Handy sufficient conditions for the convergence of the maximum likelihood estimator in observation-driven models

Randal Douc, François Roueff, Tepmony Sim

Handy sufficient conditions for the convergence of the maximum likelihood estimator in observation-driven models

Randal Douc*[1], François Roueff†[2] and Tepmony Sim‡[2]

[1]Department CITI, CNRS UMR 5157, Telecom Sudparis, Evry, France.
[2]Institut Mines-Telecom, Telecom Paristech, CNRS LTCI, Paris, France.

April 7, 2015

Abstract

This paper generalizes asymptotic properties obtained in the observation-driven times series models considered by [7] in the sense that the conditional law of each observation is also permitted to depend on the parameter. The existence of ergodic solutions and the consistency of the Maximum Likelihood Estimator (MLE) are derived under easy-to-check conditions. The obtained conditions appear to apply for a wide class of models. We illustrate our results with specific observation-driven times series, including the recently introduced NBIN-GARCH and NM-GARCH models, demonstrating the consistency of the MLE for these two models.

MSC: Primary: 62F12; Secondary: 60J05.
Keywords: consistency, ergodicity, maximum likelihood, observation-driven models, time series of counts.

1 Introduction

Observation-driven time series models have been widely used in various disciplines such as in economics, finance, epidemiology, population dynamics, etc. These models have been introduced by [4] and later considered by [19],

*randal.douc@telecom-sudparis.eu
†roueff@telecom-paristech.fr
‡sim@telecom-paristech.fr

[5], [11], [17], [9], [6] and [7]. The celebrated GARCH(1, 1) model, see [2], as well as most of the models derived from this one, see [3] for a list of some of them, are typical examples of observation-driven models. Observation-driven models have the nice feature that the associated (conditional) likelihood and its derivatives are easy to compute and the prediction is straightforward. The consistency of the maximum likelihood estimator (in short, MLE) for the class of these models can be cumbersome, except when it can be derived using computations specific to the studied model (the GARCH(1,1) case being one of the most celebrated example). When the observed variable is discrete, general consistency results have been obtained only recently in [6] or [7] (see also in [13] for the existence of stationary and ergodic solutions to some observation-driven time series models). However, the consistency result of [7] applies to some restricted class of models and does not cover the case where the distribution of the observations given the hidden variable also depends on an unknown parameter. We now introduce three simple examples, to which the results of [7] can not be directly applied. The first one is the negative binomial integer-valued GARCH (NBIN-GARCH) model, which was first introduced by [20] as a generalization of the Poisson IN-GARCH model. The NBIN-GARCH model belongs to the class of integer-valued GARCH models that account for overdispersion (i.e., variability is larger than mean) and potential heavy tails in the high values. In [20], the author applied this model to treat the data of counts of poliomyelitis cases in the USA from 1970 to 1983 reported by the Centres for Disease Control, where data overdispersion was detected. The estimation result showed that NBIN-GARCH(1, 1) outperformed among some commonly used models such as Poisson and Double Poisson models. The NBIN-GARCH(1, 1) model is formally defined as follows.

Example 1 (NBIN-GARCH(1, 1) model)**.** Consider the following recursion.

$$\begin{aligned} X_{k+1} &= \omega + aX_k + bY_k \ , \\ Y_{k+1}|X_{0:k+1}, Y_{0:k} &\sim \mathcal{NB}\left(r, \frac{X_{k+1}}{1+X_{k+1}}\right) \ , \end{aligned} \tag{1}$$

where X_k takes values in $\mathsf{X} = \mathbb{R}_+$, Y_k takes values in $\mathbb{Z}_+$ and $\theta = (\omega, a, b, r) \in (0, \infty)^4$ is an unknown parameter. In (1), $\mathcal{NB}(r, p)$ denotes the negative binomial distribution with parameters $r > 0$ and $p \in (0, 1)$, that is: if $Y \sim \mathcal{NB}(r, p)$, then $\mathbb{P}(Y = k) = \frac{\Gamma(k+r)}{k!\Gamma(r)}(1-p)^r p^k$ for all $k \geq 0$, where Γ stands for the Gamma function. Though substantial analysis on this model has been carried out in the literature, to the best of our knowledge, the consistency of the MLE has not been treated, see the end of the discussions of Section 6 in [20].

The second example is the univariate normal mixture GARCH model (NM-GARCH) proposed by [12] and later considered by [1]. The NM-GARCH model is another natural extension of GARCH processes, where

the usual Gaussian conditional distribution of the observations given the hidden volatility variable is replaced by a mixture of Gaussian distributions given a hidden vector volatility variable. The NM-GARCH model has the ability of capturing time variation in both conditional skewness and kurtosis, while the classical GARCH cannot. In [1], the NM-GARCH$(1,1)$ model was applied to examine the data of exchange rates consisting of daily prices in US dollars of three different currencies (British pound, euro and Japanese yen) from 2 January 1989 to 31 December 2002. The empirical evidence suggested the best performance of NM(2)-GARCH$(1,1)$ when compared to the classical GARCH$(1,1)$, standardized symmetric and skewed t-GARCH$(1,1)$ models applied to this same data. The definition of this model is formally stated as follows.

Example 2 (NM(d)-GARCH$(1,1)$ model)**.** Let $d \in \mathbb{N} \setminus \{0\}$ and consider the following recursion.

$$\begin{aligned} &\mathbf{X}_{k+1} = \boldsymbol{\omega} + \mathbf{A}\mathbf{X}_k + Y_k^2 \mathbf{b} \ , \\ &Y_{k+1} | \mathbf{X}_{0:k+1}, Y_{0:k} \sim G^\theta(\mathbf{X}_{k+1}; \cdot) \ , \\ &\frac{\mathrm{d}G^\theta(\mathbf{x}; \cdot)}{\mathrm{d}\nu}(y) = \sum_{\ell=1}^{d} \gamma_\ell \frac{\mathrm{e}^{-y^2/2x_\ell}}{(2\pi x_\ell)^{1/2}} \ , \quad \mathbf{x} \in (0,\infty)^d, \ y \in \mathbb{R} \ , \end{aligned} \tag{2}$$

where ν is the Lebesgue measure on $\mathbb{R}$, $\mathbf{X}_k = [X_{1,k} \dots X_{d,k}]^T$ takes values in $\mathsf{X} = \mathbb{R}_+^d$; $\boldsymbol{\gamma} = [\gamma_1 \dots \gamma_d]^T$ a d-dimensional vector of mixture coefficients belonging to the d-dimensional simplex

$$\mathsf{P}_d = \left\{ \boldsymbol{\gamma} \in \mathbb{R}_+^d \ : \ \sum_{\ell=1}^{d} \gamma_\ell = 1 \right\} \ , \tag{3}$$

$\boldsymbol{\omega}$, $\mathbf{b}$ are d-dimensional vector parameters with positive and non-negative entries, respectively and $\mathbf{A}$ is a $d \times d$ matrix parameter with non-negative entries. Here we have $\theta = (\boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{A}, \mathbf{b})$. Note that G^θ depends on θ only through the mixture coefficients $\gamma_1, \dots, \gamma_d$. If $d = 1$, we obtain the usual conditionally Gaussian GARCH(1,1) process. In such a case, since $\boldsymbol{\gamma} = \gamma_1 = 1$, G^θ no longer depends on θ. Up to our knowledge, the usual consistency proof of the MLE for the GARCH cannot be directly adapted to this model.

Finally, we consider the following new example, where a threshold is added to the usual INGARCH model in the conditional distribution.

Example 3 (Threshold INGARCH model)**.** Consider the following recursion.

$$\begin{aligned} &X_{k+1} = \omega + aX_k + bY_k \ , \\ &Y_{k+1} | X_{0:k+1}, Y_{0:k} \sim \mathcal{P}\left(X_{k+1} \wedge \tau\right) \ , \end{aligned} \tag{4}$$

where X_k takes values in $\mathsf{X} = (0, \infty)$, Y_k takes values in $\mathbb{Z}_+$ and $\theta = (\omega, a, b, \tau) \in (0, \infty)^4$ is an unknown parameter. Comparing with the usual INGARCH model, a threshold τ has been added in the conditional observation distribution. This corresponds to the practical case where the hidden variable has an influence on the observation up to this threshold.

For a well-specified model, a classical approach to establish the consistency of the MLE generally involves two main steps: first the maximum likelihood estimator (MLE) converges to the maximizing set $\Theta_\star$ of a limit criterion, and second the maximizing set indeed reduces to the true parameter $\theta_\star$, which is usually referred to as solving the *identifiability* problem. In this paper, we are interested in solving the problem involved in the first step, that is, the convergence of MLE. We extend the convergence result of MLE obtained in [7], which is valid for a restricted class of models, to a larger class of models in which the three examples introduced above are embedded. More precisely, we show the convergence of MLE in observation-driven models where the probability distributions of observations explicitly depend on the unknown parameters. Moreover, we provide very simple conditions that are easy to check, as shown by the three illustrating examples.

The paper is organized as follows. Specific definitions and notation are introduced in Section 2. Then, Section 3 contains the main contribution of the paper, that is, sufficient conditions for the existence of ergodic solutions and for the consistency of the MLE. These results are then applied in Section 4 to the three examples introduced above. Numerical experiments for the NBIN-GARCH(1, 1) model are given in Section 5. Finally, Section 6 provides the proofs of the main results, mainly inspired from [7].

2 Definitions and notation

Consider a bivariate stochastic process $\{(X_k, Y_k) : k \in \mathbb{Z}_+\}$ on $\mathsf{X} \times \mathsf{Y}$, where (X, d) is a complete and separable metric space endowed with the associated Borel σ-field $\mathcal{X}$ and $(\mathsf{Y}, \mathcal{Y})$ is a Borel space. Let (Θ, Δ), the set of parameters, be a compact metric space, $\{G^\theta : \theta \in \Theta\}$ be a family of probability kernels on $\mathsf{X} \times \mathcal{Y}$ and $\{(x, y) \mapsto \psi_y^\theta(x) : \theta \in \Theta\}$ be a family of measurable functions from $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ to $(\mathsf{X}, \mathcal{X})$. The observation-driven time series model can be formally defined as follows.

Definition 1. A time series $\{Y_k : k \in \mathbb{Z}_+\}$ valued in Y is said to be distributed according to an *observation-driven model* with parameter $\theta \in \Theta$ if there is a bivariate Markov chain $\{(X_k, Y_k) : k \in \mathbb{Z}_+\}$ on $\mathsf{X} \times \mathsf{Y}$ whose transition kernel K^θ satisfies

$$K^\theta((x, y); \mathrm{d}x'\mathrm{d}y') = \delta_{\psi_y^\theta(x)}(\mathrm{d}x')\, G^\theta(x'; \mathrm{d}y') \,, \tag{5}$$

where δ_a denotes the Dirac mass at point a. Moreover, we will say that the observation-driven time series model is dominated by some σ-finite measure

ν on $(\mathsf{Y}, \mathcal{Y})$ if for all $x \in \mathsf{X}$, the probability kernel $G^\theta(x;\cdot)$ is dominated by ν. In this case we denote by $g^\theta(x;\cdot)$ its Radon-Nikodym derivative, $g^\theta(x;y) = \frac{\mathrm{d}G^\theta(x;\cdot)}{\mathrm{d}\nu}(y)$, and we always assume that for all $(x,y) \in \mathsf{X} \times \mathsf{Y}$ and for all $\theta \in \Theta$,

$$g^\theta(x;y) > 0 \; .$$

A dominated parametric observation-driven model is thus characterized by the collection $\{(g^\theta, \psi^\theta) \; : \; \theta \in \Theta\}$. The class of observation-driven time series models is a particular case of *partially-observed Markov chains* since only Y_k's are observed, whereas X_k's are *hidden* variables. Note that our notation for observation-driven models is slightly different from that of [7] where their sequence $\{Y_k\}$ corresponds to our sequence $\{Y_{k-1}\}$. Note also that the process $\{X_k \; : \; k \geq 1\}$ by itself is a Markov chain with transition kernel defined by

$$R^\theta(x;A) = \int 1_A(\psi_y^\theta(x)) \; G^\theta(x;\mathrm{d}y), \quad x \in \mathsf{X}, \; A \in \mathcal{X} \; . \tag{6}$$

However, observation-driven time series models do not belong to the class of hidden Markov models. This can be seen in the following recursive relation, which holds for all $k \geq 0$,

$$\begin{aligned} X_{k+1} &= \psi_{Y_k}^\theta(X_k) \; , \\ Y_{k+1} \mid \mathcal{F}_k &\sim G^\theta(X_{k+1};\cdot) \; , \end{aligned} \tag{7}$$

where $\mathcal{F}_k = \sigma\left(X_\ell, X_{\ell+1}, Y_\ell \; : \; \ell \leq k, \ell \in \mathbb{Z}_+\right)$ and which can be represented graphically as below.

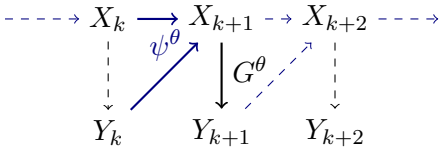

Figure 1: Graphical representation of the observation-driven model.

The most popular example is the GARCH(1,1) process, where $G^\theta(x;\cdot)$ is a centered (say Gaussian) distribution with variance x and $\psi_y^\theta(x)$ is an affine function of x and y^2. One can readily check that Examples 1 and 2 are other instances of dominated observation-driven models.

The inference about model parameter is carried out by relying on the conditional likelihood of the observations $(Y_1, \ldots, Y_n)$ given $X_1 = x$ for an arbitrary $x \in \mathsf{X}$. The corresponding conditional density function with respect to $\nu^{\otimes n}$ is, under parameter θ, for all $x \in \mathsf{X}$,

$$y_{1:n} \mapsto \prod_{k=1}^{n} g^{\theta}\left(\psi^{\theta}\langle y_{1:k-1}\rangle(x); y_k\right) , \tag{8}$$

where, for any vector $y_{1:p} = (y_1, \ldots, y_p) \in \mathsf{Y}^p$, $\psi^{\theta}\langle y_{1:p}\rangle$ is the $\mathsf{X} \to \mathsf{X}$ function obtained as the successive composition of $\psi^{\theta}_{y_1}$, $\psi^{\theta}_{y_2}$, ..., and $\psi^{\theta}_{y_p}$,

$$\psi^{\theta}\langle y_{1:p}\rangle = \psi^{\theta}_{y_p} \circ \psi^{\theta}_{y_{p-1}} \circ \cdots \circ \psi^{\theta}_{y_1} , \tag{9}$$

with the convention $\psi^{\theta}\langle y_{s:t}\rangle(x) = x$ for $s > t$. Then, the corresponding (conditional) Maximum Likelihood Estimator (MLE) $\hat{\theta}_{x,n}$ of the parameter θ, is defined by

$$\hat{\theta}_{x,n} \in \operatorname*{argmax}_{\theta\in\Theta} \mathsf{L}^{\theta}_{x,n}\langle Y_{1:n}\rangle , \tag{10}$$

where

$$\mathsf{L}^{\theta}_{x,n}\langle y_{1:n}\rangle := n^{-1} \sum_{k=1}^{n} \ln g^{\theta}\left(\psi^{\theta}\langle y_{1:k-1}\rangle(x); y_k\right) . \tag{11}$$

In this contribution, we study the convergence of $\hat{\theta}_{x,n}$ as $n \to \infty$ for some well-chosen value of x under the assumption that the model is well specified and the observations are in a steady state. This means that we assume that the observations $\{Y_k : k \in \mathbb{Z}_+\}$ are distributed according to $\tilde{\mathbb{P}}^{\theta_\star}$ with $\theta_\star \in \Theta$, where, for all $\theta \in \Theta$, $\tilde{\mathbb{P}}^{\theta}$ denotes the stationary distribution of the observation-driven time series corresponding to the parameter θ. However whether such a distribution is well defined is not always obvious. We will use the following ergodicity assumption.

(A-1) For all $\theta \in \Theta$, the transition kernel K^{θ} of the complete chain admits a unique stationary distribution π^{θ} on $\mathsf{X} \times \mathsf{Y}$.

With this assumption, we can now define $\tilde{\mathbb{P}}^{\theta}$. The following notation and definitions will be used throughout the paper.

Definition 2. For any probability distribution μ on $\mathsf{X} \times \mathsf{Y}$, we denote by $\mathbb{P}^{\theta}_{\mu}$ the distribution of the Markov chain $\{(X_k, Y_k), k \geq 0\}$ with kernel K^{θ} and initial probability mesure μ. Under Assumption (A-1), we denote by π^{θ}_1 and π^{θ}_2 the marginal distributions of π^{θ} on X and Y, respectively and by $\mathbb{P}^{\theta}$ and $\tilde{\mathbb{P}}^{\theta}$ the probability distributions defined respectively as follows.

a) $\mathbb{P}^{\theta}$ denotes the extension of $\mathbb{P}^{\theta}_{\pi^{\theta}}$ on the whole line $(\mathsf{X} \times \mathsf{Y})^{\mathbb{Z}}$.

b) $\tilde{\mathbb{P}}^{\theta}$ is the corresponding projection on the component $\mathsf{Y}^{\mathbb{Z}}$.

The probability distributions $\mathbb{P}^{\theta}$ and $\tilde{\mathbb{P}}^{\theta}$ are more formally defined by setting, for all $m \in \mathbb{Z}$ and $B \in \mathcal{Y}^{\otimes(m+\mathbb{Z}^*_+)}$,

$$\tilde{\mathbb{P}}^{\theta}\left(\mathsf{Y}^{m+\mathbb{Z}_-}\times B\right)=\mathbb{P}^{\theta}\left(\mathsf{X}^{\mathbb{Z}}\times\left(\mathsf{Y}^{m+\mathbb{Z}_-}\times B\right)\right)=\mathbb{P}^{\theta}_{\pi^{\theta}}\left(\mathsf{X}^{m+\mathbb{Z}_+^*}\times B\right)\ , \quad (12)$$

or equivalently, using the canonical functions Y_k, $k\in\mathbb{Z}$,

$$\tilde{\mathbb{P}}^{\theta}\left(Y_{m+1:\infty}\in B\right)=\mathbb{P}^{\theta}\left(Y_{m+1:\infty}\in B\right)=\mathbb{P}^{\theta}_{\pi^{\theta}}\left(Y_{m+1:\infty}\in B\right)\ . \quad (13)$$

Here and in what follows, we abusively use the same notation Y_k both for the canonical projection defined on $\mathsf{Y}^{\mathbb{Z}}$ and for the one defined on $(\mathsf{X}\times\mathsf{Y})^{\mathbb{Z}_+}$. We also use the symbols $\mathbb{E}^{\theta}$ and $\tilde{\mathbb{E}}^{\theta}$ to denote the expectations corresponding to $\mathbb{P}^{\theta}$ and $\tilde{\mathbb{P}}^{\theta}$, respectively.

3 Main results

3.1 Preliminaries

In this section, we follow the same lines as in [7] to derive the convergence of the MLE $\hat{\theta}_{x,n}$ for a general class of observation-driven models. The approach is to establish that, as the number of observations $n\to\infty$, there exists a $(\mathsf{Y}^{\mathbb{Z}},\mathcal{Y}^{\otimes\mathbb{Z}})\to(\mathbb{R},\mathcal{B}(\mathbb{R}))$ measurable function $p^{\theta}(\cdot|\cdot)$ such that the normalized log-likelihood $\mathsf{L}^{\theta}_{x,n}\langle Y_{1:n}\rangle$ defined in (11), for some appropriate value of x, can be approximated by

$$n^{-1}\sum_{k=1}^{n}\ln p^{\theta}(Y_k|Y_{-\infty:k-1})\ .$$

To define $p^{\theta}(\cdot|\cdot)$, we set, for all $y_{-\infty:1}\in\mathsf{Y}^{\mathbb{Z}_-}$, whenever the following limit is well defined,

$$p^{\theta}\left(y_1\,|\,y_{-\infty:0}\right)=\begin{cases}\lim\limits_{m\to\infty}g^{\theta}\left(\psi^{\theta}\langle y_{-m:0}\rangle(x);y_1\right) & \text{if the limit exists,}\\ \infty & \text{otherwise.}\end{cases} \quad (14)$$

By (A-1), the process Y is ergodic under $\tilde{\mathbb{P}}^{\theta_\star}$ and provided that

$$\tilde{\mathbb{E}}^{\theta_\star}\left[\ln^{+}p^{\theta}(Y_1|Y_{-\infty:0})\right]<\infty\ ,$$

it follows that

$$\lim_{n\to\infty}\mathsf{L}^{\theta}_{x,n}\langle Y_{1:n}\rangle=\tilde{\mathbb{E}}^{\theta_\star}\left[\ln p^{\theta}(Y_1|Y_{-\infty:0})\right]\ ,\quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}$$

In this paper we show that with probability tending to one, the MLE $\hat{\theta}_{x,n}$ eventually lies in a neighborhood of the set

$$\Theta_\star=\operatorname*{argmax}_{\theta\in\Theta}\tilde{\mathbb{E}}^{\theta_\star}\left[\ln p^{\theta}(Y_1|Y_{-\infty:0})\right]\ , \quad (15)$$

which only depends on $\theta_\star$. In this contribution, we provide easy-to-check sufficient conditions implying

$$\lim_{n\to\infty} \Delta(\hat{\theta}_{x,n}, \Theta_\star) = 0, \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}, \tag{16}$$

but, for the sake of brevity, we do not precisely determine the set $\Theta_\star$. Many approaches have been proposed to investigate this problem, which is often referred to as the *identifiability* problem. In particular cases, one can prove that $\Theta_\star = \{\theta_\star\}$, in which case the strong consistency of the MLE follows from (16). We will mention a general result which precises how the set $\Theta_\star$ is related to the true parameter $\theta_\star$ in Remark 3. For the moment, let us mention that we have

$$\theta_\star \in \Theta_\star \,, \tag{17}$$

provided that the following assumption holds:

(B-1) For all $\theta, \theta_\star \in \Theta$, we have

- (i) If $\theta \neq \theta_\star$, $y \mapsto p^\theta(y|Y_{-\infty:0})$ is a density function $\tilde{\mathbb{P}}^{\theta_\star}$-a.s.
- (ii) Under $\tilde{\mathbb{P}}^{\theta_\star}$, the function $y \mapsto p^{\theta_\star}(y|Y_{-\infty:0})$ is the conditional density function of Y_1 given $Y_{-\infty:0}$.

Indeed, (17) follows by writing for all $\theta \in \Theta$,

$$\begin{aligned}\tilde{\mathbb{E}}^{\theta_\star}\left[\ln p^{\theta_\star}(Y_1|Y_{-\infty:0}) - \ln p^{\theta}(Y_1|Y_{-\infty:0})\right] &= \tilde{\mathbb{E}}^{\theta_\star}\left[\ln \frac{p^{\theta_\star}(Y_1|Y_{-\infty:0})}{p^{\theta}(Y_1|Y_{-\infty:0})}\right] \\ &= \tilde{\mathbb{E}}^{\theta_\star}\left[\tilde{\mathbb{E}}^{\theta_\star}\left[\ln \frac{p^{\theta_\star}(Y_1|Y_{-\infty:0})}{p^{\theta}(Y_1|Y_{-\infty:0})}\middle| Y_{-\infty:0}\right]\right] ,\end{aligned}$$

which is nonnegative under (B-1) since it is the expectation of a conditional Kullback-Leibler divergence.

3.2 Convergence of the MLE

In this part, we always assume that (A-1) holds. The following is a list of additional assumptions on which our convergence result relies.

(A-2) There exists a function $\bar{V} : \mathsf{X} \to \mathbb{R}_+$ such that, for all $\theta \in \Theta$, $\pi_1^\theta(\bar{V}) < \infty$.

Remark 1. Assumption (A-2) is usually obtained as a byproduct of the proof of Assumption (A-1), see Section 3.3. It is here stated as an assumption for convenience.

The following set of conditions can readily be checked on g^θ and ψ^θ.

(B-2) For all $y \in \mathsf{Y}$, the function $(\theta, x) \mapsto g^\theta(x; y)$ is continuous on $\Theta \times \mathsf{X}$.

(B-3) For all $y \in \mathsf{Y}$, the function $(\theta, x) \mapsto \psi_y^\theta(x)$ is continuous on $\Theta \times \mathsf{X}$.

The function $\bar{V}$ appearing in (B-4)(viii) below is the same one as in Assumption (A-2). Moreover, in this condition and throughout the paper we write $f \lesssim V$ for a real-valued function f and a nonnegative function V defined on the same space X, whenever there exists a positive constant c such that $|f(x)| \leq cV(x)$ for all $x \in \mathsf{X}$.

(B-4) There exist $x_1 \in \mathsf{X}$, a closed set $\mathsf{X}_1 \subseteq \mathsf{X}$, $\varrho \in (0,1)$, $C \geq 0$ and measurable functions $\bar{\psi} : \mathsf{X}_1 \to \mathbb{R}_+$, $H : \mathbb{R}_+ \to \mathbb{R}_+$ and $\bar{\phi} : \mathsf{Y} \to \mathbb{R}_+$ such that the following assertions hold.

(i) For all $\theta \in \Theta$ and $(x,y) \in \mathsf{X} \times \mathsf{Y}$, $\psi_y^\theta(x) \in \mathsf{X}_1$.

(ii) $\sup_{(\theta,x,y) \in \Theta \times \mathsf{X}_1 \times \mathsf{Y}} g^\theta(x;y) < \infty$.

(iii) For all $\theta \in \Theta$, $n \in \mathbb{Z}_+$, $x \in \mathsf{X}$, and $y_{1:n} \in \mathsf{Y}^n$,

$$\mathrm{d}\left(\psi^\theta\langle y_{1:n}\rangle(x_1), \psi^\theta\langle y_{1:n}\rangle(x)\right) \leq \varrho^n \, \bar{\psi}(x) \,, \tag{18}$$

(iv) $\bar{\psi}$ is locally bounded.

(v) For all $\theta \in \Theta$ and $y \in \mathsf{Y}$, $\bar{\psi}(\psi_y^\theta(x_1)) \leq \bar{\phi}(y)$.

(vi) For all $\theta \in \Theta$ and $(x, x', y) \in \mathsf{X}_1 \times \mathsf{X}_1 \times \mathsf{Y}$,

$$\left|\ln \frac{g^\theta(x;y)}{g^\theta(x';y)}\right| \leq H(\mathrm{d}(x,x')) \, \mathrm{e}^{C\,(\mathrm{d}(x_1,x) \vee \mathrm{d}(x_1,x'))} \, \bar{\phi}(y) \,, \tag{19}$$

(vii) $H(u) = O(u)$ as $u \to 0$.

(viii) If $C = 0$, then, for all $\theta \in \Theta$,

$$G^\theta \ln^+ \bar{\phi} \lesssim \bar{V} \,, \tag{20}$$

otherwise, for all $\theta \in \Theta$,

$$G^\theta \bar{\phi} \lesssim \bar{V} \,. \tag{21}$$

Let us now state our main result as follows.

Theorem 3. *Assume that (A-1), (A-2), (B-2), (B-3) and (B-4) hold. Then, letting $x_1 \in \mathsf{X}$ as in (B-4), the function $p^\theta(\cdot|\cdot)$ defined by (14) with $x = x_1$ satisfies (B-1) and the convergence (16) of the MLE holds with the set $\Theta_\star$ defined by (15).*

For convenience, the proof is postponed to Section 6.1.

Remark 2. As noticed in [7], the techniques used to prove Theorem 3 also apply in the misspecified case, where Y is not distributed according to $\tilde{\mathbb{P}}^{\theta_\star}$. We do not pursue in this direction in this contribution.

The consistency of the MLE then follows from Theorem 3 by the following remark.

Remark 3. In many specific cases, one can show that $\Theta_\star$ defined by (15) is the singleton $\{\theta_\star\}$. However this task appears to be quite difficult in some cases such as Example 3. Instead one can use [8, Section 4.2], where it is shown that the assumptions of Theorem 3 imply that $\Theta_\star$ is exactly the set of parameters θ such that $\tilde{\mathbb{P}}^\theta = \tilde{\mathbb{P}}^{\theta_\star}$. Thus we can conclude that the MLE converges to the *equivalence class* of the true parameter. This type of consistency has been introduced by [14] in the context of hidden Markov models in order to disentangle the proof of the consistency from the problem of identifiability. Recall that the model is identifiable if and only if the equivalent classes $\{\theta \ : \ \tilde{\mathbb{P}}^\theta = \tilde{\mathbb{P}}^{\theta_\star}\}$ reduce to singletons $\{\theta_\star\}$ for all $\theta_\star \in \Theta$.

3.3 Ergodicity

In this section, the observation-driven model is studied to prove the condition (A-1). Since this is a "for all θ (...)" condition, to save space and alleviate the notational burden, we will drop the superscript θ from, for example, G^θ, R^θ and ψ^θ and respectively write G, R and ψ, instead.

Ergodicity of Markov chains are usually studied using ψ-irreducibility. This approach is well known to be quite efficient when dealing with fully dominated models, see [15]. It is not at all the same picture for observation-driven models, where other tools need to be invoked, see [10, 7]. Since the ergodicity is studied for a given parameter θ, the ergodicity results of [7] directly apply, even though observation-driven models are restricted to the case where g does not depend on the unknown parameter θ in this reference. Our main contribution here is to focus on an easy-to-check list of assumptions yielding the ergodicity conditions (A-1) and (A-2). We also provide a lemma (Lemma 5) which gives the construction of the instrumental functions α and ϕ used in the list of assumptions.

(A-3) The measurable space (X, d) is a locally compact, complete and separable metric space and its associated σ-field $\mathcal{X}$ is the Borel σ-field.

(A-4) There exist $(\lambda, \beta) \in (0,1) \times \mathbb{R}_+$ and a measurable function $V : \mathsf{X} \to \mathbb{R}_+$ such that $RV \leq \lambda V + \beta$ and $\{V \leq M\}$ is compact for any $M > 0$.

(A-5) The Markov kernel R is weak Feller, that is, for any continuous and bounded function f defined on X, Rf is continuous and bounded on X.

(A-6) The Markov kernel R has a reachable point, that is, there exists $x_0 \in \mathsf{X}$ such that, for any $x \in \mathsf{X}$ and any neighborhood $\mathcal{N}$ of x_0, $R^m(x; \mathcal{N}) > 0$ for at least one positive integer m.

(A-7) We have $\sup_{\substack{(x,x',y)\in \mathsf{X}^2\times\mathsf{Y}\\ x\neq x'}} \frac{\mathrm{d}(\psi_y(x),\psi_y(x'))}{\mathrm{d}(x,x')} < 1.$

(A-8) There exist a measurable function α from X^2 to $[0,1]$, a measurable function $\phi : \mathsf{X}^2 \to \mathsf{X}$ and a measurable function $W : \mathsf{X}^2 \to [1,\infty)$ such that the following assertions hold.

(i) For all $(x,x') \in \mathsf{X}^2$ and $y \in \mathsf{Y}$,

$$\min\left\{g(x;y), g(x';y)\right\} \geq \alpha(x,x') g\left(\phi(x,x');y\right) \ . \tag{22}$$

(ii) For all $x \in \mathsf{X}$, $W(x,\cdot)$ is finitely bounded in a neighborhood of x, that is, there exists $\gamma_x > 0$ such that $\sup_{x'\in \mathsf{B}(x,\gamma_x)} W(x,x') < \infty$.

(iii) For all $(x,x') \in \mathsf{X}^2$, $1-\alpha(x,x') \leq \mathrm{d}(x,x')W(x,x')$.

(iv) $\sup\left(\int_{\mathsf{Y}} W(\psi_y(x),\psi_y(x'))\, G(\phi(x,x');\mathrm{d}y) - W(x,x')\right) < \infty$, where the sup is taken over all $(x,x') \in \mathsf{X}^2$.

We can now state the main ergodicity result.

Theorem 4. *Conditions (A-3), (A-4), (A-5), (A-6), (A-7) and (A-8) imply that K admits a unique stationary distribution π on $\mathsf{X}\times\mathsf{Y}$. Moreover $\pi_1 \bar{V} < \infty$ for every $\bar{V} : \mathsf{X} \to \mathbb{R}_+$ such that $\bar{V} \lesssim V$.*

The proof of Theorem 4 is postponed to Section 6.2 for convenience.

The first conclusion of Theorem 4 can directly be applied for all $\theta \in \Theta$ to check (A-1). The second conclusion can be used to check (A-2). In doing so, one must take care of the fact that although V may depend on θ, $\bar{V}$ does not.

Assumptions (A-4), (A-5) and (A-6) have to be checked directly on the Markov kernel R defined by (6). To this end it can be useful to define, for any given $x \in \mathsf{X}$, the distribution

$$\bar{\mathbb{P}}_x := \mathbb{P}_{\delta_x \otimes G(x;\cdot)} \tag{23}$$

on $(\mathsf{X}\times\mathsf{Y})^{\mathbb{Z}_+}$, where $\mathbb{P}_\mu$ is defined for any distribution μ on $\mathsf{X}\times\mathsf{Y}$ as in Definition 2. Then the first component process $\{X_k,\, k \in \mathbb{Z}_+\}$ associated to $\bar{\mathbb{P}}_x$ is a Markov chain with Markov kernel R and initial distribution δ_x.

We now provide a general framework for constructing α and ϕ that appear in (A-8).

Lemma 5. *Suppose that $\mathsf{X} = \mathsf{C}^{\mathsf{S}}$ for some measurable space $(\mathsf{S},\mathcal{S})$ and $\mathsf{C} \subseteq \mathbb{R}$. Thus for all $x \in \mathsf{X}$, we write $x = (x_s)_{s\in\mathsf{S}}$, where $x_s \in \mathsf{C}$ for all $s \in \mathsf{S}$. Suppose moreover that for all $x = (x_s)_{s\in\mathsf{S}} \in \mathsf{X}$, we can express the conditional density $g(x;\cdot)$ as a mixture of densities of the form $j(x_s)h(x_s;\cdot)$*

over $s \in \mathsf{S}$. This means that for all $t \in \mathsf{C}$, $y \mapsto j(t)h(t;y)$ is a density with respect to ν and there exists a probability measure μ on $(\mathsf{S}, \mathcal{S})$ such that

$$g(x;y) = \int_{\mathsf{S}} j(x_s)h(x_s;y)\mu(\mathrm{d}s) \;, \quad y \in \mathsf{Y} \;. \tag{24}$$

We moreover assume that h takes non-negative values and that one of the two following assumptions holds.

(F-1) For all $y \in \mathsf{Y}$, the function $h(\cdot;y) : t \mapsto h(t;y)$ is non-decreasing.

(F-2) For all $y \in \mathsf{Y}$, the function $h(\cdot;y) : t \mapsto h(t;y)$ is non-increasing.

For all $(x,x') \in \mathsf{X}^2$, denoting $x \wedge x' := (\min\{x_s, x'_s\})_{s\in\mathsf{S}}$ and $x \vee x' := (\max\{x_s, x'_s\})_{s\in\mathsf{S}}$, we define $\alpha(x,x')$ and $\phi(x,x')$ as

$$\begin{cases} \alpha(x,x') = \inf\limits_{s\in\mathsf{S}} \left\{ \dfrac{j(x_s \vee x'_s)}{j(x_s \wedge x'_s)} \right\} & \text{and} \quad \phi(x,x') = x \wedge x' \quad \text{under (F-1)} \;; \\ \alpha(x,x') = \inf\limits_{s\in\mathsf{S}} \left\{ \dfrac{j(x_s \wedge x'_s)}{j(x_s \vee x'_s)} \right\} & \text{and} \quad \phi(x,x') = x \vee x' \quad \text{under (F-2)} \;. \end{cases}$$

Then α and ϕ defined above satisfy (A-8)(i).

Proof. We only prove this result under Condition (F-1). The proof is similar under (F-2).

Since for all $t \in \mathsf{C}$, $y \mapsto j(t)h(t;y)$ is a density with respect to ν, we have

$$j(t) = \left(\int h(t;y)\nu(\mathrm{d}y) \right)^{-1} > 0 \;.$$

Thus j is non-increasing on C. Clearly, the defined α takes values on $[0,1]$ and ϕ defines a function from X^2 to X. For all $(x,x') \in \mathsf{X}^2$ and $y \in \mathsf{Y}$, we have

$$\begin{aligned} g(x;y) &= \int_{\mathsf{S}} j(x_s)h(x_s;y)\mu(\mathrm{d}s) \\ &\geq \int_{\mathsf{S}} j(x_s \vee x'_s)h(x_s \wedge x'_s;y)\mu(\mathrm{d}s) \\ &\geq \int_{\mathsf{S}} \frac{j(x_s \vee x'_s)}{j(x_s \wedge x'_s)} j(x_s \wedge x'_s)h(x_s \wedge x'_s;y)\mu(\mathrm{d}s) \\ &\geq \int_{\mathsf{S}} \inf_{s\in\mathsf{S}} \left\{ \frac{j(x_s \vee x'_s)}{j(x_s \wedge x'_s)} \right\} j(x_s \wedge x'_s)h(x_s \wedge x'_s;y)\mu(\mathrm{d}s) \\ &= \alpha(x,x')g(\phi(x,x');y) \;. \end{aligned}$$

By symmetry of α and ϕ, we get (22) and thus (A-8)(i) holds. □

4 Examples

Let us now apply these results to prove the convergence of MLE of Examples 1, 2 and 3.

4.1 NBIN-GARCH model

Example 1 is a specific case of Definition 1 where ν is the counting measure on $\mathsf{Y} = \mathbb{N}$,

$$\psi_y^\theta(x) = \omega + ax + by \ , \tag{25}$$

$$g^\theta(x;y) = \frac{\Gamma(y+r)}{y!\Gamma(r)} \left(\frac{1}{1+x}\right)^r \left(\frac{x}{1+x}\right)^y \ , \tag{26}$$

with $\theta = (\omega, a, b, r)$ in a compact subset Θ of $(0,\infty)^4$ and $\mathsf{X} = (0,\infty)$.

In [20, Theorem 1], the equation satisfied by the mean of the observations $\mu_k = \mathbb{E}[Y_k]$ is derived and is shown to admit a constant solution if and only if

$$rb + a < 1 \ . \tag{27}$$

This clearly implies that this condition is necessary to have a stationary solution $\{Y_k\}$ with finite mean. However it does not imply the existence of such a solution. In fact, the following result shows that (27) is indeed a necessary and sufficient condition to have a stationary solution $\{Y_k\}$ with finite mean. It also shows that all the assumptions of Theorem 3 hold, which, with Remark 3, provides the consistency of the MLE $\hat{\theta}_{x_1,n}$ for any $x_1 \in \mathsf{X}$.

Theorem 6. *Suppose that all $\theta = (\omega, a, b, r)$ in Θ satisfy Condition (27). Then Assumptions (A-1), (A-2), (B-2), (B-3) and (B-4) hold with $\bar{V}$ being defined as the identity function on X and with any $x_1 \in \mathsf{X}$.*

Proof. For convenience, we divide the proof into two steps.
Step 1. We first prove Assumptions (A-1) and (A-2) by applying Theorem 4. We set $\bar{V}(x) = V(x) = x$ and thus we only need to check (A-3), (A-4), (A-5), (A-6), (A-7) and (A-8). Condition (A-3) holds. We have for all $\theta \in \Theta$,

$$RV(x) = \omega + (a+br)x = (a+br)V(x) + \omega,$$

which yields (A-4). The fact that the kernel R is weak Feller easily follows by observing that, as $p \to p'$, $\mathcal{NB}(r,p)$ converges weakly to $\mathcal{NB}(r,p')$, so (A-5) holds.

We now prove (A-6). Let $x_\infty = \omega/(1-a)$. Let $x \in \mathbb{R}$ and define recursively the sequence $x_0 = x, x_k = \omega + ax_{k-1}$ for all positive integers k. Since $0 < a < 1$, this sequence converges to the fixed point x_∞. Therefore, defining $\mathbb{P}_x$ as in (23), for any neighborhood $\mathcal{N}$ of x_∞, there exists some n such that $x_n \in \mathcal{N}$ and we have

$$R^n(x;\mathcal{N}) = \bar{\mathbb{P}}_x\left(X_n \in \mathcal{N}\right) \geq \bar{\mathbb{P}}_x\left(X_k = x_k \text{ for all } k = 1,\ldots,n\right)$$
$$= \bar{\mathbb{P}}_x\left(Y_0 = \ldots = Y_{n-1} = 0\right) > 0.$$

So (A-6) holds. Assumption (A-7) holds since we have for all $(x, x', y) \in \mathsf{X}^2 \times \mathsf{Y}$ with $x \neq x'$,

$$\frac{|\psi_y(x) - \psi_y(x')|}{|x - x'|} = a < 1 \ .$$

To prove (A-8), we apply Lemma 5 with $\mathsf{C} = \mathsf{X}$, $\mathsf{S} = \{1\}$ (so μ boils down to the Dirac measure on $\{1\}$). For all $(x, y) \in \mathsf{X} \times \mathsf{Y}$, let $j(x) = \left(\frac{1}{1+x}\right)^r$ and $h(x;y) = \frac{\Gamma(y+r)}{y!\Gamma(r)}\left(\frac{x}{1+x}\right)^y$. Indeed, h satisfies (F-1). Thus by Lemma 5, for all $(x, x') \in \mathsf{X}^2$ and $y \in \mathsf{Y}$, we get that

$$\alpha(x, x') = \left(\frac{1 + x \wedge x'}{1 + x \vee x'}\right)^r \in (0, 1] \quad \text{and} \quad \phi(x, x') = x \wedge x'$$

satisfy (A-8)(i). For any given $r > 0$, let a function $W : \mathsf{X}^2 \to [1, \infty)$ be defined by, for all $(x, x') \in \mathsf{X}^2$, $W(x, x') = 1 \vee r$. By definition of W, as a constant function, (A-8)(ii) and (A-8)(iv) clearly hold. Moreover, (A-8)(iii) holds since for all $(x, x') \in \mathsf{X}^2$, we have that

$$1 - \alpha(x, x') \leq (1 \vee r)|x - x'| = W(x, x')|x - x'| \ .$$

Therefore, (A-8) holds, which completes **Step 1**.

Step 2. We now prove (B-2), (B-3) and (B-4). By assumption on Θ, then there exists $(\underline{\omega}, \bar{\omega}, \underline{b}, \bar{b}, \underline{r}, \bar{r}, \underline{\alpha}, \bar{\alpha}) \in (0, \infty)^6 \times (0, 1)^2$ such that

$$\underline{\omega} \leq \omega \leq \bar{\omega}, \ \leq b \leq \bar{b}, \ \underline{r} \leq r \leq \bar{r}, \ \underline{\alpha} \leq a + br \leq \bar{\alpha} \ .$$

Clearly, (B-2) and (B-3) hold by definitions of $\psi_y^\theta(x)$ and $g^\theta(x;y)$. It remains to check (B-4) for a well-chosen closed subset X_1 and any $x_1 \in \mathsf{X}$. Let $\mathsf{X}_1 = [\underline{\omega}, \infty) \subset \mathsf{X}$ so that (B-4)(i) holds. By noting that for all $(\theta, x, y) \in \Theta \times \mathsf{X} \times \mathsf{Y}$, $g^\theta(x;y) \leq 1$, we have (B-4)(ii). From (9) and (25), we have for all $s \leq t$, $y_{s:t} \in \mathsf{Y}^{t-s+1}$, $x \in \mathsf{X}$ and $\theta \in \Theta$,

$$\psi^\theta\langle y_{s:t}\rangle(x) = \omega\left(\frac{1 - a^{t-s+1}}{1 - a}\right) + a^{t-s+1}x + b\sum_{j=0}^{t-s} a^j y_{t-j} \ . \tag{28}$$

Using (28), we have, for all $\theta \in \Theta$, $x \in \mathsf{X}$ and $y_{1:n} \in \mathsf{Y}^n$,

$$\left|\psi^\theta\langle y_{1:n}\rangle(x_1) - \psi^\theta\langle y_{1:n}\rangle(x)\right| = a^n\,|x_1 - x| \leq \bar{\alpha}^n\,|x_1 - x| \ .$$

This gives (B-4)(iii) and (B-4)(iv) by setting $\varrho = \bar{\alpha} < 1$ and $\bar{\psi}(x) = |x_1 - x|$. Next we set $\bar{\phi}$, H and C to meet Conditions (B-4)(v) and (B-4)(vi) and (B-4)(vii). Let us write, for all $\theta \in \Theta$ and $y \in \mathsf{Y}$,

$$\left|x_1 - \psi_y^\theta(x_1)\right| \le \omega + (1+a)x_1 + by \le \bar{\omega} + (1+\bar{\alpha})x_1 + \bar{b}y$$

and, for all $(x, x') \in \mathsf{X}_1^2 = [\underline{\omega}, \infty)^2$,

$$\begin{aligned}\left|\ln \frac{g^\theta(x;y)}{g^\theta(x';y)}\right| &= \left|(r+y)\left[\ln(1+x') - \ln(1+x)\right] + y\left[\ln x - \ln x'\right]\right| \\ &\le \left[(r+y)(1+\underline{\omega})^{-1} + y\,\underline{\omega}^{-1}\right]\ |x - x'| \\ &\le \left[\overline{r} + y\,(1+\underline{\omega}^{-1})\right]\ |x - x'|\,.\end{aligned}$$

Setting $\bar{\phi}(y) = \bar{\omega} \vee \bar{r} + (1+\bar{\alpha})x_1 + \left(\bar{b} \vee (1+\underline{\omega}^{-1})\right) y$, $H(x) = x$ and $C = 0$ then yield Conditions (B-4)(v), (B-4)(vi) and (B-4)(vii). Now (B-4)(viii) follows from

$$\int \ln^+ y\ G^\theta(x, \mathrm{d}y) \le \int y\ G^\theta(x, \mathrm{d}y) = rx \le \overline{r}\bar{V}(x)\,.$$

This concludes the proof. □

4.2 NM-GARCH model

The NM(d)-GARCH(1,1) of Example 2 is a specific case of Definition 1 where $\mathsf{X} = \mathbb{R}_+^d$ and ν is the Lebesgue measure on $\mathsf{Y} = \mathbb{R}$,

$$\psi_y^\theta(\mathbf{x}) = \boldsymbol{\omega} + \mathbf{A}\mathbf{x} + y^2\mathbf{b}\,, \tag{29}$$

$$g^\theta(\mathbf{x};y) = \sum_{\ell=1}^d \gamma_\ell \frac{\mathrm{e}^{-y^2/2x_\ell}}{(2\pi x_\ell)^{1/2}}\,, \quad (\mathbf{x}, y) \in \mathsf{X} \times \mathsf{Y}\,, \tag{30}$$

and $\theta = (\boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{A}, \mathbf{b}) \in \Theta$, a compact subset of $\mathsf{P}_d \times (0,\infty)^d \times \mathbb{R}_+^{d\times d} \times \mathbb{R}_+^d$, with P_d defined by (3).

In [12], it is shown that the equation satisfied by the variance of a univariate NM(d)-GARCH(1,1) process admits a constant solution if and only if

$$|\lambda|_{\max}(\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T) < 1\,, \tag{31}$$

where, for any square matrix $\mathbf{M}$, $|\lambda|_{\max}(\mathbf{M})$ denotes the spectral radius of $\mathbf{M}$. It follows that the existence of a weakly stationary solution implies (31) but it does not say anything about the existence of stationary or weakly stationary solution. The result below shows that (31) is indeed a sufficient condition for the existence of a stationary solution with finite variance. It moreover provides with Theorem 3 and Remark 3 the consistency of the MLE $\hat{\theta}_{\mathbf{x}_1,n}$ for any $\mathbf{x}_1 \in \mathsf{X}$.

Theorem 7. *Suppose that all $\theta = (\boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{A}, \mathbf{b})$ in Θ satisfy Condition (31). Then Assumptions (A-1), (A-2), (B-2), (B-3) and (B-4) hold with $\bar{V}$ being defined as any norm on X.*

Proof. In this proof section, we set

$$\bar{V}(\mathbf{x}) = |\mathbf{x}| = \sum_{\ell=1}^{d} |x_\ell| \,, \tag{32}$$

for all $\mathbf{x} = (x_\ell) \in \mathsf{X}$. As in Theorem 6, we divide the proof into two steps.
Step 1. We first show that Assumptions (A-1) and (A-2) hold with the above $\bar{V}$ by applying Theorem 4. Define V on X by setting

$$V(\mathbf{x}) = (\mathbf{1} + \mathbf{x}_0)^T \mathbf{x} \,,$$

where $\mathbf{1}$ is the vector of X with all entries equal to 1 and $\mathbf{x}_0$ is defined by

$$\mathbf{1} + \mathbf{x}_0 = (\mathbf{I} - \left(\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T\right)^T)^{-1}\mathbf{1} \,.$$

We indeed note that by Condition (31) the above inversion is well defined and moreover

$$(\mathbf{I} - (\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T)^T)^{-1} = \mathbf{I} + \sum_{k \geq 1} \left(\mathbf{A}^T + \boldsymbol{\gamma}\mathbf{b}^T\right)^k \,,$$

and, since $\mathbf{A}$, $\mathbf{b}$, $\boldsymbol{\gamma}$ all have non-negative entries, it follows that $\mathbf{x}_0$ has non-negative entries. Thus, for all $\mathbf{x} = (x_\ell) \in \mathsf{X}$,

$$\bar{V}(\mathbf{x}) = \mathbf{1}^T \mathbf{x} \leq V(\mathbf{x}) \,,$$

so that $\bar{V} \lesssim V$. Hence by Theorem 4, we thus only need to check (A-3), (A-4), (A-5), (A-6), (A-7) and (A-8) with V defined as above for a given $\theta = (\boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{A}, \mathbf{b}) \in \Theta$ (so we drop θ in the notation in the remaining of **Step 1**). Condition (A-3) holds for any metric d associated to a norm on the finite dimensional space X. (The precise choice of d is postponed to the verification of (A-7).) We have

$$\begin{aligned}
RV(\mathbf{x}) &= \int V(\boldsymbol{\omega} + \mathbf{A}\mathbf{x} + y^2\mathbf{b})\, G(\mathbf{x}, \mathrm{d}y) \\
&= (\mathbf{1} + \mathbf{x}_0)^T \boldsymbol{\omega} + (\mathbf{1} + \mathbf{x}_0)^T \left(\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T\right) \mathbf{x} \\
&= V(\boldsymbol{\omega}) + \mathbf{1}^T (\mathbf{I} - \left(\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T\right))^{-1} \left(\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T - \mathbf{I} + \mathbf{I}\right) \mathbf{x} \\
&= V(\boldsymbol{\omega}) + \mathbf{x}_0^T \mathbf{x} \\
&\leq V(\boldsymbol{\omega}) + \lambda V(\mathbf{x}) \,,
\end{aligned}$$

where we set $\lambda = \max_\ell \{\mathbf{x}_{0,\ell}/(1 + \mathbf{x}_{0,\ell})\} < 1$. Hence (A-4) holds. Condition (A-5) easily follows from the continuity of the Gaussian distribution with

respect to its variance parameter. We now prove (A-6). From (9) and (29), we have for all $n \geq 1$, $y_{0:n-1} \in \mathsf{Y}^n$ and $\mathbf{x} \in \mathsf{X}$,

$$\psi^\theta \langle y_{0:n-1} \rangle(\mathbf{x}) = \mathbf{A}^n \mathbf{x} + \sum_{j=0}^{n-1} \mathbf{A}^j (\boldsymbol{\omega} + y_{n-1-j}^2 \mathbf{b}) \; . \tag{33}$$

Let us use the norm

$$\|\mathbf{M}\| = \max_j \sum_i |\mathbf{M}_{i,j}| = \sup_{|\mathbf{x}| \leq 1} |\mathbf{M}\mathbf{x}|$$

on $d \times d$ matrices. Note that by (31), there exists $\delta \in (0,1)$ and $c > 0$ such that, for any $k \geq 1$,

$$\left\| \left(\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T \right)^k \right\| \leq c \, \delta^k \; . \tag{34}$$

Using that $\mathbf{A}$, $\mathbf{b}$, $\boldsymbol{\gamma}$ all have nonnegative entries, we have

$$\left\| \mathbf{A}^k \right\| \leq \left\| \left(\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T \right)^k \right\| \; . \tag{35}$$

Hence $(\mathbf{I} - \mathbf{A})^{-1} = \mathbf{I} + \sum_{k \geq 1} \mathbf{A}^k$ is well defined and we set $\mathbf{x}_\infty = (\mathbf{I} - \mathbf{A})^{-1} \boldsymbol{\omega}$ so that, with (29), we have

$$\psi^\theta \langle y_{0:n-1} \rangle(\mathbf{x}) - \mathbf{x}_\infty = \mathbf{A}^n \mathbf{x} + \sum_{j \geq n} \mathbf{A}^j \boldsymbol{\omega} + \sum_{j=0}^{n-1} y_{n-1-j}^2 \mathbf{A}^j \mathbf{b} \; .$$

Then, using definition (23), we get that, $\bar{\mathbb{P}}_{\mathbf{x}}$-a.s., for all $n \geq 1$,

$$\begin{aligned} |\mathbf{X}_n - \mathbf{x}_\infty| &= |\psi \langle Y_{0:n-1} \rangle(\mathbf{x}) - \mathbf{x}_\infty| \\ &\leq |\mathbf{A}^n (\mathbf{x} - \mathbf{x}_\infty)| + \sum_{j \geq n} \left| \mathbf{A}^j \boldsymbol{\omega} \right| + \left(\max_{0 \leq j \leq n-1} Y_j^2 \right) \sum_{j=0}^{n-1} \left| \mathbf{A}^j \mathbf{b} \right| \; . \end{aligned}$$

With (34) and (35), this implies

$$\bar{\mathbb{P}}_{\mathbf{x}} \left(|\mathbf{X}_n - \mathbf{x}_\infty| \leq c \left[\delta^n \left(|\mathbf{x} - \mathbf{x}_\infty| + \frac{|\boldsymbol{\omega}|}{1 - \delta} \right) + \frac{|\mathbf{b}|}{1 - \delta} \max_{0 \leq j \leq n-1} Y_j^2 \right] \right) = 1 \; .$$

To obtain (A-6), it is sufficient to observe that, since g takes positive values in (30), for any positive ϵ, $\mathbf{x} \in \mathsf{X}$ and any $n \geq 1$,

$$\bar{\mathbb{P}}_{\mathbf{x}} \left(\max_{0 \leq j \leq n-1} Y_j^2 < \epsilon \right) > 0 \; .$$

Next we prove (A-7). We have

$$\psi_y(\mathbf{x}) - \psi_y(\mathbf{x}') = \mathbf{A}(\mathbf{x} - \mathbf{x}') \; .$$

Since (34) and (35) imply that $|\lambda|_{\max}(\mathbf{A}) < 1$, there exists a vector norm which makes $\mathbf{A}$ strictly contracting. Choosing the metric d on X as the one derived from this norm, we get (A-7). To show (A-8), we again rely on Lemma 5. Let us set $\mathsf{C} = (0, \infty)$ and $\mathsf{S} = \{1, \dots, d\}$ and define the probability measure μ on S by $\mu(\{s\}) = \gamma_s$, for all $s \in \mathsf{S}$. For all $(t, y) \in \mathsf{C} \times \mathsf{Y}$, let $j(t) = \frac{1}{(2\pi t)^{1/2}}$ and $h(t; y) = \exp\left(-y^2/2t\right)$. Obviously, Relation (24) holds and h satisfies (F-1). Hence, Lemma 5 implies that α and ϕ defined respectively for all $\mathbf{x} = (x_1, \dots, x_d),\ \mathbf{x}' = (x'_1, \dots, x'_d) \in \mathsf{X}$ by

$$\alpha(\mathbf{x}, \mathbf{x}') = \min_{1 \le \ell \le d} \left\{ \left(\frac{x_\ell \wedge x'_\ell}{x_\ell \vee x'_\ell} \right)^{\frac{1}{2}} \right\} \in (0, 1] \text{ and } \phi(\mathbf{x}, \mathbf{x}') = (x_1 \wedge x'_1, \dots, x_d \wedge x'_d),$$

satisfy (A-8)(i). For $\mathbf{x} = (x_1, \dots, x_d),\ \mathbf{x}' = (x'_1, \dots, x'_d) \in \mathsf{X}$, we have

$$\begin{aligned} 1 - \alpha(\mathbf{x}, \mathbf{x}') &= 1 - \min_{1 \le \ell \le d} \left\{ \left(1 - \frac{|x_\ell - x'_\ell|}{x_\ell \vee x'_\ell} \right)^{\frac{1}{2}} \right\} \\ &\le \max_{1 \le \ell \le d} \left\{ \frac{|x_\ell - x'_\ell|}{x_\ell \vee x'_\ell} \right\} \\ &\le \min_{1 \le \ell \le d} (x_\ell^{-1} \wedge x_\ell'^{-1})\ |\mathbf{x} - \mathbf{x}'| \\ &\le W(\mathbf{x}, \mathbf{x}')\ \mathrm{d}(\mathbf{x}, \mathbf{x}')\ , \end{aligned}$$

where d is the metric previously defined and W is defined by $W(\mathbf{x}, \mathbf{x}') = 1 \vee \left(c_{\mathrm{d}} \min_{1 \le \ell \le d} (x_\ell^{-1} \wedge x_\ell'^{-1})\right)$ with $c_{\mathrm{d}} > 0$ is conveniently chosen (such a constant exists since d is the metric associated to a norm and X has finite dimension). Then (A-8)(ii) and (A-8)(iii) hold and, since for all $y \in \mathsf{Y}$ and $x \in \mathsf{X}$, $\psi_y(\mathbf{x})$ has all its entries bounded from below by the positive entries of $\boldsymbol{\omega}$, $W(\psi_y(\mathbf{x}), \psi_y(\mathbf{x}'))$ is uniformly bounded over $(\mathbf{x}, \mathbf{x}', y) \in \mathsf{X} \times \mathsf{X} \times \mathsf{Y}$ and (A-8)(iv) holds. This completes **Step 1**.

Step 2 We now show that Assumptions (B-2), (B-3) and (B-4) hold.

Clearly, (B-2) and (B-3) hold by definitions of $\psi_y^\theta(\mathbf{x})$ and $g^\theta(\mathbf{x}; y)$. It remains to show (B-4). Since Θ is compact, then

$$\underline{\omega} \le \min_{1 \le \ell \le d} \boldsymbol{\omega}_\ell,\ |\boldsymbol{\omega}| \le \overline{\omega},\ \underline{b} \le |\mathbf{b}| \le \bar{b},\ |\lambda|_{\max}(\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T) \le \bar{\rho},\ \left\|\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T\right\| \le L$$

for some $(\underline{\omega}, \overline{\omega}, \underline{b}, \bar{b}, \bar{\rho}) \in (0, \infty)^4 \times (0, 1)$ and $L > 0$. By [16, Lemma 12], we note that this implies that, for all $\bar{\delta} \in (\bar{\rho}, 1)$, there exists $\bar{C} > 0$ such that for all $k \ge 1$ and all $\theta \in \Theta$,

$$\left\| \left(\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T\right)^k \right\| \le \bar{C}\, \bar{\delta}^k\ . \tag{36}$$

We set $\mathsf{X}_1 = [\underline{\omega}, \infty)^d \subset \mathsf{X}$ so that (B-4)(i) holds. Moreover, for all $(\theta, \mathbf{x}, y) \in \Theta \times \mathsf{X}_1 \times \mathsf{Y}$, $g^\theta(\mathbf{x}; y) \le (2\pi\underline{\omega})^{-1/2}$. Thus, Condition (B-4)(ii) holds. Now let

$x_1 \in \mathsf{X}$. Using (33), (36) and (35), we have, for all $\mathbf{x} \in \mathsf{X}$, $y_{1:n} \in \mathsf{Y}^n$ and $\theta \in \Theta$,

$$\begin{aligned}\left|\psi^\theta\langle y_{1:n}\rangle(\mathbf{x}_1) - \psi^\theta\langle y_{1:n}\rangle(\mathbf{x})\right| &= |\mathbf{A}^n(\mathbf{x}_1 - \mathbf{x})| \\ &\leq \bar{C}\,\bar{\delta}^n\,|\mathbf{x}_1 - \mathbf{x}|\ .\end{aligned}$$

Using that the norm defining d is equivalent to the norm $|\cdot|$, we get (B-4)(iii) with

$$\bar{\psi}(\mathbf{x}) = \bar{C}'\,|\mathbf{x}_1 - \mathbf{x}|\ ,$$

for some positive constant $\bar{C}'$. Hence (B-4)(iv) holds and since

$$\left|\mathbf{x}_1 - \psi_y^\theta(\mathbf{x}_1)\right| \leq (L+1)\,|\mathbf{x}_1| + \overline{\omega} + y^2\bar{b}\ ,$$

we also get (B-4)(v) provided that

$$\bar{\phi}(y) \geq (L+1)\,|\mathbf{x}_1| + \overline{\omega} + y^2\bar{b}\ . \tag{37}$$

It is straightforward to show that, for all $\theta \in \Theta$, $\mathbf{x} \in \mathsf{X}_1$, $y \in \mathbb{R}$, and $\ell \in \{1, \ldots, d\}$,

$$\left|\frac{\partial \ln g^\theta}{\partial x_\ell}(\mathbf{x}; y)\right| \leq \frac{1}{2}\left(\frac{y^2}{\underline{\omega}^2} + \frac{1}{\underline{\omega}}\right)\ .$$

Thus, by the mean value theorem, for all $\theta \in \Theta$, $(\mathbf{x}, \mathbf{x}') \in \mathsf{X}_1 \times \mathsf{X}_1$ and $y \in \mathsf{Y}$,

$$\left|\ln g^\theta(\mathbf{x}; y) - \ln g^\theta(\mathbf{x}'; y)\right| \leq \frac{1}{2}\left(\frac{y^2}{\underline{\omega}^2} + \frac{1}{\underline{\omega}}\right)|\mathbf{x} - \mathbf{x}'|\ .$$

We thus obtain (B-4)(v), (B-4)(vi) and (B-4)(vii) by setting $C = 0$,

$$H(u) = \sup_{d(\mathbf{x},\mathbf{x}') \leq u} |\mathbf{x} - \mathbf{x}'|\ ,$$

and

$$\bar{\phi}(y) = (L+1)\,|\mathbf{x}_1| + \overline{\omega} + 1/(2\underline{\omega}) + y^2(\bar{b} + \underline{\omega}^2)\ .$$

In addition, for all $\theta \in \Theta$ and $\mathbf{x} \in \mathsf{X}$, we have

$$\int y^2 G^\theta(x, \mathrm{d}y) = \boldsymbol{\gamma}^T\mathbf{x}\ .$$

Hence, using (32) with the above definitions, we obtain (B-4)(viii) and the proof is concluded. □

4.3 The Threshold INGARCH model

The threshold INGARCH(1, 1) in Example 3 is a specific case of Definition 1 where ν is the counting measure on $\mathsf{Y} = \mathbb{Z}_+$,

$$\psi_y^\theta(x) = \omega + ax + by \,, \tag{38}$$

$$g^\theta(x; y) = \mathrm{e}^{-(x\wedge\tau)} \frac{(x \wedge \tau)^y}{y!} \,, \tag{39}$$

with $\theta = (\omega, a, b, \tau)$ in a compact subset Θ of $(0, \infty)^4$ and $\mathsf{X} = (0, \infty)$. In this model, if $a < 1$, we then have the ergodicity and consistency results as stated in Theorem 8 below.

Theorem 8. *Suppose that all $\theta = (\omega, a, b, \tau)$ in Θ satisfy $a < 1$. Then Assumptions (A-1), (A-2), (B-2), (B-3) and (B-4) hold with $\bar{V}$ being defined as the identity function on X and with any $x_1 \in \mathsf{X}$.*

Proof. As in the proofs of the two theorems above, for convenience, we divide the proof into two steps.

Step 1. We first prove Assumptions (A-1) and (A-2) by applying Theorem 4. We set $\bar{V}(x) = V(x) = x$ and thus we only need to check (A-3), (A-4), (A-5), (A-6), (A-7) and (A-8). Condition (A-3) holds with the usual metric on $\mathbb{R}$. We have for all $\theta \in \Theta$,

$$RV(x) = \omega + ax + b(x \wedge \tau) \le aV(x) + (\omega + b\tau),$$

which yields (A-4). The fact that the kernel R is weak Feller easily follows by observing that, as $x \to x'$, $\mathcal{P}(x)$ converges weakly to $\mathcal{P}(x')$ and the map $x \mapsto x \wedge \tau$ is continuous, so (A-5) holds.

The proof of (A-6) is similar to the NBIN-GARCH case of Theorem 6 and is thus omitted. Assumption (A-7) holds since we have for all $(x, x', y) \in \mathsf{X}^2 \times \mathsf{Y}$ with $x \neq x'$,

$$\frac{|\psi_y(x) - \psi_y(x')|}{|x - x'|} = a < 1 \,.$$

To prove (A-8), we apply Lemma 5 with $\mathsf{C} = \mathsf{X}$, $\mathsf{S} = \{1\}$ (so μ boils down to the Dirac measure on $\{1\}$). For all $(x, y) \in \mathsf{X} \times \mathsf{Y}$, let $j(x) = \mathrm{e}^{-(x\wedge\tau)}$ and $h(x; y) = \frac{(x\wedge\tau)^y}{y!}$. Then h indeed satisfies (F-1). Thus by Lemma 5, for all $(x, x') \in \mathsf{X}^2$ and $y \in \mathsf{Y}$, we get that

$$\alpha(x, x') = \mathrm{e}^{-(x\vee x')\wedge\tau + (x\wedge x')\wedge\tau} \in (0, 1] \quad \text{and} \quad \phi(x, x') = x \wedge x'$$

satisfy (A-8)(i).

Let $W(x, x') = 1$ for all $(x, x') \in \mathsf{X}^2$, which is a constant function. Thus (A-8)(ii) and (A-8)(iv) clearly hold. Moreover, (A-8)(iii) holds since for all $(x, x') \in \mathsf{X}^2$, we have that

$$1 - \alpha(x, x') \le x \vee x' - x \wedge x' = |x - x'| = W(x, x')|x - x'| \,.$$

Therefore, (A-8) holds, which completes **Step 1**.
Step 2. We now prove (B-2), (B-3) and (B-4). By assumption on Θ, then there exists $(\underline{\omega}, \bar{\omega}, \underline{b}, \bar{b}, \underline{\tau}, \bar{\tau}, \underline{\alpha}, \bar{\alpha}) \in (0,\infty)^6 \times (0,1)^2$ such that

$$\underline{\omega} \le \omega \le \bar{\omega},\ \le b \le \bar{b},\ \underline{\tau} \le \tau \le \bar{r},\ \underline{\alpha} \le a \le \bar{\alpha}\ .$$

Clearly, (B-2) and (B-3) hold by definitions of $\psi_y^\theta(x)$ and $g^\theta(x;y)$. It remains to check (B-4) for a well-chosen closed subset X_1 and any $x_1 \in \mathsf{X}$. Let $\mathsf{X}_1 = [\underline{\omega}, \infty) \subset \mathsf{X}$ so that (B-4)(i) holds. By noting that for all $(\theta, x, y) \in \Theta \times \mathsf{X} \times \mathsf{Y}$, $g^\theta(x;y) \le 1$, we have (B-4)(ii). From (9) and (38), we have for all $s \le t$, $y_{s:t} \in \mathsf{Y}^{t-s+1}$, $x \in \mathsf{X}$ and $\theta \in \Theta$,

$$\psi^\theta\langle y_{s:t}\rangle(x) = \omega\left(\frac{1-a^{t-s+1}}{1-a}\right) + a^{t-s+1}x + b\sum_{j=0}^{t-s} a^j y_{t-j}\ . \tag{40}$$

Using (40), we have, for all $\theta \in \Theta$, $x \in \mathsf{X}$ and $y_{1:n} \in \mathsf{Y}^n$,

$$\left|\psi^\theta\langle y_{1:n}\rangle(x_1) - \psi^\theta\langle y_{1:n}\rangle(x)\right| = a^n\,|x_1 - x| \le \bar{\alpha}^n\,|x_1 - x|\ .$$

This gives (B-4)(iii) and (B-4)(iv) by setting $\varrho = \bar{\alpha} < 1$ and $\bar{\psi}(x) = |x_1 - x|$. Next we set $\bar{\phi}$, H and C to meet Conditions (B-4)(v) and (B-4)(vi) and (B-4)(vii). Let us write, for all $\theta \in \Theta$ and $y \in \mathsf{Y}$,

$$\left|x_1 - \psi_y^\theta(x_1)\right| \le \omega + (1+a)x_1 + by \le \bar{\omega} + (1+\bar{\alpha})x_1 + \bar{b}y$$

and, for all $(x, x') \in \mathsf{X}_1^2 = [\underline{\omega}, \infty)^2$,

$$\begin{aligned}\left|\ln g^\theta(x;y) - \ln g^\theta(x';y)\right| &= \left|\left(x' \wedge \tau - x \wedge \tau\right) + y\left(\ln(x \wedge \tau) - \ln(x' \wedge \tau)\right)\right| \\ &\le \left(1 + (\underline{\omega} \wedge \underline{\tau})^{-1} y\right)\ |x - x'|\ .\end{aligned}$$

Setting $\bar{\phi}(y) = 1 + \bar{\omega} + (1+\bar{\alpha})x_1 + \left(\bar{b} \vee (\underline{\omega} \wedge \underline{\tau})^{-1}\right) y$, $H(x) = x$ and $C = 0$ then yield Conditions (B-4)(v), (B-4)(vi) and (B-4)(vii). Now (B-4)(viii) follows from

$$\int \ln^+ y\ G^\theta(x, \mathrm{d}y) \le \int y\ G^\theta(x, \mathrm{d}y) = x \wedge \tau \le \bar{V}(x)\ .$$

This concludes the proof. □

5 Numerical experiments

5.1 Numerical procedure

In this part we provide a numerical method for computing the (conditional) MLE $\hat{\theta}_{x,n}$ for the parameter $\theta = (\omega, a, b, r)$ in the NBIN-GARCH(1, 1) model introduced in Example 1 and studied in Section 4.1. It is convenient to write

$\theta = (\vartheta, r)$ with $\vartheta = (\omega, a, b)$ and then to write $\psi_y^\vartheta(x)$ and $g^r(x;y)$ instead of $\psi_y^\theta(x)$ and $g^\theta(x;y)$ in (25) and (26), respectively. In contrast to the approach used in [20], we allow the component r to be any positive real number, rather than a discrete one and to be unknown as well. We thus maximize jointly with respect to the parameters ϑ and r the log-likelihood function $\mathsf{L}_{x,n}^\theta\langle y_{1:n}\rangle = \mathsf{L}_{x,n}^{(\vartheta,r)}\langle y_{1:n}\rangle$. In practice, one does not rely on a compact set Θ of parameters as in Theorem 6. Instead the maximization is performed over all parameters $\omega > 0$, $a > 0$, $b > 0$, $r > 0$ such that the stability constraint $a + br < 1$ holds (taken from (27)). We use the constrained nonlinear optimization function `auglag` (Augmented Lagrangian Minimization Algorithm) from the package `alabama` (Augmented Lagrangian Adaptive Barrier Minimization Algorithm) in `R`. For this purpose we provide an initial parameter point and a numerical computation of the normalized log-likelihood function $\mathsf{L}_{x,n}^\theta\langle y_{1:n}\rangle$ and of its gradient. The initial point is obtained by applying a conditional least square (CLS) estimation based on an ARMA$(1,1)$ representation of the model, see [20, Section 3]. The computation of the log-likelihood and of its derivatives are derived as follows. For all $x \in \mathsf{X}$, denoting $u_k^\vartheta = \psi^\vartheta\langle y_{1:k-1}\rangle(x)$ for all $k \geq 2$ and $u_1^\vartheta = x$, we have

$$\begin{aligned}\mathsf{L}_{x,n}^{(\vartheta,r)}\langle y_{1:n}\rangle &= n^{-1}\sum_{k=1}^n \ln g^r\left(\psi^\vartheta\langle y_{1:k-1}\rangle(x); y_k\right)\\ &= n^{-1}\ln g^r(x, y_1) + n^{-1}\sum_{k=2}^n \ln g^r\left(u_k^\vartheta; y_k\right) .\end{aligned}$$

The computation of u_k^ϑ for all $k \geq 2$ is done iteratively by observing that $u_k^\vartheta = \psi^\vartheta\langle y_{k-1}\rangle(u_{k-1}^\vartheta)$ and the computation of $\mathsf{L}_{x,n}^{(\vartheta,r)}\langle y_{1:n}\rangle$ is deduced. The computation of the derivatives with respect to parameter $\theta = (\vartheta, r)$ of the function $\mathsf{L}_{x,n}^{(\vartheta,r)}\langle y_{1:n}\rangle$ are then obtained in two steps. First, for $k \geq 2$, the derivative of u_k^ϑ with respect to ϑ are obtained iteratively by $\partial u_1^\vartheta/\partial\vartheta = 0$ and

$$\frac{\partial u_k^\vartheta}{\partial\vartheta} = (1, u_{k-1}^\vartheta, Y_{k-1}) + a\frac{\partial u_{k-1}^\vartheta}{\partial\vartheta} .$$

Then the derivatives of $\mathsf{L}_{x,n}^{(\vartheta,r)}\langle y_{1:n}\rangle$ with respect to ϑ and r are given by

$$\frac{\partial \mathsf{L}_{x,n}^{(\vartheta,r)}}{\partial\vartheta} = n^{-1}\sum_{k=1}^n \frac{\partial \ln g^r}{\partial x}\left(u_k^\vartheta; y_k\right)\frac{\partial u_k^\vartheta}{\partial\vartheta} = n^{-1}\sum_{k=2}^n \left(\frac{y_k}{u_k^\vartheta} - \frac{y_k + r}{1 + u_k^\vartheta}\right)\frac{\partial u_k^\vartheta}{\partial\vartheta}$$

and

$$\begin{aligned}\frac{\partial \mathsf{L}_{x,n}^{(\vartheta,r)}}{\partial r} &= n^{-1}\sum_{k=1}^n \frac{\partial \ln g^r}{\partial r}\left(u_k^\vartheta; y_k\right)\\ &= n^{-1}\sum_{k=1}^n \left(\Gamma_2(r + y_k) - \ln(1 + u_k^\vartheta)\right) - \Gamma_2(r) ,\end{aligned}$$

respectively, where Γ_2 is the digamma function $\Gamma_2(r) = \frac{\mathrm{d}}{\mathrm{d}r}\ln \circ \Gamma(r)$, $r > 0$.

5.2 Simulation study

We consider two NBIN-GARCH(1, 1) models with parameters:

(M.1) $\theta_\star = (\omega_\star, a_\star, b_\star, r_\star) = (3,\ .2,\ .2,\ 2)$ and

(M.2) $\theta_\star = (\omega_\star, a_\star, b_\star, r_\star) = (3,\ .35,\ .1,\ 1.5)$.

We simulated $m = 200$ data sets for each sample size $n = 2^7, 2^8, 2^9$ and 2^{10}. In Figure 2, we display the obtained boxplots of the difference of the normalized log-likelihood function evaluated respectively at MLE and at the true value $\theta_\star$. As predicted by the theory, this difference appears to converge to 0 as the number of observations $n \to \infty$. For the NBIN-GARCH(1, 1) model, it can be shown that $\Theta_\star = \{\theta_\star\}$, which implies the convergence of the MLE to the true parameter. We can observe this behavior for each component of the MLE for the two models in Figure 3 and Figure 4. We also report the Monte Carlo mean along with the mean absolute deviation error (MADE): MADE $= m^{-1}\sum_{j=1}^{m} |\hat{\theta}_{x,n}^{j} - {\theta_\star}^{j}|$ as an evaluation criterion for the estimated parameter in Table 1.

Table 1: Mean of estimates, MADEs (within parentheses) for the NBIN-GARCH(1, 1) models

Model	Parameter	Sample size n			
		$n = 2^7$	$n = 2^8$	$n = 2^9$	$n = 2^{10}$
(M.1)	$\hat{\omega}$	3.311(.973)	3.212(.719)	3.108(.507)	3.062(.372)
	$\hat{a}$	.165(.138)	.173(.113)	.187(.076)	.193(.055)
	$\hat{b}$	.194(.049)	.195(.034)	.197(.025)	.200(.018)
	$\hat{r}$	2.045(.241)	2.035(.166)	2.020(.112)	2.011(.074)
(M.2)	$\hat{\omega}$	3.525(1.325)	3.362(1.258)	3.326(1.041)	3.167(.761)
	$\hat{a}$	.252(.227)	.290(.213)	.296(.170)	.319(.136)
	$\hat{b}$	.092(.056)	.097(.039)	.098(.028)	.100(.022)
	$\hat{r}$	1.563(.175)	1.539(.129)	1.520(.093)	1.513(.066)

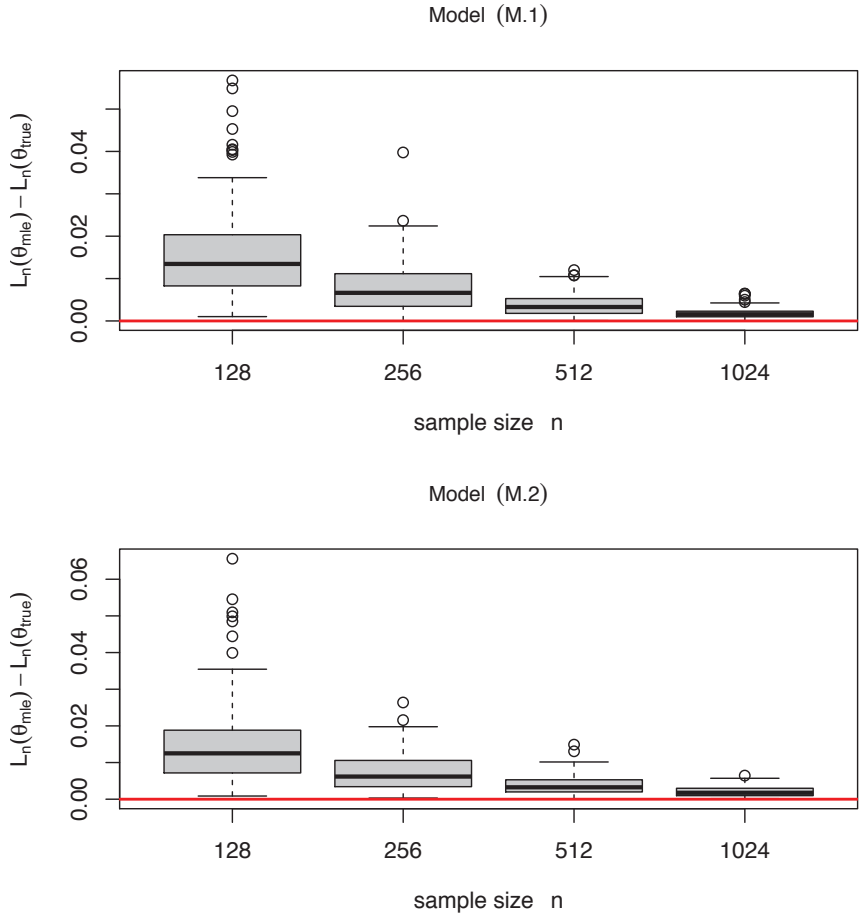

Figure 2: Boxplots of the differences of log-likelihood functions evaluated at the estimated MLE and the true value for Models (M.1) and (M.2) with sample sizes $n = 2^7, 2^8, 2^9$ and $n = 2^{10}$, respectively. The red "continuous" line indicates the position of zero.

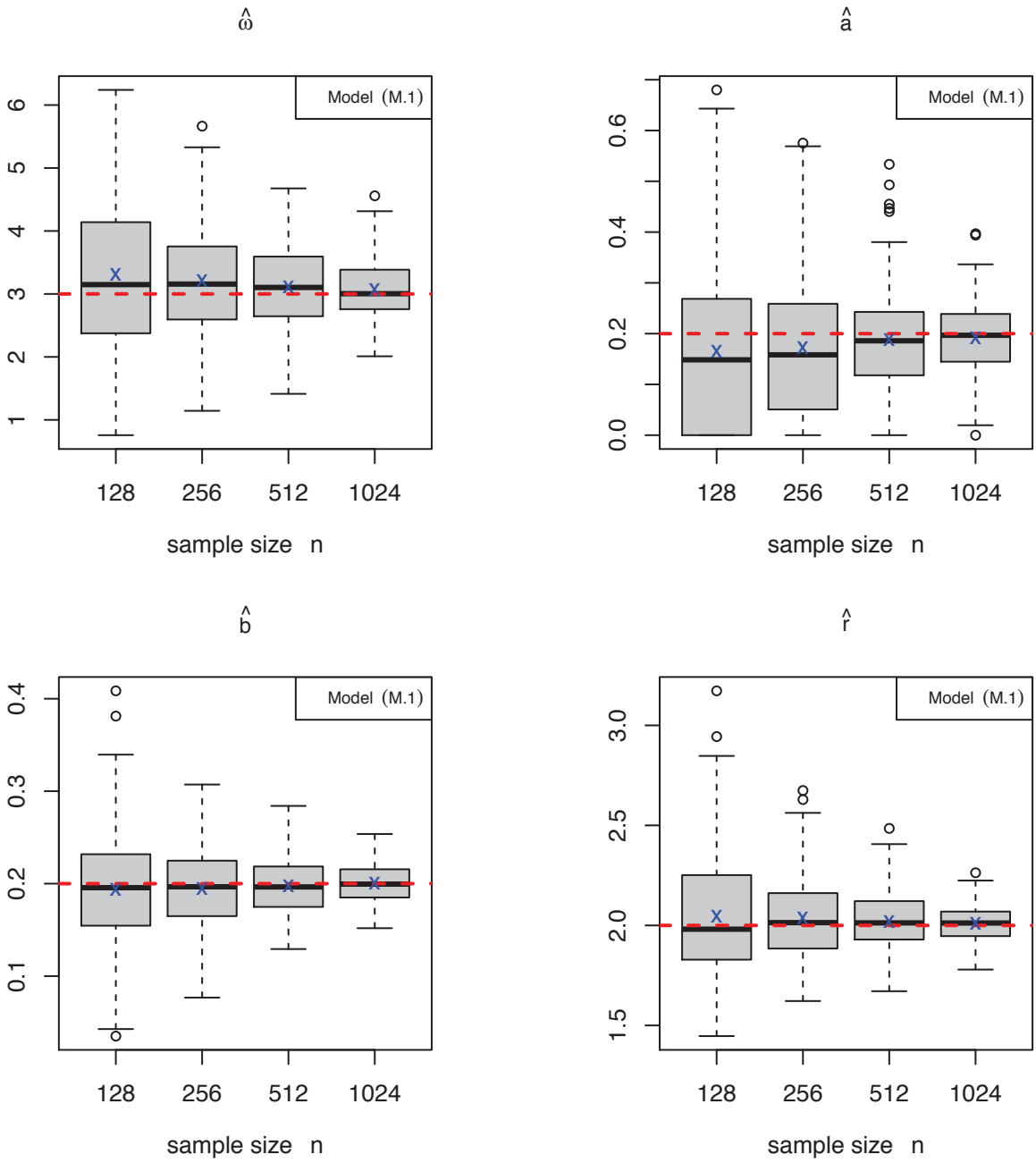

Figure 3: Boxplots of the estimated MLE for Model (M.1) with sample sizes $n = 2^7, 2^8, 2^9$ and $n = 2^{10}$, respectively. The red "dashed" line indicates the true value of the parameter and the blue "x" indicates the location of the Monte Carlo mean of the MLE.

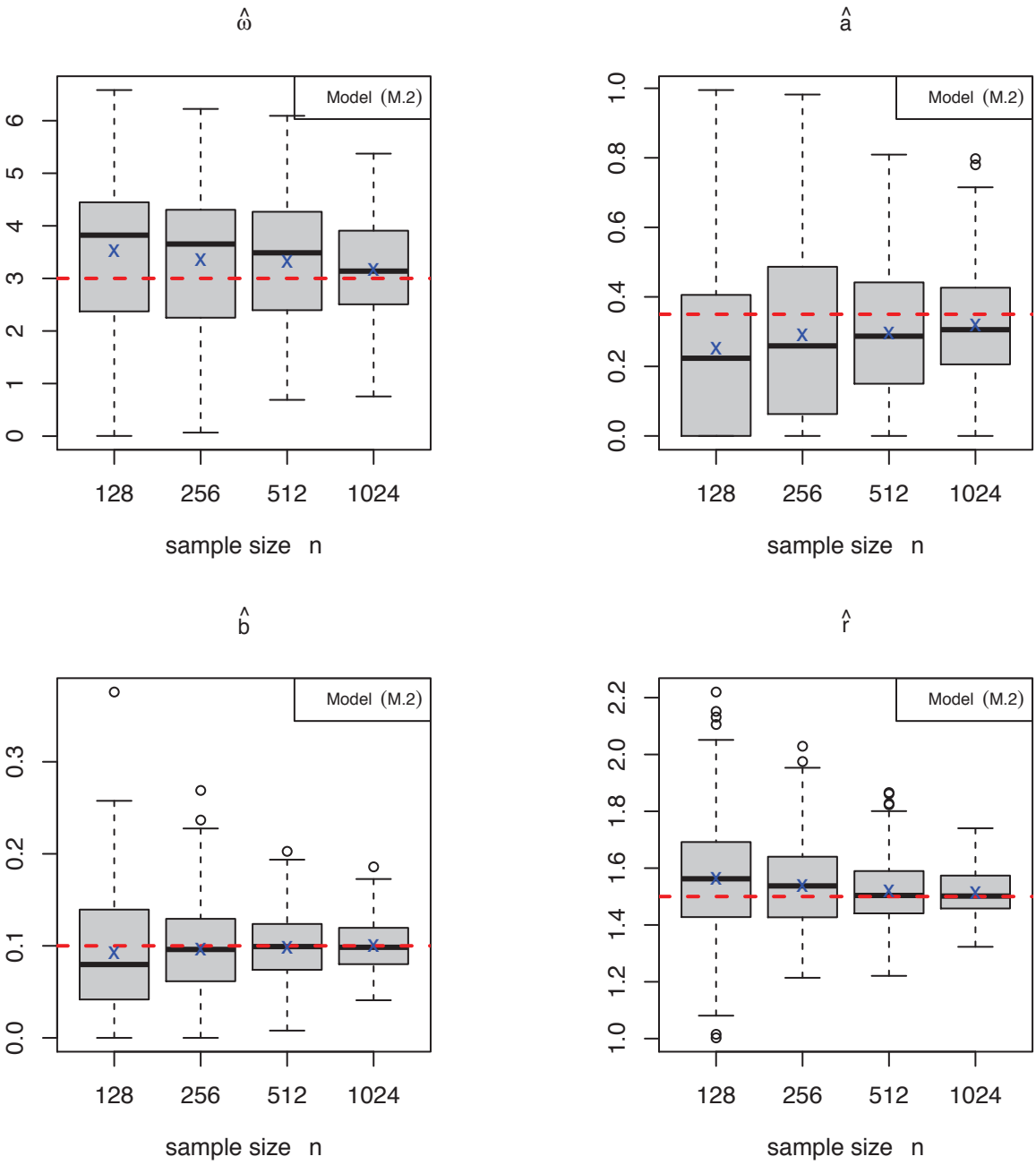

Figure 4: Same as Figure 3 but for Model (M.2).

6 Postponed proofs

6.1 Convergence of the MLE

Assumptions (A-1) and (A-2) are supposed to hold throughout this section. The proof of Theorem 3 relies on the approach introduced in [18], which was already used in [7] for a restricted class of observation-driven models. Our main contribution here is to provide the handy conditions listed in Assumption (B-4). We first show that our conditions imply (B-1) and the following one.

(B-5) There exists $x_1 \in \mathsf{X}$ such that, for all $\theta, \theta_\star \in \Theta$, $p^\theta(Y_1 \,|\, Y_{-\infty:0})$ defined as in (14) with $x = x_1$ is finite $\tilde{\mathbb{P}}^{\theta_\star}$-a.s. Moreover, for all $\theta_\star \in \Theta$, we have

$$\lim_{k\to\infty} \sup_{\theta\in\Theta} \left| \ln \frac{g^\theta(\psi^\theta\langle Y_{1:k-1}\rangle(x_1); Y_k)}{p^\theta(Y_k \,|\, Y_{-\infty:k-1})} \right| = 0 \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{41}$$

Indeed we have the following lemma.

Lemma 9. *Assumptions (B-2), (B-3) and (B-4) imply (B-5) and (B-1).*

Proof. See 6.3. □

Now the proof of Theorem 3 directly follows from the following lemma.

Lemma 10. *Assume that (B-2), (B-3) and (B-4)(i)–(ii) hold and that x_1 satisfies (B-5). Then $\Theta_\star$ defined by (15) is a non-empty closed subset of Θ and (16) holds.*

Proof. By [7, Theorem 33], to obtain (16), it is sufficient to show that, for all $\theta_\star \in \Theta$, the two following assertions hold.

(a) $\tilde{\mathbb{E}}^{\theta_\star}\left[\sup_{\theta\in\Theta} \ln^+ p^\theta(Y_1 \,|\, Y_{-\infty:0})\right] < \infty$,

(b) the function $\theta \mapsto \ln p^\theta(Y_1 \,|\, Y_{-\infty:0})$ is continuous on Θ, $\tilde{\mathbb{P}}^{\theta_\star}$-a.s.

In (B-5), $p^\theta(Y_1 \,|\, Y_{-\infty:0})$ is defined $\tilde{\mathbb{P}}^{\theta_\star}$-a.s. as the limit in (14) with $x = x_1$. So, $\tilde{\mathbb{P}}^{\theta_\star}$-a.s., by (B-4)(i)–(ii), $p^\theta(Y_1 \,|\, Y_{-\infty:0})$ is bounded by the finite constant appearing in (B-4)(ii). Hence Condition (a) holds.

Condition (b) then follows from (41). Since almost sure convergence implies the convergence in probability and $\tilde{\mathbb{P}}^{\theta_\star}$ is shift invariant, the random sequence

$$U_m := \sup_{\theta\in\Theta} \left| \ln \frac{g^\theta(\psi^\theta\langle Y_{-m:0}\rangle(x_1); Y_1)}{p^\theta(Y_1 \,|\, Y_{-\infty:0})} \right| , \quad m \in \mathbb{Z}_+ ,$$

converges to zero in $\tilde{\mathbb{P}}^{\theta_\star}$-probability. Then there exists a subsequence of (U_m) which converges $\tilde{\mathbb{P}}^{\theta_\star}$-a.s. to zero. Hence, interpreting this convergence as a uniform (in θ) convergence of $\ln g^\theta(\psi^\theta\langle Y_{-m:0}\rangle(x_1); Y_1)$ to $\ln p^\theta(Y_1 \,|\, Y_{-\infty:0})$ to conclude that (b) holds, it is sufficient to show that $\theta \mapsto \ln g^\theta(\psi^\theta\langle Y_{-m:0}\rangle(x_1); Y_1)$ is continuous for all m $\tilde{\mathbb{P}}^{\theta_\star}$-a.s. This is indeed the case by (B-2) and (B-3) and since $g^\theta(x; y)$ is positive. □

6.2 Ergodicity

For proving Theorem 4, we first recall a more general set of conditions derived in [7], which are based on the following definition.

Definition 11. Let $\bar{G}$ be a probability kernel from X^2 to $\mathcal{Y}^{\otimes 2} \otimes \mathcal{P}(\{0,1\})$ satisfying the following marginal conditions, for all $(x,x') \in \mathsf{X}^2$ and $B \in \mathcal{Y}$,

$$\begin{cases} \bar{G}((x,x'); B \times \mathsf{Y} \times \{0,1\}) = G(x;B)\,, \\ \bar{G}((x,x'); \mathsf{Y} \times B \times \{0,1\}) = G(x';B)\,, \end{cases} \tag{42}$$

and such that the following coupling condition holds

$$\bar{G}((x,x'); \{(y,y) \ : \ y \in \mathsf{Y}\} \times \{1\}) = \bar{G}((x,x'); \mathsf{Y}^2 \times \{1\})\,. \tag{43}$$

Define the following quantities successively.

- The trace measure of $\bar{G}((x,x');\cdot)$ on the set $\{(y,y) \ : \ y \in \mathsf{Y}\} \times \{1\}$ is denoted by

$$\check{G}((x,x');B) = \bar{G}((x,x'); \{(y,y) \ : \ y \in B\} \times \{1\}), \quad B \in \mathcal{Y}\,. \tag{44}$$

- The probability kernel $\bar{R}$ from $(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$ to $(\mathsf{X}^2 \times \{0,1\}, \mathcal{X}^{\otimes 2} \otimes \mathcal{P}(\{0,1\})$ is defined for all $x,x' \in \mathsf{X}^2$ and $A \in \mathcal{X}^{\otimes 2}$ by

$$\bar{R}((x,x'); A \times \{1\}) = \int_{\mathsf{Y}} 1_A(\psi_y(x), \psi_y(x'))\, \check{G}((x,x'); \mathrm{d}y)\,. \tag{45}$$

- The measurable function α from X^2 to $[0,1]$ is defined by

$$\alpha(x,x') = \bar{R}((x,x'); \mathsf{X}^2 \times \{1\}) = \bar{G}((x,x'); \mathsf{Y}^2 \times \{1\})\,. \tag{46}$$

- The kernel $\hat{R}$ is defined for all $(x,x') \in \mathsf{X}^2$ and $A \in \mathcal{X}^{\otimes 2}$ by

$$\hat{R}((x,x'); A) = \begin{cases} \dfrac{\bar{R}((x,x'); A \times \{1\})}{\alpha(x,x')} & \text{if } \alpha(x,x') > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{47}$$

We can now introduce the so-called *contracting condition* which yields ergodicity.

(A-9) There exists a kernel $\bar{G}$ yielding α and $\hat{R}$ as in Definition 11, a measurable function $W : \mathsf{X}^2 \to [1,\infty)$ satisfying Conditions (A-8)(ii) and (A-8)(iii) and real numbers $(D, \zeta_1, \zeta_2, \rho) \in (\mathbb{R}_+)^3 \times (0,1)$ such that for all $(x,x') \in \mathsf{X}^2$ and, for all $n \geq 1$,

$$\hat{R}^n((x,x'); \mathrm{d}) \leq D\rho^n \mathrm{d}(x,x')\,, \tag{48}$$

$$\hat{R}^n((x,x'); \mathrm{d} \times W) \leq D\rho^n \mathrm{d}^{\zeta_1}(x,x') W^{\zeta_2}(x,x')\,. \tag{49}$$

Under Conditions (A-3), (A-4), (A-5), (A-6) and (A-9) and by combining Theorem 6, Proposition 8 and Lemma 7 in [7], we immediately obtain the following result.

Theorem 12. *Assume (A-3), (A-4), (A-5), (A-6) and (A-9). Then the Markov kernel K admits a unique invariant distribution π and $\pi_1(\bar{V}) < \infty$ for any $\bar{V} : \mathsf{X} \to \mathbb{R}_+$ such that $\bar{V} \lesssim V$.*

Assumptions (A-3), (A-4), (A-5) and (A-6) are quite usual and easy to check. The key point to obtain ergodicity is thus to construct $\bar{G}$ satisfying (A-9). For this, we can also rely on the following result which is quoted from [7, Lemma 9].

Lemma 13. *Assume that there exists $(\rho, \beta) \in (0,1) \times \mathbb{R}$ such that for all $(x, x') \in \mathsf{X}^2$,*

$$\hat{R}\left((x,x'); \left\{(x_1,x_1') \in \mathsf{X}^2 \;:\; \mathrm{d}(x_1,x_1') > \rho\, \mathrm{d}(x,x')\right\}\right) = 0\,, \tag{50}$$

$$\hat{R}W \le W + \beta\,. \tag{51}$$

Then, (48) and (49) hold.

Now we can prove that our set of conditions is sufficient.

Proof of Theorem 4. We only need to show that (A-7) and (A-8) imply (A-9). We preface our proof by the following lemma.

Lemma 14. *Assume (A-8)(i). Then one can define a kernel $\bar{G}$ as in Definition 11 with the same α given in (46). Moreover, the kernel $\hat{R}$ defined by (47) satisfies, for all $(x, x') \in \mathsf{X}^2$ such that $\alpha(x, x') > 0$ and all measurable functions $f : \mathsf{X}^2 \to \mathbb{R}_+$,*

$$\hat{R}((x,x'); f) = G(\phi(x,x'); \tilde{f}) \quad \textit{with} \quad \tilde{f}(y) = f(\psi_y(x), \psi_y(x'))\,. \tag{52}$$

Let us conclude the proof of Theorem 4 before proving this lemma. By Lemma 14 and Lemma 13, it remains to check that (50) and (51) hold for all $(x, x') \in \mathsf{X}^2$. Observe that by definition of $\hat{R}$, Condition (A-8)(iv) is equivalent to

$$\sup_{(x,x') \in \mathsf{X}^2} \left(\hat{R}W(x,x') - W(x,x')\right) < \infty\,.$$

so we can find $\beta \in \mathbb{R}$ such that (51) holds for all $(x, x') \in \mathsf{X}^2$.

Now, let $(x, x') \in \mathsf{X}^2$ and let (X, X') be distributed according to $\hat{R}((x,x');\cdot)$ which is defined in (52). When $x = x'$, then $\mathrm{d}(X, X') = 0$, implying that Condition (50) holds with any nonnegative ρ. For $x \neq x'$, let ρ be defined by

$$\rho = \sup_{\substack{(x,x',y) \in \mathsf{X}^2 \times \mathsf{Y} \\ x \neq x'}} \frac{\mathrm{d}(\psi_y(x), \psi_y(x'))}{\mathrm{d}(x,x')}\,, \tag{53}$$

which is in $(0,1)$ by (A-7). Then

$$\frac{\mathrm{d}(X,X')}{\mathrm{d}(x,x')} = \frac{\mathrm{d}(\psi_Y(x),\psi_Y(x'))}{\mathrm{d}(x,x')} \leq \rho \ .$$

Therefore, Condition (50) holds for all $(x,x') \in \mathsf{X}^2$ with ρ as in (53). □

We conclude this section with the postponed

Proof of Lemma 14. Let $(x,x') \in \mathsf{X}^2$. We define $\bar{G}((x,x');\cdot)$ as the distribution of (Y,Y',U) drawn as follows. We first draw a random variable $\bar{Y}$ taking values in Y with density $g(\phi(x,x');\cdot)$ with respect to ν. Then we define (Y,Y',U) by separating the two cases, $\alpha(x,x') = 1$ and $\alpha(x,x') < 1$.

- Suppose that $\alpha(x,x') = 1$. Then from (A-8)(i), we have

  $$G(x;\cdot) = G(x';\cdot) = G(\phi(x,x');\cdot) \ .$$

  In this case, we set $(Y,Y',U) = (\bar{Y},\bar{Y},1)$.

- Suppose now that $\alpha(x,x') < 1$. Then, using (22), the functions

  $$(1-\alpha(x,x'))^{-1}\left[g(x;\cdot) - \alpha(x,x')g(\phi(x,x');\cdot)\right]$$

  and

  $$(1-\alpha(x,x'))^{-1}\left[g(x';\cdot) - \alpha(x,x')g(\phi(x,x');\cdot)\right] \ ,$$

  are probability density functions with respect to ν and we let Λ and Λ' be two independent random variables taking values in Y drawn with these two density functions, respectively. In this case we draw U independently according to a Bernoulli variable with mean $\alpha(x,x')$ and set

  $$(Y,Y') = \begin{cases} (\bar{Y},\bar{Y}) & \text{if } U = 1 \ , \\ (\Lambda,\Lambda') & \text{if } U = 0 \ . \end{cases}$$

One can easily check that the so defined kernel $\bar{G}$ satisfies (42) and (43). Moreover, for all $(x,x') \in \mathsf{X}^2$,

$$\bar{G}((x,x');\mathsf{Y}^2 \times \{1\}) = \mathbb{P}(U=1) = \alpha(x,x') \ ,$$

which is compatible with (46). The kernel $\hat{R}$ is defined by setting $\hat{R}((x,x');\cdot)$ as the conditional distribution of $(X,X') = (\psi_Y(x),\psi_Y(x'))$ given that $U = 1$. To complete the proof of Lemma 14, observe that for any measurable $f : \mathsf{X}^2 \to \mathbb{R}_+$, we have, for all $(x,x') \in \mathsf{X}^2$ such that $\alpha(x,x') > 0$,

$$\begin{aligned} \hat{R}((x,x');f) &= \mathbb{E}\left[f(\psi_Y(x),\psi_Y(x')) \mid U = 1\right] \\ &= \mathbb{E}\left[f(\psi_{\bar{Y}}(x),\psi_{\bar{Y}}(x'))\right] \\ &= G(\phi(x,x');\tilde{f}) \ , \end{aligned}$$

where $\tilde{f}(y) = f(\psi_y(x),\psi_y(x'))$ for all $y \in \mathsf{Y}$.

□

6.3 Proof of Lemma 9

Under (A-2), Assumptions (B-4)(viii) implies that for all $\theta \in \Theta$,

$$\pi_2^\theta \left(\ln^+(\bar{\phi})\right) < \infty \,, \tag{54}$$

and if moreover $C > 0$,

$$\pi_2^\theta \left(\bar{\phi}\right) < \infty \,. \tag{55}$$

For proving Lemma 9, we will also make use of [7, Lemma 34] which we quote here for convenience.

Lemma 15. *Let $\{U_n\}_{n\in\mathbb{Z}_+}$ be a stationary sequence of real-valued random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that $\mathbb{E}(\ln^+ |U_0|) < \infty$. Then, for all $\eta \in (0,1)$,*

$$\lim_{k\to\infty} \eta^k U_k = 0 \,, \quad \mathbb{P}\text{-a.s.}$$

Proof of Lemma 9. We first show that $p^\theta(y|Y_{-\infty:0})$ in (14) is finite for $x = x_1$ $\tilde{\mathbb{P}}^{\theta_\star}$-a.s. By (B-2), this follows by writing

$$p^\theta \left(y_1 \,|\, y_{-\infty:0}\right) = g^\theta \left(\psi^\theta \langle y_{-\infty:0}\rangle; y_1\right) \,, \tag{56}$$

if, for all $\theta, \theta_\star \in \Theta$, the limit

$$\psi^\theta \langle Y_{-\infty:0}\rangle = \lim_{m\to\infty} \psi^\theta \langle Y_{-m:0}\rangle(x_1) \quad \text{is well defined} \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{57}$$

For all $\theta \in \Theta$, $m \geq 0$, $x \in \mathsf{X}$ and $y_{-m:0} \in \mathsf{Y}^{m+1}$, using (B-4)(iii), we have

$$\mathrm{d}(\psi^\theta \langle y_{-m:0}\rangle(x_1), \psi^\theta \langle y_{-m:0}\rangle(x)) \leq \varrho^{m+1} \, \bar{\psi}(x) \,. \tag{58}$$

Taking $x = \psi^\theta_{y_{-m-1}}(x_1)$ and using (B-4)(v), we obtain, for all $y_{-m-1:0} \in \mathsf{Y}^{m+2}$,

$$\mathrm{d}(\psi^\theta \langle y_{-m:0}\rangle(x_1), \psi^\theta \langle y_{-m-1:0}\rangle(x_1)) \leq \varrho^{m+1} \, \bar{\phi}\left(y_{-m-1}\right) \,.$$

Using (54) and Lemma 15, we have that

$$\forall \eta \in (0,1), \quad \sum_{k\in\mathbb{Z}} \eta^{|k|} \bar{\phi}\left(Y_k\right) < \infty \,, \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \,, \tag{59}$$

and thus $\left(\psi^\theta \langle Y_{-m:0}\rangle(x_1)\right)_{m\geq 0}$ is a Cauchy sequence $\tilde{\mathbb{P}}^{\theta_\star}$-a.s. Its limit exists $\tilde{\mathbb{P}}^{\theta_\star}$-a.s., since (X, d) is assumed to be complete, which defines the X-valued random variable $\psi^\theta \langle Y_{-\infty:0}\rangle$ for all $\theta, \theta_\star \in \Theta$ when Y has distribution $\tilde{\mathbb{P}}^{\theta_\star}$-a.s. Thus (57) holds and we further obtain that

$$\begin{aligned}
\sup_{\theta\in\Theta} \mathrm{d}(\psi^\theta \langle Y_{-k:0}\rangle(x_1), x_1) &\leq \sup_{\theta\in\Theta} \sum_{m=0}^{k} \mathrm{d}(\psi^\theta \langle Y_{-m:0}\rangle(x_1), \psi^\theta \langle Y_{-m+1:0}\rangle(x_1)) \\
&\leq \sum_{m\geq 0} \varrho^m \, \bar{\phi}\left(Y_{-m}\right) < \infty \,, \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}
\end{aligned} \tag{60}$$

so that, letting $k \to \infty$,

$$\sup_{\theta\in\Theta} \mathrm{d}(\psi^\theta\langle Y_{-\infty:0}\rangle, x_1) \le \sum_{m\ge 0} \varrho^m\, \bar\phi\,(Y_{-m}) < \infty\ , \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{61}$$

Let us now prove (B-1). Relation (56) directly yields (B-1)(i). Let us prove (B-1)(ii), hence consider the case $\theta = \theta_\star$. Using (58), we have

$$\mathrm{d}(\psi^{\theta_\star}\langle Y_{-m:0}\rangle(x_1), \psi^{\theta_\star}\langle Y_{-m:0}\rangle(X_{-m})) \le \varrho^{m+1}\ \bar\psi(X_{-m}) \quad \mathbb{P}^{\theta_\star}\text{-a.s.}$$

Since $\{\bar\psi(X_{-m})\}_{m\ge 0}$ is stationary under $\mathbb{P}^{\theta_\star}$, it is bounded in probability, and since $\varrho < 1$, for all $\epsilon > 0$, we have

$$\lim_{m\to\infty} \mathbb{P}^{\theta_\star}\left(\mathrm{d}\left(\psi^{\theta_\star}\langle Y_{-m:0}\rangle(X_{-m}), \psi^{\theta_\star}\langle Y_{-m:0}\rangle(x)\right) > \epsilon\right) = 0\ . \tag{62}$$

Note that for all $m \ge 1$, $\psi^{\theta_\star}\langle Y_{-m:0}\rangle(X_{-m}) = X_1$ $\mathbb{P}^{\theta_\star}$-a.s., hence we get that

$$\psi^{\theta_\star}\langle Y_{-\infty:0}\rangle = X_1 \quad \mathbb{P}^{\theta_\star}\text{-a.s.} \tag{63}$$

To complete the proof of (B-1)(ii), we need to show that, under $\tilde{\mathbb{P}}^{\theta_\star}$, $y \mapsto g^{\theta_\star}(\psi^{\theta_\star}\langle Y_{-\infty:0}\rangle; y) = g^{\theta_\star}(X_1; y)$ is the conditional density of Y_1 given $Y_{-\infty:0}$, that is, for any $B \in \mathcal{Y}$,

$$\int 1_B(y) g^{\theta_\star}\,(X_1; y)\,\nu(\mathrm{d}y) = \mathbb{P}^{\theta_\star}\,(Y_1 \in B \,|\, Y_{-\infty:0})\ .$$

Now, note that, by defintion of $\mathbb{P}^{\theta_\star}$,

$$\int 1_B(y) g^{\theta_\star}\,(X_1; y)\,\nu(\mathrm{d}y) = \mathbb{P}^{\theta_\star}\,(Y_1 \in B \mid X_1) = \mathbb{P}^{\theta_\star}\,(Y_1 \in B \mid X_1, Y_{-\infty:0})\ .$$

But since (63) implies that X_1 is $\sigma(Y_{-\infty:0})$-measurable, X_1 can be removed in the last conditioning, which concludes the proof (B-1)(ii).

Finally, it remains to show the uniform convergence (41) in (B-5). By (B-3) and (57), we have, for all $\theta, \theta_\star \in \Theta$, $k \in \mathbb{Z}_+$,

$$\psi^\theta\langle Y_{-\infty:k-1}\rangle = \psi^\theta\langle Y_{1:k-1}\rangle\left(\psi^\theta\langle Y_{-\infty:0}\rangle\right)\ , \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{64}$$

From (B-4)(iii) and (64), we get

$$\mathrm{d}(\psi^\theta\langle Y_{1:k-1}\rangle(x_1), \psi^\theta\langle Y_{-\infty:k-1}\rangle) \le \varrho^{k-1}\bar\psi\left(\psi^\theta\langle Y_{-\infty:0}\rangle\right)\ , \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}$$

On the other hand (B-4)(iv) and (61) imply

$$\sup_{\theta\in\Theta} \bar\psi\left(\psi^\theta\langle Y_{-\infty:0}\rangle\right) < \infty\ , \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}\ , \tag{65}$$

which, with the previous display, yields,

$$\sup_{\theta\in\Theta} \mathrm{d}(\psi^\theta\langle Y_{1:k-1}\rangle(x_1), \psi^\theta\langle Y_{-\infty:k-1}\rangle) = O_{k\to\infty}\left(\varrho^k\right) \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \tag{66}$$

Since X_1 is closed and satisfies Condition (B-4)(i), we have that, $\psi^\theta\langle Y_{1:k-1}\rangle(x_1)$ and $\psi^\theta\langle Y_{-\infty:k-1}\rangle$ are in X_1 for all $k \geq 2$. Thus Condition (B-4)(vi) gives that

$$\sup_{\theta\in\Theta}\left|\ln\frac{g^\theta(\psi^\theta\langle Y_{1:k-1}\rangle(x_1); Y_k)}{g^\theta(\psi^\theta\langle Y_{-\infty:k-1}\rangle; Y_k)}\right| \leq A_k(1)\times A_k(2)\times A_k(3)\times A_k(4) \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}\,,$$

where

$$\begin{aligned}
A_k(1) &= \sup_{\theta\in\Theta} H\left(\mathrm{d}(\psi^\theta\langle Y_{1:k-1}\rangle(x_1), \psi^\theta\langle Y_{-\infty:k-1}\rangle)\right)\\
A_k(2) &= \sup_{\theta\in\Theta} \mathrm{e}^{C\,\mathrm{d}\left(x_1, \psi^\theta\langle Y_{-\infty:k-1}\rangle\right)}\\
A_k(3) &= \sup_{\theta\in\Theta} \mathrm{e}^{C\,\mathrm{d}\left(x_1, \psi^\theta\langle Y_{1:k-1}\rangle(x_1)\right)}\\
A_k(4) &= \bar{\phi}(Y_k)\,.
\end{aligned}$$

By (66) and (B-4)(vii), we have

$$A_k(1) = O_{k\to\infty}\left(\varrho^k\right) \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}$$

With (59), this yields (41) in the case where $C = 0$. For $C > 0$, we further observe that, by (61) and (55), we have, for all $\theta_\star \in \Theta$ and $k \in \mathbb{Z}_+$,

$$\tilde{\mathbb{E}}^{\theta_\star}\left[\ln^+ A_k(2)\right] \leq \tilde{\mathbb{E}}^{\theta_\star}\left[C\sum_{m\geq 0}^{\infty}\varrho^m\bar{\phi}\left(Y_{-m+k-1}\right)\right] = \frac{C\pi_2^{\theta_\star}\left(\bar{\phi}\right)}{1-\varrho} < \infty\,.$$

Then Lemma 15 implies that, $\tilde{\mathbb{P}}^{\theta_\star}$-a.s., $A_k(2) = O(\eta^{-k})$ for any $\eta \in (0,1)$. The same property applies similarly to $A_k(3)$ by using (60) in place of (61). This yields (41) in the case where $C > 0$, which concludes the proof. □

Acknowledgement

We are thankful to the editor-in-charge and anonymous referee for the insightful comments and the helpful suggestions that lead to improve this paper.

References

[1] Carol Alexander and Emese Lazar. Normal mixture garch (1, 1): Applications to exchange rate modelling. *Journal of Applied Econometrics*, 21(3):307–336, 2006.

[2] T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, 31:307–327, 1986.

[3] Tim Bollerslev. Glossary to arch (garch). Technical report, CREATES Research Paper, September 2008.

[4] DR Cox. Statistical analysis of time-series: some recent developments. *Scand. J. Statist.*, 8(2):93–115, 1981.

[5] RA Davis, WTM Dunsmuir, and SB Streett. Observation-driven models for Poisson counts. *Biometrika*, 90(4):777–790, DEC 2003.

[6] R.A. Davis and H. Liu. Theory and inference for a class of observation-driven models with application to time series of counts. *Preprint, arXiv:1204.3915*, 2012.

[7] 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(7):2620–2647, 2013.

[8] Randal Douc, Francois Roueff, and Tepmony Sim. The maximizing set of the asymptotic normalized log-likelihood for partially observed markov chains. 2014.

[9] Paul Doukhan, Konstantinos Fokianos, and Dag Tjøstheim. On weak dependence conditions for Poisson autoregressions. *Statist. Probab. Lett.*, 82(5):942–948, 2012.

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

[11] Konstantinos Fokianos, Anders Rahbek, and Dag Tjøstheim. Poisson autoregression. *J. Am. Statist. Assoc.*, 104(488):1430–1439, 2009. With electronic supplementary materials available online.

[12] Markus Haas, Stefan Mittnik, and Marc S Paolella. Mixed normal conditional heteroskedasticity. *Journal of Financial Econometrics*, 2(2):211–250, 2004.

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

[14] B. G. Leroux. Maximum-likelihood estimation for hidden Markov models. *Stoch. Proc. Appl.*, 40:127–143, 1992.

[15] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, London, 2009.

[16] E. Moulines, P. Priouret, and F. Roueff. On recursive estimation for time varying autoregressive processes. *Ann. Statist.*, 33(6):2610–2654, 2005. available at [arXiv].

[17] Michael H. Neumann. Absolute regularity and ergodicity of Poisson count processes. *Bernoulli*, 17(4):1268–1284, NOV 2011.

[18] J. Pfanzagl. On the measurability and consistency of minimum contrast estimates. *Metrica*, 14:249–272, 1969.

[19] S. Streett. *Some observation driven models for time series of counts*. PhD thesis, Colorado State University, Department of Statistics, 2000.

[20] Fukang Zhu. A negative binomial integer-valued GARCH model. *J. Time Series Anal.*, 32(1):54–67, 2011.