



**HAL**  
open science

# Practical conditions for the convergence of the maximum likelihood estimator in observation-driven models

Randal Douc, François Roueff, Tepmony Sim

## ► To cite this version:

Randal Douc, François Roueff, Tepmony Sim. Practical conditions for the convergence of the maximum likelihood estimator in observation-driven models. 2014. hal-01078073v1

**HAL Id: hal-01078073**

**<https://hal.science/hal-01078073v1>**

Preprint submitted on 7 Nov 2014 (v1), last revised 3 Jun 2015 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Practical conditions for the convergence of the maximum likelihood estimator in observation-driven models

R. Douc<sup>a</sup>, F. Roueff<sup>b</sup>, T. Sim<sup>b</sup>

<sup>a</sup>*Department CITI, CNRS UMR 5157, Telecom Sudparis, Evry. France.*

<sup>b</sup>*Telecom Paristech, Institut Telecom, CNRS LTCI, Paris. France.*

---

## Abstract

This paper generalizes asymptotic properties obtained in the observation-driven times series models considered by Douc et al. (2013) 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, however, appear to apply for a wide class of models. We illustrate our results with two recently introduced observation-driven times series, namely, the NBIN-GARCH and the NM-GARCH, demonstrating the consistency of the MLE for these two models.

*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 Cox (1981) and later considered by Streett (2000), Davis et al. (2003), Fokianos et al. (2009), Neumann (2011), Doukhan et al. (2012), Davis and Liu (2012) and Douc et al. (2013). The celebrated GARCH(1, 1) model, see Bollerslev (1986), as well as most of the models derived from this one, see Bollerslev (2008) 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

---

*Email addresses:* [randal.douc@telecom-sudparis.eu](mailto:randal.douc@telecom-sudparis.eu) (R. Douc),  
[roueff@telecom-paristech.fr](mailto:roueff@telecom-paristech.fr) (F. Roueff), [sim@telecom-paristech.fr](mailto:sim@telecom-paristech.fr) (T. Sim)

example). When the observed variable is discrete, general results concerning the consistency of the MLE have been obtained recently in Henderson et al. (2011) and Douc et al. (2013). However, the consistency result of Douc et al. (2013) 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 two simple examples, to which the results of Douc et al. (2013) can not be directly applied. The first one is the negative binomial integer-valued GARCH (NBIN-GARCH) model, which was first introduced by Zhu (2011) as a generalization of the Poisson IN-GARCH model.

**Example 1** (Definition of the NBIN-GARCH(1, 1) model).

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

where  $X_k$  takes values in  $\mathbb{X} = \mathbb{R}_+$ ,  $Y_k$  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  $\Gamma$  stands for the Gamma function.

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. To the best of our knowledge the consistency of the MLE has not been treated for this model.

The second example is the univariate normal mixture GARCH model (NM-GARCH) proposed by Haas et al. (2004) and later considered by Alexander and Lazar (2006). It is formally defined by

**Example 2** (Definition of the NM( $d$ )-GARCH(1, 1) model where  $d \in \mathbb{N} \setminus \{0\}$ ).

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

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

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

$\boldsymbol{\omega}$ ,  $\mathbf{b}$  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  $G^\theta$  depends on  $\theta$  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  $\boldsymbol{\gamma} = \gamma_1 = 1$ ,  $G^\theta$  no longer depends on  $\theta$ .

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. Up to our knowledge, the usual consistency proof of the MLE for the GARCH cannot be directly adapted to this model.

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  $\Theta_*$  of a limit criterion, and second the maximizing set indeed reduces to the true parameter  $\theta_*$ , 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 Douc et al. (2013), which is valid for a restricted class of models, to a larger class of models in which the two 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 in practice, as shown by the two 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 two examples introduced above. Finally, Section 5 provides the proofs of the main results, mainly inspired from Douc et al. (2013). Appendix A contains the proof of a technical lemma.

## 2. Definitions and notation

Consider a bivariate stochastic process  $\{(X_k, Y_k) : k \in \mathbb{Z}_+\}$  on  $\mathsf{X} \times \mathsf{Y}$ , where  $(\mathsf{X}, d)$  is a complete and separable metric space endowed with the associated Borel  $\sigma$ -field  $\mathcal{X}$  and  $(\mathsf{Y}, \mathcal{Y})$  is a Borel space. Let  $(\Theta, \Delta)$ , 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  $\mathsf{Y}$  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  $K^\theta$  satisfies

$$K^\theta((x, y); dx' dy') = \delta_{\psi_y^\theta(x)}(dx') G^\theta(x'; dy'), \quad (4)$$

where  $\delta_a$  denotes the Dirac mass at point  $a$ . Moreover, we will say that the observation-driven time series model is dominated by some  $\sigma$ -finite measure  $\nu$  on  $(\mathsf{Y}, \mathcal{Y})$  if for all  $x \in \mathsf{X}$ , the probability kernel  $G^\theta(x; \cdot)$  is dominated by  $\nu$ . In this case we denote by  $g^\theta(x; \cdot)$  its Radon-Nikodym derivative,  $g^\theta(x; y) = \frac{dG^\theta(x; \cdot)}{d\nu}(y)$ , 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 Douc et al. (2013) 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

$$R^\theta(x; A) = \int \mathbb{1}_A(\psi_y^\theta(x)) G^\theta(x; dy), \quad x \in \mathsf{X}, A \in \mathcal{X}. \quad (5)$$

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  $k \geq 0$ ,

$$\begin{aligned} X_{k+1} &= \psi_{Y_k}^\theta(X_k), \\ Y_{k+1} | \mathcal{F}_k &\sim G^\theta(X_{k+1}; \cdot), \end{aligned} \quad (6)$$

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

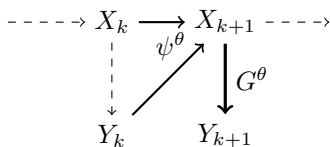


Figure 1: Graphical representation of the observation-driven model.

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  $(Y_1, \dots, Y_n)$  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  $\theta$ , for all  $x \in \mathsf{X}$ ,

$$y_{1:n} \mapsto \prod_{k=1}^n g^\theta(\psi^\theta\langle y_{1:k-1} \rangle(x); y_k) , \quad (7)$$

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

$$\psi^\theta\langle y_{1:p} \rangle = \psi_{y_p}^\theta \circ \psi_{y_{p-1}}^\theta \circ \dots \circ \psi_{y_1}^\theta , \quad (8)$$

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  $\theta$ , is defined by

$$\hat{\theta}_{x,n} \in \operatorname{argmax}_{\theta \in \Theta} \mathsf{L}_{x,n}^\theta\langle Y_{1:n} \rangle , \quad (9)$$

where

$$\mathsf{L}_{x,n}^\theta\langle y_{1:n} \rangle := n^{-1} \sum_{k=1}^n \ln g^\theta(\psi^\theta\langle y_{1:k-1} \rangle(x); y_k) . \quad (10)$$

In this contribution, we study the convergence of  $\hat{\theta}_{x,n}$  as  $n \rightarrow \infty$  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  $\theta$ . 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  $K^\theta$  of the complete chain admits a unique stationary distribution  $\pi^\theta$  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  $\mu$  on  $\mathsf{X} \times \mathsf{Y}$ , we denote by  $\mathbb{P}_\mu^\theta$  the distribution of the Markov chain  $\{(X_k, Y_k), k \geq 0\}$  with kernel  $K^\theta$  and initial probability measure  $\mu$ . Under Assumption (A-1), we denote by  $\pi_1^\theta$  and  $\pi_2^\theta$  the marginal distributions of  $\pi^\theta$  on  $\mathsf{X}$  and  $\mathsf{Y}$ , 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  $\mathbb{P}_{\pi^\theta}^\theta$  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  $B \in \mathcal{Y}^{\otimes(m+\mathbb{Z}_+^*)}$ ,

$$\tilde{\mathbb{P}}^\theta (\mathcal{Y}^{m+\mathbb{Z}^-} \times B) = \mathbb{P}^\theta (\mathcal{X}^{\mathbb{Z}} \times (\mathcal{Y}^{m+\mathbb{Z}^-} \times B)) = \mathbb{P}_{\pi^\theta}^\theta (\mathcal{X}^{m+\mathbb{Z}_+^*} \times B) , \quad (11)$$

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

$$\tilde{\mathbb{P}}^\theta (Y_{m+1:\infty} \in B) = \mathbb{P}^\theta (Y_{m+1:\infty} \in B) = \mathbb{P}_{\pi^\theta}^\theta (Y_{m+1:\infty} \in B) . \quad (12)$$

Here and in what follows, we abusively use the same notation  $Y_k$  both for the canonical projection defined on  $\mathcal{Y}^{\mathbb{Z}}$  and for the one defined on  $(\mathcal{X} \times \mathcal{Y})^{\mathbb{Z}_+}$ . 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 Douc et al. (2013) 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  $n \rightarrow \infty$ , there exists a  $(\mathcal{Y}^{\mathbb{Z}}, \mathcal{Y}^{\otimes \mathbb{Z}}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$  measurable function  $p^{\theta, \theta^*}(\cdot | \cdot)$  such that the normalized log-likelihood  $\mathbb{L}_{x,n}^\theta \langle Y_{1:n} \rangle$  defined in (10), for some appropriate value of  $x$ , can be approximated by

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

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

$$p^{\theta, \theta^*} (y_1 | y_{-\infty:0}) = \lim_{m \rightarrow \infty} g^\theta (\psi^\theta \langle y_{-m:0} \rangle (x); y_1) . \quad (13)$$

Then, by (A-1), the process  $Y$  is ergodic under  $\tilde{\mathbb{P}}^{\theta^*}$  and provided that

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

it can be shown that

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

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

$$\Theta_\star = \operatorname{argmax}_{\theta \in \Theta} \tilde{\mathbb{E}}^{\theta^*} [\ln p^{\theta, \theta^*} (Y_1 | Y_{-\infty:0})] , \quad (14)$$

which only depends on  $\theta_*$ . In this contribution, we provide practical conditions implying

$$\lim_{n \rightarrow \infty} \Delta(\hat{\theta}_{x,n}, \Theta_*) = 0, \quad \tilde{\mathbb{P}}^{\theta_*}\text{-a.s.}, \quad (15)$$

but, for the sake of brevity, we do not precisely determine the set  $\Theta_*$ . 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  $\Theta_* = \{\theta_*\}$ , in which case the strong consistency of the MLE follows from (15). We will mention a general result which precises how the set  $\Theta_*$  is related to the true parameter  $\theta_*$  in Remark 2. For the moment, let us mention that we have

$$\theta_* \in \Theta_*, \quad (16)$$

provided that the following assumption holds:

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

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

Indeed, (16) follows by writing for all  $\theta \in \Theta$ ,

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

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  $\bar{V} : \mathsf{X} \rightarrow \mathbb{R}_+$  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  $g^\theta$  and  $\psi^\theta$ .

(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  $x \in X$ .

(B-4) There exist  $x_1 \in X$ , a closed set  $X_1 \subseteq X$ ,  $\varrho \in (0, 1)$ ,  $C \geq 0$  and measurable functions  $\bar{\psi} : X_1 \rightarrow \mathbb{R}_+$ ,  $H : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  and  $\bar{\phi} : Y \rightarrow \mathbb{R}_+$  such that the following assertions hold.

(i) For all  $\theta \in \Theta$  and  $(x, y) \in X \times Y$ ,  $\psi_y^\theta(x) \in X_1$ .

(ii)  $\sup_{(\theta, x, y) \in \Theta \times X_1 \times Y} g^\theta(x; y) < \infty$ .

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

$$d(\psi^\theta \langle y_{1:n} \rangle(x_1), \psi^\theta \langle y_{1:n} \rangle(x)) \leq \varrho^n \bar{\psi}(x), \quad (17)$$

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

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

(vi) For all  $\theta \in \Theta$  and  $(x, x', y) \in X_1 \times X_1 \times Y$ ,

$$\left| \ln \frac{g^\theta(x; y)}{g^\theta(x'; y)} \right| \leq H(d(x, x')) e^{C(d(x_1, x) \vee d(x_1, x'))} \bar{\phi}(y), \quad (18)$$

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

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

$$G^\theta \ln^+ \bar{\phi} \lesssim \bar{V}, \quad (19)$$

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

$$G^\theta \bar{\phi} \lesssim \bar{V}. \quad (20)$$

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  $x_1 \in X$  as in (B-4), the function  $p^{\theta, \theta_\star}(\cdot | \cdot)$  defined by the limit (13) with  $x = x_1$  satisfies (B-1) and the convergence (15) of the MLE holds with the set  $\Theta_\star$  defined by (14).*

For convenience, the proof is postponed to Section 5.1. The consistency of the MLE then follows from Theorem 3 by the following remark.

**Remark 2.** In (Douc et al., 2014, Section 4.2), it is shown that the assumptions of Theorem 3 imply that any  $\theta \in \Theta_\star$  is 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 was already introduced in Leroux (1992). Note that the usual consistency then immediately follows from the identifiability of the parameter:  $\tilde{\mathbb{P}}^\theta = \tilde{\mathbb{P}}^{\theta_\star} \Rightarrow \theta = \theta_\star$ .

### 3.3. Ergodicity

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

Ergodicity of Markov chains are usually studied using  $\psi$ -irreducibility. This approach is well known to be quite efficient when dealing with fully dominated models, see Meyn and Tweedie (2009). It is not at all the same picture for observation-driven models, where other tools need to be invoked, see Fokianos and Tjøstheim (2011); Douc et al. (2013). Since the ergodicity is studied for a given parameter  $\theta$ , the ergodicity results of Douc et al. (2013) directly apply, even though observation-driven models are restricted to the case where  $g$  does not depend on the unknown parameter  $\theta$  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  $\alpha$  and  $\phi$  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  $\sigma$ -field  $\mathcal{X}$  is the Borel  $\sigma$ -field.

(A-4) There exist  $(\lambda, \beta) \in (0, 1) \times \mathbb{R}_+$  and a measurable function  $V : X \rightarrow \mathbb{R}_+$  such that  $RV \leq \lambda V + \beta$ .

(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  $x_0 \in X$  such that, for any  $x \in X$  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 
$$\sup_{\substack{(x, x', y) \in X^2 \times Y \\ x \neq x'}} \frac{d(\psi_y(x), \psi_y(x'))}{d(x, x')} < 1.$$

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

(i) For all  $(x, x') \in X^2$ ,  $\alpha(x, x') = 1$  if and only if  $x = x'$ .

(ii) For all  $x \in X$ ,  $\phi(x, x) = x$ .

(iii) For all  $(x, x') \in X^2$  and  $y \in Y$ ,

$$\min \{g(x; y), g(x'; y)\} \geq \alpha(x, x')g(\phi(x, x'); y) . \quad (21)$$

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

- (v) For all  $(x, x') \in \mathsf{X}^2$ ,  $1 - \alpha(x, x') \leq d(x, x')W(x, x')$ .
- (vi)  $\sup \left( \int_{\mathsf{Y}} W(\psi_y(x), \psi_y(x')) G(\phi(x, x'); dy) - W(x, x') \right) < \infty$ , 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  $\pi$  on  $\mathsf{X} \times \mathsf{Y}$ . Moreover  $\pi_1 \bar{V} < \infty$  for every  $\bar{V} : \mathsf{X} \rightarrow \mathbb{R}_+$  such that  $\bar{V} \lesssim V$ .*

The proof of Theorem 4 is postponed to Section 5.2 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  $\theta$ ,  $\bar{V}$  does not.

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

$$\bar{\mathbb{P}}_x := \mathbb{P}_{\delta_x \otimes G(x; \cdot)} \quad (22)$$

on  $(\mathsf{X} \times \mathsf{Y})^{\mathbb{Z}_+}$ , where  $\mathbb{P}_\mu$  is defined for any distribution  $\mu$  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  $\delta_x$ .

We now provide a general framework for constructing  $\alpha$  and  $\phi$  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  $\nu$  and there exists a probability measure  $\mu$  on  $(\mathsf{S}, \mathcal{S})$  such that*

$$g(x; y) = \int_{\mathsf{S}} j(x_s)h(x_s; y)\mu(ds), \quad y \in \mathsf{Y}. \quad (23)$$

We moreover assume that  $h$  takes non-negative values,  $j$  is one-to-one 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

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

Then  $\alpha$  and  $\phi$  defined above satisfy (A-8)(i), (A-8)(ii) and (A-8)(iii).

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

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

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

Thus  $j$  is strictly decreasing on  $\mathbb{C}$ . Clearly, the defined  $\alpha$  and  $\phi$  are positive and continuous functions and satisfy (A-8)(i) and (A-8)(ii), respectively. For all  $(x, x') \in \mathbb{X}^2$  and  $y \in \mathbb{Y}$ , we have

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

By symmetry of  $\alpha$  and  $\phi$ , we get (21) and thus (A-8)(ii) holds.  $\square$

#### 4. Examples

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

##### 4.1. NBIN-GARCH model

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

$$\psi_y^\theta(x) = \omega + ax + by, \quad (24)$$

$$g^\theta(x; y) = \frac{\Gamma(y+r)}{y! \Gamma(r)} \left( \frac{1}{1+x} \right)^r \left( \frac{x}{1+x} \right)^y, \quad (25)$$

with  $\theta = (\omega, a, b, r)$  in a compact subset  $\Theta$  of  $(0, \infty)^2 \times \mathbb{R}_+^2$  and  $\mathbb{X} = (0, \infty)$ .

In (Zhu, 2011, 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

$$rb + a < 1. \quad (26)$$

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 (26) 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 2, provides the consistency of the MLE  $\hat{\theta}_{x_1, n}$  for any  $x_1 \in \mathbb{X}$ .

**Theorem 6.** *Suppose that all  $\theta = (\omega, a, b, r)$  in  $\Theta$  satisfy Condition (26). Then Assumptions (A-1), (A-2), (B-2), (B-3) and (B-4) hold with  $\bar{V}$  being defined as the identity function on  $\mathsf{X}$  and with any  $x_1 \in \mathsf{X}$ .*

*Proof.* For convenience, we divide the proof in 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  $p \rightarrow p'$ ,  $\mathcal{NB}(r, p)$  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  $x_\infty$ . Therefore, defining  $\mathbb{P}_x$  as in (22), for any neighborhood  $\mathcal{N}$  of  $x_\infty$ , there exists some  $n$  such that  $x_n \in \mathcal{N}$  and we have

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

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  $\mu$  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,  $j$  is one-to-one and  $h$  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), (A-8)(ii) and (A-8)(iii). For any given  $r > 0$ , let a function  $W : \mathsf{X}^2 \rightarrow [1, \infty)$  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)(iv) and (A-8)(vi) clearly hold. Moreover, (A-8)(v) 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  $\Theta$ , 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

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

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  $x_1 \in X$ . Let  $X_1 = [\underline{\omega}, \infty) \subset X$  so that (B-4)(i) holds. By noting that for all  $(\theta, x, y) \in \Theta \times X \times Y$ ,  $g^\theta(x; y) \leq 1$ , we have (B-4)(ii). From (8) and (24), we have for all  $s \leq t$ ,  $y_{s:t} \in Y^{t-s+1}$ ,  $x \in X$  and  $\theta \in \Theta$ ,

$$\psi^\theta \langle y_{s:t} \rangle (x) = \omega \left( \frac{1 - a^{t-s+1}}{1 - a} \right) + a^{t-s+1}x + b \sum_{j=0}^{t-s} a^j y_{t-j}. \quad (27)$$

Using (27), we have, for all  $\theta \in \Theta$ ,  $x \in X$  and  $y_{1:n} \in Y^n$ ,

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

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  $y \in Y$ ,

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

and, for all  $(x, x') \in X_1^2 = [\underline{\omega}, \infty)^2$ ,

$$\begin{aligned} |\ln g^\theta(x; y) - \ln g^\theta(x'; y)| &= |(r + y) [\ln(1 + x') - \ln(1 + x)] + y [\ln x - \ln x']| \\ &\leq [(r + y)(1 + \underline{\omega})^{-1} + y \underline{\omega}^{-1}] |x - x'| \\ &\leq [\bar{r} + y(1 + \underline{\omega}^{-1})] |x - x'|. \end{aligned}$$

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

$$\int \ln^+ y G^\theta(x, dy) \leq \int y G^\theta(x, dy) = rx \leq \bar{r}\bar{V}(x).$$

This concludes the proof.  $\square$

#### 4.2. NM-GARCH model

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

$$\psi_y^\theta(\mathbf{x}) = \boldsymbol{\omega} + \mathbf{A}\mathbf{x} + y^2 \mathbf{b}, \quad (28)$$

$$g^\theta(\mathbf{x}; y) = \sum_{\ell=1}^d \gamma_\ell \frac{e^{-y^2/2x_\ell}}{(2\pi x_\ell)^{1/2}}, \quad (\mathbf{x}, y) \in X \times Y, \quad (29)$$

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

In Haas et al. (2004), 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

$$|\lambda|_{\max}(\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T) < 1, \quad (30)$$

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 (30) but it does not say anything about the existence of stationary or weakly stationary solution. The result below shows that (30) is indeed a sufficient condition for the existence of a stationary solution with finite variance. It moreover provides with Theorem 3 and Remark 2 the consistency of the MLE  $\hat{\theta}_{\mathbf{x}_1, n}$  for any  $\mathbf{x}_1 \in \mathbf{X}$ .

**Theorem 7.** *Suppose that all  $\theta = (\boldsymbol{\gamma}, \boldsymbol{\omega}, \mathbf{A}, \mathbf{b})$  in  $\Theta$  satisfy Condition (30). Then Assumptions (A-1), (A-2), (B-2), (B-3) and (B-4) hold with  $\bar{V}$  being defined as any norm on  $\mathbf{X}$ .*

*Proof.* In this proof section, we set

$$\bar{V}(\mathbf{x}) = |\mathbf{x}| = \sum_{\ell=1}^d |x_\ell|, \quad (31)$$

for all  $\mathbf{x} = (x_\ell) \in \mathbf{X}$ . As in Theorem 6, we divide the proof in 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  $\mathbf{X}$  by setting

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

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

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

We indeed note that by Condition (30) the above inversion is well defined and moreover

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

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  $\mathbf{x} = (x_\ell) \in \mathbf{X}$ ,

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

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  $\theta$  in the notation in the remaining of **Step 1**). Condition (A-3) holds for any distance  $d$  associated to a norm on the finite dimensional space  $\mathbf{X}$ .

(The precise choice of  $d$  is postponed to the verification of (A-7).) We have

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

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 (8) and (28), we have for all  $n \geq 1$ ,  $y_{0:n-1} \in \mathcal{Y}^n$  and  $\mathbf{x} \in \mathcal{X}$ ,

$$\psi^\theta \langle y_{0:n-1} \rangle (\mathbf{x}) = \mathbf{A}^n \mathbf{x} + \sum_{j=0}^{n-1} \mathbf{A}^j (\boldsymbol{\omega} + y_{n-1-j}^2 \mathbf{b}) . \quad (32)$$

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 (30), there exists  $\delta \in (0, 1)$  and  $c > 0$  such that, for any  $k \geq 1$ ,

$$\left\| (\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T)^k \right\| \leq c \delta^k . \quad (33)$$

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

$$\|\mathbf{A}^k\| \leq \left\| (\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T)^k \right\| . \quad (34)$$

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

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

With (33) and (34), 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 (29), for any positive  $\epsilon$ ,  $\mathbf{x} \in \mathbf{X}$  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 (33) and (34) imply that  $|\lambda|_{\max}(\mathbf{A}) < 1$ , there exists a vector norm which makes  $\mathbf{A}$  strictly contracting. Choosing the metric  $d$  on  $\mathbf{X}$  as the distance corresponding to this norm, we get (A-7). To show (A-8), we again rely on Lemma 5. Let us set  $\mathbf{C} = (0, \infty)$  and  $\mathbf{S} = \{1, \dots, d\}$  and define the probability measure  $\mu$  on  $\mathbf{S}$  by  $\mu(\{s\}) = \gamma_s$ , for all  $s \in \mathbf{S}$ . For all  $(t, y) \in \mathbf{C} \times \mathbf{Y}$ , let  $j(t) = \frac{1}{(2\pi t)^{1/2}}$  and  $h(t; y) = \exp(-y^2/2t)$ . Obviously,  $j$  is one-to-one, Relation (23) holds and  $h$  satisfies (F-1). Hence, Lemma 5 implies that  $\alpha$  and  $\phi$  defined respectively for all  $\mathbf{x} = (x_1, \dots, x_d)$ ,  $\mathbf{x}' = (x'_1, \dots, x'_d) \in \mathbf{X}$  by

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

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

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

where  $d$  is the distance previously defined and  $W$  is defined by  $W(\mathbf{x}, \mathbf{x}') = 1 \vee (c_d \min_{1 \leq \ell \leq d} (x_\ell^{-1} \wedge x'_\ell^{-1}))$  with  $c_d > 0$  is conveniently chosen (such a constant exists since  $d$  is the distance associated to a norm and  $\mathbf{X}$  has finite dimension). Then (A-8)(iv) and (A-8)(v) hold and, since for all  $y \in \mathbf{Y}$  and  $x \in \mathbf{X}$ ,  $\psi_y(\mathbf{x})$  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  $(\mathbf{x}, \mathbf{x}', y) \in \mathbf{X} \times \mathbf{X} \times \mathbf{Y}$  and (A-8)(vi) 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  $\Theta$  is compact, then

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

for some  $(\underline{\omega}, \bar{\omega}, \underline{b}, \bar{b}, \bar{\rho}) \in (0, \infty)^4 \times (0, 1)$  and  $L > 0$ . By (Moulines et al., 2005, 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  $k \geq 1$  and all  $\theta \in \Theta$ ,

$$\left\| (\mathbf{A} + \mathbf{b}\boldsymbol{\gamma}^T)^k \right\| \leq \bar{C} \bar{\delta}^k. \quad (35)$$

We set  $\mathbf{X}_1 = [\underline{\omega}, \infty)^d \subset \mathbf{X}$  so that (B-4)(i) holds. Moreover, for all  $(\theta, \mathbf{x}, y) \in \Theta \times \mathbf{X}_1 \times \mathbf{Y}$ ,  $g^\theta(\mathbf{x}; y) \leq (2\pi\underline{\omega})^{-1/2}$ . Thus, Condition (B-4)(ii) holds. Now let  $x_1 \in \mathbf{X}$ . Using (32), (35) and (34), we have, for all  $\mathbf{x} \in \mathbf{X}$ ,  $y_{1:n} \in \mathbf{Y}^n$  and  $\theta \in \Theta$ ,

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

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

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

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

$$\bar{\phi}(y) \geq (L+1)|\mathbf{x}_1| + \bar{\omega} + y^2\bar{b}. \quad (36)$$

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

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

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

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

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

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

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

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

Hence, using (31) with the above definitions, we obtain (B-4)(viii) and the proof is concluded.  $\square$

## 5. Postponed proofs

### 5.1. 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 Pfanzagl (1969), which was already used in Douc et al. (2013) for a restricted class of observation-driven models. Our main contribution here is to provide the practical conditions of Assumption (B-4). We divide the proof in two steps. The first step is to show that our conditions imply (B-1) and the following one.

(B-5) There exists  $x_1 \in \mathsf{X}$  such that, for all  $\theta, \theta_* \in \Theta$ ,  $p^{\theta, \theta_*}(Y_1 | Y_{-\infty:0})$  is well defined by (13) with  $x = x_1$ ,  $\tilde{\mathbb{P}}^{\theta_*}$ -a.s. Moreover, for all  $\theta_* \in \Theta$ , we have

$$\limsup_{k \rightarrow \infty} \sup_{\theta \in \Theta} \left| \ln \frac{g^\theta(\psi^\theta \langle Y_{1:k-1} \rangle(x_1); Y_k)}{p^{\theta, \theta_*}(Y_k | Y_{-\infty:k-1})} \right| = 0 \quad \tilde{\mathbb{P}}^{\theta_*}\text{-a.s.} \quad (37)$$

Indeed we have the following lemma.

**Lemma 8.** *Assumptions (B-2), (B-3) and (B-4) imply (B-5) and (B-1).*

*Proof.* See Appendix A. □

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

**Lemma 9.** *Assume that (B-2), (B-3) and (B-4)(i)–(ii) hold and that  $x_1$  satisfies (B-5). Then  $\Theta_*$  defined by (14) is a non-empty closed subset of  $\Theta$  and (15) holds.*

*Proof.* By (Douc et al., 2013, Theorem 33), to obtain (15), it is sufficient to show that, for all  $\theta_* \in \Theta$ , the two following assertions hold.

- (a)  $\tilde{\mathbb{E}}^{\theta_*} [\sup_{\theta \in \Theta} \ln^+ p^{\theta, \theta_*}(Y_1 | Y_{-\infty:0})] < \infty$ ,
- (b) the function  $\theta \mapsto \ln p^{\theta, \theta_*}(Y_1 | Y_{-\infty:0})$  is continuous on  $\Theta$ ,  $\tilde{\mathbb{P}}^{\theta_*}$ -a.s.

In (B-5),  $p^{\theta, \theta_*}(Y_1 | Y_{-\infty:0})$  is defined as the limit (13) with  $x = x_1$ . So,  $\tilde{\mathbb{P}}^{\theta_*}$ -a.s., by (B-4)(i)–(ii),  $p^{\theta, \theta_*}(Y_1 | Y_{-\infty:0})$  is bounded by the finite constant appearing in (B-4)(ii). Hence Condition (a) holds.

Condition (b) then follows from (37). Since almost sure convergence implies the convergence in probability and  $\tilde{\mathbb{P}}^{\theta_*}$  is shift invariant, the random sequence

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

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

### 5.2. Ergodicity

For proving Theorem 4, we first recall a more general set of conditions derived in Douc et al. (2013), which are based on the following definition.

**Definition 10.** Let  $\bar{G}$  be a probability kernel from  $\mathsf{X}^2$  to  $\mathcal{Y}^{\otimes 2} \otimes \mathcal{P}(\{0, 1\})$  satisfying the following marginal conditions, for all  $(x, x') \in \mathsf{X}^2$  and  $B \in \mathcal{Y}$ ,

$$\begin{cases} \bar{G}((x, x'); B \times \mathsf{Y} \times \{0, 1\}) = G(x; B), \\ \bar{G}((x, x'); \mathsf{Y} \times B \times \{0, 1\}) = G(x'; B), \end{cases} \quad (38)$$

and such that the following coupling condition holds

$$\bar{G}((x, x'); \{(y, y) : y \in \mathsf{Y}\} \times \{1\}) = \bar{G}((x, x'); \mathsf{Y}^2 \times \{1\}). \quad (39)$$

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

$$\check{G}((x, x'); B) = \bar{G}((x, x'); \{(y, y) : y \in B\} \times \{1\}), \quad B \in \mathcal{Y}. \quad (40)$$

- 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

$$\bar{R}((x, x'); A \times \{1\}) = \int_{\mathsf{Y}} \mathbb{1}_A(\psi_y(x), \psi_y(x')) \check{G}((x, x'); dy). \quad (41)$$

- The measurable function  $\alpha$  from  $\mathsf{X}^2$  to  $(0, 1]$  is defined by

$$\alpha(x, x') = \bar{R}((x, x'); \mathsf{X}^2 \times \{1\}) = \bar{G}((x, x'); \mathsf{Y}^2 \times \{1\}). \quad (42)$$

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

$$\hat{R}((x, x'); A) = \frac{\bar{R}((x, x'); A \times \{1\})}{\alpha(x, x')}. \quad (43)$$

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

- (A-9) There exists a kernel  $\bar{G}$  yielding  $\alpha$  and  $\hat{R}$  as in Definition 10, a measurable function  $W : \mathsf{X}^2 \rightarrow [1, \infty)$  satisfying Conditions (A-8)(v) and (A-8)(iv) 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  $n \geq 1$ ,

$$\hat{R}^n((x, x'); d) \leq D \rho^n d(x, x'), \quad (44)$$

$$\hat{R}^n((x, x'); d \otimes W) \leq D \rho^n d^{\zeta_1}(x, x') W^{\zeta_2}(x, x'). \quad (45)$$

Under Conditions (A-3), (A-4), (A-5), (A-6) and (A-9) and by combining Theorem 6, Proposition 8 and Lemma 7 in Douc et al. (2013), we immediately obtain the following result.

**Theorem 11.** *Assume (A-3), (A-4), (A-5), (A-6) and (A-9). Then the Markov kernel  $K$  admits a unique invariant distribution  $\pi$  and  $\pi_1(\bar{V}) < \infty$  for any  $\bar{V} : \mathsf{X} \rightarrow \mathbb{R}_+$  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 (Douc et al., 2013, Lemma 9).

**Lemma 12.** *Assume that there exists  $(\rho, \beta) \in (0, 1) \times \mathbb{R}$  such that for all  $(x, x') \in \mathsf{X}^2$ ,*

$$\hat{R}((x, x'); \{(x_1, x'_1) \in \mathsf{X}^2 : d(x_1, x'_1) \leq \rho d(x, x')\}) = 1, \quad (46)$$

$$\hat{R}W \leq W + \beta. \quad (47)$$

Then, (44) and (45) hold.

Now we can prove our Practical conditions.

*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 13.** *Assume (A-8)(i) and (A-8)(ii). Then one can define a kernel  $\bar{G}$  as in Definition 10 with the same  $\alpha$  given in (42). Moreover, the kernel  $\hat{R}$  is defined by, for all  $(x, x') \in \mathsf{X}^2$  and all measurable functions  $f : \mathsf{X}^2 \rightarrow \mathbb{R}_+$ ,*

$$\hat{R}((x, x'); f) = G(\phi(x, x'); \tilde{f}) \quad \text{with} \quad \tilde{f}(y) = f(\psi_y(x), \psi_y(x')). \quad (48)$$

Let us conclude the proof of Theorem 4 before proving this lemma. By Lemma 13 and Lemma 12, it remains to check that (46) and (47) hold for all  $(x, x') \in \mathsf{X}^2$ . Observe that by definition of  $\hat{R}$ , Condition (A-8)(vi) 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 (47) 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 (48). When  $x = x'$ , then  $d(X, X') = 0$ , implying that Condition (46) holds with any nonnegative  $\rho$ . For  $x \neq x'$ , let  $\rho$  be defined by

$$\rho = \sup_{\substack{(x, x', y) \in \mathsf{X}^2 \times \mathsf{Y} \\ x \neq x'}} \frac{d(\psi_y(x), \psi_y(x'))}{d(x, x')}, \quad (49)$$

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

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

Therefore, Condition (46) holds for all  $(x, x') \in \mathsf{X}^2$  with  $\rho$  as in (49).  $\square$

We conclude this section with the postponed

*Proof of Lemma 13.* 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  $\mathsf{Y}$  with density  $g(\phi(x, x'); \cdot)$  with respect to  $\nu$ . Then we define  $(Y, Y', U)$  by separating the two cases  $x = x'$  and  $x \neq x'$ .

- Suppose that  $x = x'$ . Using (A-8)(ii), for all  $y \in \mathsf{Y}$ ,

$$g(x; y) = g(x'; y) = g(\phi(x, x'); y) .$$

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

- Suppose now that  $x \neq x'$ . From (A-8)(i),  $\alpha(x, x') < 1$  and, using (21), the functions

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

and

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

are probability density functions with respect to  $\nu$  and we let  $\Lambda$  and  $\Lambda'$  be two independent random variables taking values in  $\mathsf{Y}$  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 easily checks that the so defined kernel  $\bar{G}$  satisfies (38) and (39). Moreover, for all  $(x, x') \in \mathsf{X}^2$ ,

$$\bar{G}((x, x'); \mathsf{Y}^2 \times \{1\}) = \mathbb{P}(U = 1) = \begin{cases} 1 & \text{if } x = x' , \\ \alpha(x, x') & \text{if } x \neq x' , \end{cases}$$

hence, by (A-8)(i), is equal to  $\alpha(x, x')$  in all cases, which is compatible with (42).

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 13, observe that for any measurable  $f : \mathsf{X}^2 \rightarrow \mathbb{R}_+$ , we have, for all  $(x, x') \in \mathsf{X}^2$ ,

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

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

## Appendix A. Proof of Lemma 8.

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

$$\pi_2^\theta(\ln^+(\bar{\phi})) < \infty, \quad (\text{A.1})$$

and if moreover  $C > 0$ ,

$$\pi_2^\theta(\bar{\phi}) < \infty. \quad (\text{A.2})$$

For proving Lemma 8, we will also make use of (Douc et al., 2013, Lemma 34) which we quote here for convenience.

**Lemma 14.** *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  $\eta \in (0, 1)$ ,*

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

*Proof of Lemma 8.* We first show that  $p^{\theta, \theta_\star}$  in (13) is well defined for  $x = x_1$ . By (B-2), this follows by setting

$$p^{\theta, \theta_\star}(y_1 | y_{-\infty:0}) = g^\theta(\psi^{\theta, \theta_\star}\langle y_{-\infty:0} \rangle; y_1), \quad (\text{A.3})$$

if, for all  $\theta, \theta_\star \in \Theta$ , the following limit is well defined,

$$\lim_{m \rightarrow \infty} \psi^\theta \langle Y_{-m:0} \rangle(x_1) = \psi^{\theta, \theta_\star} \langle Y_{-\infty:0} \rangle, \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \quad (\text{A.4})$$

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

$$d(\psi^\theta \langle y_{-m:0} \rangle(x_1), \psi^\theta \langle y_{-m:0} \rangle(x)) \leq \varrho^{m+1} \bar{\psi}(x). \quad (\text{A.5})$$

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

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

Using (A.1) and Lemma 14, we have that

$$\forall \eta \in (0, 1), \quad \sum_{k \in \mathbb{Z}} \eta^{|k|} \bar{\phi}(Y_k) < \infty, \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.}, \quad (\text{A.6})$$

and thus  $(\psi^\theta \langle Y_{-m:0} \rangle(x_1))_{m \geq 0}$  is a Cauchy sequence  $\tilde{\mathbb{P}}^{\theta_\star}$ -a.s. Its limit exists  $\tilde{\mathbb{P}}^{\theta_\star}$ -a.s., which, since  $(\mathbf{X}, d)$  is assumed to be complete, defines the  $\mathbf{X}$ -valued random variable  $\psi^{\theta, \theta_\star} \langle Y_{-\infty:0} \rangle$  for all  $\theta, \theta_\star \in \Theta$ . Thus (A.4) holds and we further obtain that

$$\begin{aligned} \sup_{\theta \in \Theta} d(\psi^{\theta, \theta_\star} \langle Y_{-k:0} \rangle, x_1) &\leq \sup_{\theta \in \Theta} \sum_{m=0}^k d(\psi^\theta \langle Y_{-m:0} \rangle(x_1), \psi^\theta \langle Y_{-m+1:0} \rangle(x_1)) \\ &\leq \sum_{m \geq 0} \varrho^m \bar{\phi}(Y_{-m}) < \infty, \quad \tilde{\mathbb{P}}^{\theta_\star}\text{-a.s.} \end{aligned} \quad (\text{A.7})$$

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

$$\sup_{\theta \in \Theta} d(\psi^{\theta, \theta_*} \langle Y_{-\infty:0} \rangle, x_1) \leq \sum_{m \geq 0} \varrho^m \bar{\phi}(Y_{-m}) < \infty, \quad \tilde{\mathbb{P}}^{\theta_*}\text{-a.s.} \quad (\text{A.8})$$

Let us now prove (B-1). Relation (A.3) directly yields (B-1)(i). Let us prove (B-1)(ii), hence consider the case  $\theta = \theta_*$ . Using (A.5), we have

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

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

$$\lim_{m \rightarrow \infty} \mathbb{P}^{\theta_*} (d(\psi^{\theta_*} \langle Y_{-m:0} \rangle(X_{-m}), \psi^{\theta_*} \langle Y_{-m:0} \rangle(x)) > \epsilon) = 0. \quad (\text{A.9})$$

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

$$\psi^{\theta_*, \theta_*} \langle Y_{-\infty:0} \rangle = X_1 \quad \mathbb{P}^{\theta_*}\text{-a.s.} \quad (\text{A.10})$$

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

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

Now, note that, by definition of  $\mathbb{P}^{\theta_*}$ ,

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

But since (A.10) 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 (37) in (B-5). By (B-3) and (A.4), we have, for all  $\theta, \theta_* \in \Theta$ ,  $k \in \mathbb{Z}_+$ ,

$$\psi^{\theta, \theta_*} \langle Y_{-\infty:k-1} \rangle = \psi^\theta \langle Y_{1:k-1} \rangle (\psi^{\theta, \theta_*} \langle Y_{-\infty:0} \rangle), \quad \tilde{\mathbb{P}}^{\theta_*}\text{-a.s.} \quad (\text{A.11})$$

From (B-4)(iii) and (A.11), we get

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

On the other hand (B-4)(iv) and (A.8) imply

$$\sup_{\theta \in \Theta} \bar{\psi}(\psi^{\theta, \theta_*} \langle Y_{-\infty:0} \rangle) < \infty, \quad \tilde{\mathbb{P}}^{\theta_*}\text{-a.s.}, \quad (\text{A.12})$$

which, with the previous display, yields,

$$\sup_{\theta \in \Theta} d(\psi^\theta \langle Y_{1:k-1} \rangle(x_1), \psi^{\theta, \theta_*} \langle Y_{-\infty:k-1} \rangle) = O_{k \rightarrow \infty}(\varrho^k) \quad \tilde{\mathbb{P}}^{\theta_*}\text{-a.s.} \quad (\text{A.13})$$



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

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

where

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

By (A.13) and (B-4)(vii), we have

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

With (A.6), this yields (37) in the case where  $C = 0$ . For  $C > 0$ , we further observe that, by (A.8) and (A.2), we have, for all  $\theta_\star \in \Theta$  and  $k \in \mathbb{Z}_+$ ,

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

Then Lemma 14 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 (A.7) in place of (A.8). This yields (37) in the case where  $C > 0$ , which concludes the proof.  $\square$

## Bibliography

- Alexander, C., Lazar, E., 2006. Normal mixture garch (1, 1): Applications to exchange rate modelling. *Journal of Applied Econometrics* 21 (3), 307–336.
- Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* 31, 307–327.
- Bollerslev, T., September 2008. Glossary to arch (garch). Tech. rep., CREATES Research Paper.
- Cox, D., 1981. Statistical analysis of time-series: some recent developments. *Scand. J. Statist.* 8 (2), 93–115.
- Davis, R., Dunsmuir, W., Streett, S., DEC 2003. Observation-driven models for Poisson counts. *Biometrika* 90 (4), 777–790.

- Davis, R., Liu, H., 2012. Theory and inference for a class of observation-driven models with application to time series of counts. Preprint, arXiv:1204.3915.
- Douc, R., Doukhan, P., Moulines, E., 2013. Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Processes and Their Applications* 123 (7), 2620–2647.
- Douc, R., Roueff, F., Sim, T., 2014. Equivalent-class consistency of the maximum likelihood estimator for partially-observed markov chains, in preparation.
- Doukhan, P., Fokianos, K., Tjøstheim, D., 2012. On weak dependence conditions for Poisson autoregressions. *Statist. Probab. Lett.* 82 (5), 942–948.
- Fokianos, K., Rahbek, A., Tjøstheim, D., 2009. Poisson autoregression. *J. Am. Statist. Assoc.* 104 (488), 1430–1439, with electronic supplementary materials available online.  
URL <http://dx.doi.org/10.1198/jasa.2009.tm08270>
- Fokianos, K., Tjøstheim, D., 2011. Log-linear poisson autoregression. *J. of Multivariate Analysis* 102 (3), 563–578.
- Haas, M., Mittnik, S., Paolella, M. S., 2004. Mixed normal conditional heteroskedasticity. *Journal of Financial Econometrics* 2 (2), 211–250.
- Henderson, S. G., Matteson, D., Woodard, D., 2011. Stationarity of generalized autoregressive moving average models. *Electronic Journal of Statistics* 5, 800–828.
- Leroux, B. G., 1992. Maximum-likelihood estimation for hidden Markov models. *Stoch. Proc. Appl.* 40, 127–143.
- Meyn, S. P., Tweedie, R. L., 2009. *Markov Chains and Stochastic Stability*. Cambridge University Press, London.
- Moulines, E., Priouret, P., Roueff, F., 2005. On recursive estimation for time varying autoregressive processes. *Ann. Statist.* 33 (6), 2610–2654, available at [arXiv].
- Neumann, M. H., NOV 2011. Absolute regularity and ergodicity of Poisson count processes. *Bernoulli* 17 (4), 1268–1284.
- Pfanzagl, J., 1969. On the measurability and consistency of minimum contrast estimates. *Metrika* 14, 249–272.
- Streett, S., 2000. Some observation driven models for time series of counts. Ph.D. thesis, Colorado State University, Department of Statistics.
- Zhu, F., 2011. A negative binomial integer-valued GARCH model. *J. Time Series Anal.* 32 (1), 54–67.  
URL <http://dx.doi.org/10.1111/j.1467-9892.2010.00684.x>