Electronic Journal of Statistics
Vol. 11 (2017) 2707–2740
ISSN: 1935-7524
DOI: 10.1214/17-EJS1299

Asymptotic properties of quasi-maximum likelihood estimators in observation-driven time series models*

Randal Douc

Department CITI, CNRS UMR 5157, Telecom Sudparis
Evry, France
e-mail: randal.douc@telecom-sudparis.eu

Konstantinos Fokianos

Department of Mathematics & Statistics, University of Cyprus
PO BOX 20537, Nicosia, Cyprus
e-mail: fokianos@ucy.ac.cy

and

Eric Moulines

Department LTCI, CNRS UMR 5141, Telecom Paristech
Paris, France
e-mail: eric.moulines@polytechnique.edu

Abstract: We study a general class of quasi-maximum likelihood estimators for observation-driven time series models. Our main focus is on models related to the exponential family of distributions like Poisson based models for count time series or duration models. However, the proposed approach is more general and covers a variety of time series models including the ordinary GARCH model which has been studied extensively in the literature. We provide general conditions under which quasi-maximum likelihood estimators can be analyzed for this class of time series models and we prove that these estimators are consistent and asymptotically normally distributed regardless of the true data generating process. We illustrate our results using classical examples of quasi-maximum likelihood estimation including standard GARCH models, duration models, Poisson type autoregressions and ARMA models with GARCH errors. Our contribution unifies the existing theory and gives conditions for proving consistency and asymptotic normality in a variety of situations.

MSC 2010 subject classifications: Primary 62M10, 60J05; secondary 62M05.

Keywords and phrases: Asymptotic normality, consistency, count time series, duration models, GARCH models, Kullback-Leibler divergence, maximum likelihood, stationarity.

Received April 2016.

*We would like to thank the Editor, Associate Editor and a reviewer for several useful comments.

Contents

1. Introduction

The aim of this work is to offer a systematic and unified way of studying quasi-maximum likelihood inference for a large class of time series models which are called *observation-driven models*. The terminology was introduced by [9] to signify their main ingredient; the evolution of the observations relies on a hidden process which in turn is driven by some model based dynamics. Observation-driven models can be employed for modeling various types of data including high-frequency financial tick data, epidemiological data, environmental and climate data, to mention only a few of their applications. Their wide applicability is based on the fact that they can accommodate various dependence structures met in practice. Some well-known examples are the GARCH models [6], ARMA-GARCH models (for more see [21] and [42] and the references therein) and duration models. Furthermore, count time series and binary time series models have close connections with the aforementioned models and they are actually covered by the framework we study, see [17].

The success of the observation-driven models stems from the fact that they are based on generalized linear methodology, see [35] and [30]. The combination of likelihood based inference and generalized linear models provide a systematic framework for the analysis of quantitative as well as qualitative time series data. Indeed, estimation, goodness of fit tests, diagnostics and prediction are implemented in a straightforward manner because computations can be carried out using a number of existing software packages. Furthermore, both positive and negative association can be taken into account by a suitable choice of model parametrization.

Observation-driven models are defined as follows. Suppose that $(\mathsf{X}, d_{\mathsf{X}})$ and $(\mathsf{Y}, d_{\mathsf{Y}})$ are two Polish spaces equipped with their Borel sigma-fields $\mathcal{X}$ and $\mathcal{Y}$.

Let (Θ, d_Θ) and (Φ, d_Φ) be two compact metric spaces. Consider $\{Q^\phi\,,\ \phi \in \Phi\}$ a family of Markov kernels on $(\mathsf{X} \times \mathsf{Y}) \times \mathcal{Y}$ indexed by $\phi \in \Phi$. Assume in addition that for all $(\phi, x, y) \in \Phi \times \mathsf{X} \times \mathsf{Y}$, $Q^\phi(x, y; \cdot)$ is dominated by some σ-finite measure μ on $(\mathsf{Y}, \mathcal{Y})$ and denote by $q^\phi(x, y; \cdot)$ its Radon-Nikodym derivative, that is

$$q^\phi(x, y; y') = \frac{\mathrm{d}Q^\phi(x, y; \cdot)}{\mathrm{d}\mu}(y')\ .$$

Assume further that $\left\{(x, y, y') \mapsto f^\theta_{y,y'}(x) : \theta \in \Theta\right\}$ is a family of measurable functions from $(\mathsf{X} \times \mathsf{Y}^2, \mathcal{X} \otimes \mathcal{Y}^{\otimes 2})$ to $(\mathsf{X}, \mathcal{X})$. We denote by $(Y_0, \ldots, Y_n)$ the observed data and we define observation-driven models as follows:

Definition 1.1 (Generalized Observation-Driven model)**.** We say that the process $\{(X_t, Y_t)\,,\ t \in \mathbb{N}\}$ on $((\mathsf{X} \times \mathsf{Y})^{\mathbb{N}}, (\mathcal{X} \otimes \mathcal{Y})^{\otimes \mathbb{N}})$ is a (generalized) *observation-driven time series model* if for all $A \in \mathcal{Y}$,

$$\begin{aligned} &\mathbb{P}^\theta\left[Y_t \in A \;\middle|\; \mathcal{F}_{t-1}\right] = Q^{\varphi(\theta)}(X_{t-1}, Y_{t-1}; A) = \int_A q^{\varphi(\theta)}(X_{t-1}, Y_{t-1}; y)\mu(\mathrm{d}y)\ , \\ &X_t = f^\theta_{Y_{t-1:t}}(X_{t-1})\ , \end{aligned} \tag{1.1}$$

where $\mathcal{F}_t = \sigma(Y_{0:t}, X_{0:t})$, $y_{s:t} = (y_s, \ldots, y_t)$ for $s \leq t$, and $\varphi : \Theta \to \Phi$ is a measurable function from Θ to Φ.

The dependence graph between the various random variables, appearing in equation (1.1), is illustrated in Figure 1. It can be noted that the response Y_t depends on Y_{t-1} and X_{t-1} through the kernel $Q^{\varphi(\theta)}$.

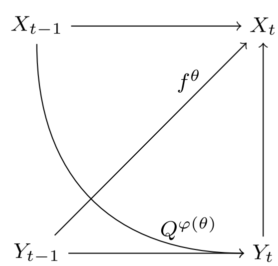

Fig 1. *The dependence graph of a generalized observation-driven model.*

Theoretical study of the properties of these models has been given a great deal of attention in the literature. It is beyond our intention to give a systematic review in this direction. Our primary aim is to study the properties of the Quasi Maximum Likelihood Estimators (QMLE) for estimating the unknown parameter θ. The QMLE is a standard methodology for inference in the class of models introduced by (1.1). Indeed, as described below, Example 2.1 refers to the standard GARCH(1,1) model which is routinely fitted by employing a Gaussian likelihood regardless of the assumed error distributions. Several other examples will be discussed throughout this work, including ARMA-GARCH examples; see Example 2.5.

As a remark, we note that the idea of quasi-likelihood inference was originated by [43] in the context of generalized linear models for independent data, and it was further developed in [30]. It should be noted that the quasi-likelihood is a special case of the methodology of estimating functions; see for example the texts by [23] and [25]. This contribution offers verifiable conditions for obtaining consistent and asymptotically normally distributed quasi-maximum likelihood based estimators for the parameter vector (1.1) even in the case where the kernel Q is misspecified. More precisely, our main results are the following:

- Theorems 3.1 and 4.1 which show strong consistency of the QMLE. In particular, assumption (**A**5) is instrumental on showing the general consistency result toward a set of a parameters given by Theorem 3.1. This fact implies (see Theorem 4.1) the classical strong consistency result toward the "true parameter". The main assumption (**A**5) is verified for several classes of models.
- Theorems 4.2 and 4.3 show asymptotic normality of QMLE; the corresponding assumptions are quite natural and have been extensively used in this context. Once again, assumption (**A**5) is fundamental for linking the martingale methodology to the Taylor expansion of the log-likelihood.

The paper is organized as follows: Section 2 discusses examples of observation-driven models and shows their wide applicability. Section 3 sets up the general notation that is used throughout this work and discusses convergence of the QMLE under model misspecification. Section 4 shows that the asymptotic distribution of the QMLE is normal and discusses conditions under which this fact holds true. All results are applied to the examples of observation-driven models presented in Section 2. Section 5 gives an empirical illustration while the Appendices contain the proofs of our results.

2. Examples of observation-driven models

In classical state-space models, also referred as parameter-driven models, the observations $\{Y_t\,,\ t \in \mathbb{N}\}$ are modelled hierarchically given a hidden process $\{X_t\,,\ t \in \mathbb{N}\}$ which has its own (most often Markovian) dynamic structure, see [28] or [12], for instance. In the Bayesian setting, the process $\{X_t\,,\ t \in \mathbb{N}\}$ may be thought as the dynamical parameter and the distribution of observations is specified conditionally on this parameter. Well-known examples include linear state-space models [28], [44], or hidden Markov models.

Suppose that $\{Y_t\,,\ t \in \mathbb{N}\}$ denotes the observed time series and let $\{X_t\,,\ t \in \mathbb{N}\}$ be an unobserved process. The dichotomy between observation-driven models and parameter-driven models was suggested by [9] who classified these processes according to whether their dynamics are driven by the observed data themselves or by an unobserved process (see also [10]); parameter-driven models are discussed in [44], or [28] and [12] for instance. The generalized observation-driven model, introduced by Definition 1.1, is linked now to several standard examples by identifying suitably the observations and the corresponding latent process.

Example 2.1. *Recall the standard GARCH model ([6])*

$$Y_t = \sigma_{t-1}\epsilon_t \,, \quad \sigma_t^2 = d + a\sigma_{t-1}^2 + bY_t^2 \,, \tag{2.1}$$

where $b > 0$ and $\{\epsilon_t \,,\ t \in \mathbb{N}^*\}$ is a sequence of i.i.d. standard normal random variables. In this example, the latent process X_t is the volatility process σ_t^2 and the conditional distribution of Y_t given $Y_0, \ldots, Y_{t-1}, X_0$ is Gaussian with zero mean and time changing variance X_t. The latent process $\{X_t \,,\ t \in \mathbb{N}\}$ can be recovered by back substitution of the second equation of (2.1)*:*

$$X_t = \sigma_t^2 = d\frac{1-a^t}{1-a} + a^t X_0 + b\sum_{i=0}^{t-1} a^i Y_{t-i}^2 \,,$$

for some starting value X_0. The last display shows that the hidden volatility process is determined by the initial value X_0 and the past observations; this is precisely the reason why model (2.1) *belongs to the class of observation-driven models.*

There are several challenging problems associated with the model specification (2.1)*. In this paper, we will give conditions for obtaining asymptotically normally distributed maximum likelihood based estimators for the parameter vector (d, a, b) when the distribution of $\{\epsilon_t \,,\ t \in \mathbb{N}\}$ is misspecified. For GARCH models, such questions have been addressed by numerous authors including [29], [4], [20], [26], [33], [2], among others. The general framework developed is based on a different point of view which unifies these works.*

The model of Example 2.1 can be extended further by replacing the second equation of (2.1) by a non-linear model, such as $\sigma_t^2 = f^\theta_{Y_t}(\sigma_{t-1}^2)$. Again by repeated substitution, we can express X_t as a function of X_0 and $Y_0, \ldots, Y_{t-1}$, that is $X_t = \sigma_t^2 = f^\theta_{Y_t} \circ \cdots \circ f^\theta_{Y_0}(\sigma_0^2)$. Smooth transition autoregressive models, such as $\sigma_t^2 = f(\sigma_{t-1}^2) + h(Y_t)$ (see [41] and [42]), are examples of non-linear specification of the volatility process.

Considering (2.1), the conditional density of Y_t given $X_{t-1} = x$ is given by $y \mapsto q(x, y) = \frac{1}{\sqrt{x}} g(y/\sqrt{x})$ where g is the density of ϵ_t. Then $X_t = \sigma_t^2$ is updated according to $\sigma_t^2 = f^\theta_{Y_t}(\sigma_{t-1}^2)$ where

$$f^\theta_y(x) = d + ax + by^2 \,, \quad \theta = (d, a, b) \,.$$

More generally, let us define the simplified observation-driven model as follows.

Definition 2.1 (Simplified Observation-Driven model)**.** We say that the process $\{(X_t, Y_t) \,,\ t \in \mathbb{N}\}$ on $((\mathsf{X} \times \mathsf{Y})^{\mathbb{N}}, (\mathcal{X} \otimes \mathcal{Y})^{\otimes \mathbb{N}})$ is a (simplified) *observation-driven time series model* if for all $A \in \mathcal{Y}$,

$$\begin{aligned} &\mathbb{P}\left[Y_t \in A \,\middle|\, \mathcal{F}_{t-1}\right] = Q(X_{t-1}, A) = \int_A q(X_{t-1}, y)\mu(\mathrm{d}y) \,, \\ &X_t = f^\theta_{Y_t}(X_{t-1}) \,, \end{aligned} \tag{2.2}$$

where $\mathcal{F}_t = \sigma(Y_{0:t}, X_{0:t}\}$, $y_{s:t} = (y_s, \dots, y_t)$ for $s \le t$. Recall that Q is a Markov kernel defined on $\mathsf{X} \times \mathcal{Y}$ and dominated by some measure μ on $(\mathsf{Y}, \mathcal{Y})$ with associated transition density $q(x,y) = \mathrm{d}Q(x,\cdot)/\mathrm{d}\mu(y)$ and $(x,y) \mapsto f_y^\theta(x)$ is a measurable function from $\mathsf{X} \times \mathsf{Y}$ to X, which is parameterized by $\theta \in \Theta$.

Remark 2.1. *Shifting the time index of the observations by setting $Z_t = Y_{t-1}$ for $t \ge 1$, we obtain*

$$\begin{aligned}
&\mathbb{P}\left[Z_t \in A \mid \sigma(X_{0:t}, Z_{0:t-1})\right] = Q(X_t, A) = \int_A q(X_t, y)\mu(\mathrm{d}y) \;, \\
&X_{t+1} = f^\theta_{Z_t}(X_t) \;.
\end{aligned}$$

This expression of the transition among variables corresponds to the classical conventions used in the existing time series literature. Nevertheless, the advantage of shifting the time index in (2.2) is that Definition 2.1 clearly generalizes to Definition 1.1, which covers many examples used in practice.

Example 2.2. *A popular class of models that describe time intervals between consecutive observations is that of duration models, see [14]. These models have been found quite useful in financial applications; in particular they have been applied to the analysis and modeling of duration dynamics between trades, as they fit adequately with intraday market activity, see [24], for instance. To be specific, suppose that Y_t denotes the duration between two consecutive observations. Then, a duration model is specified by*

$$Y_t = \psi_{t-1}\epsilon_t \;, \quad \psi_t = d + a\psi_{t-1} + bY_t \;, \tag{2.3}$$

where $b > 0$ and $\{\epsilon_t \,,\ t \in \mathbb{N}^*\}$ is a sequence of i.i.d. nonnegative random variables with mean one. Various models fall within the above framework. For specifying the distribution of ϵ_t's, we can employ the exponential distribution with mean one, or alternatively, the Gamma distribution, suitably rescaled to have mean one. Note that model (2.3) is similar to the GARCH models discussed in Example 2.1. The unobserved process X_t is given by ψ_t which is equal to the conditional mean of Y_t given its past. This is quite analogous to GARCH(1,1) where the volatility σ_t^2 is the expected value of Y_t^2 given its past. The recursion (2.3) can be rewritten as an observation-driven transition (2.2) by setting $X_t = \psi_t$ and

$$q(x,y) = \frac{1}{x} g(y/x) \;, \quad f_y^\theta(x) = d + ax + by \;,$$

where g is the density of ϵ_t and $\theta = (d, a, b)$.

For estimating the parameter vector θ, [13] has suggested the use of QMLE by assuming that $\{\epsilon_t \,,\ t \in \mathbb{N}\}$ is a sequence of i.i.d. exponential random variables with mean one. This work includes this specification and gives conditions for obtaining asymptotically normally distributed estimators in the case of model (2.3).

Example 2.3. *Assume that we observe a binary time series $\{Y_t \,,\ t \in \mathbb{N}^*\}$. Let $\lambda_t = \ln(p_t/(1-p_t))$ where $p_t = \mathbb{P}(Y_{t+1} = 1 \mid \mathcal{F}_t)$ and let us consider the* *following observation-driven model*

$$Y_t = \mathbb{1}\{U_t \le p_{t-1}\}\,, \quad \lambda_t = d + a\lambda_{t-1} + bY_t\,, \tag{2.4}$$

where $b > 0$, $\mathbb{1}\{\cdot\}$ is the indicator function and $\{U_t\,,\ t \in \mathbb{N}^*\}$ is a sequence of i.i.d. standard uniform random variables. (2.4) *introduces an observation-driven model for binary time series where the hidden process X_t is equal to λ_t; see [38], [34]. Recall that for a Bernoulli random variable with success probability p the canonical link is given by the inverse logistic cdf, that is $\ln p/(1-p)$. The logistic model has been widely used in numerous applications. An alternative model is given by the probit link which is defined by means of $\pi_t = \Phi^{-1}(p_t)$, where $\Phi(\cdot)$ is the cdf of the standard normal random variable. For the complete specification of the probit model, we replace λ_t by π_t in the second equation of* (2.4)*; see [46], [39] and [27] among others. More generally, we can consider observation-driven models for binary time series by letting $X_t = F^{-1}(p_t)$, where $F(\cdot)$ is the cdf of continuous random variable.*

Example 2.4. *Several models have been proposed for the analysis of count time series; see [40], [10], [16], [18] and [19], among others. The linear model for the analysis of count time series is based on the specification*

$$Y_t = N_t(\lambda_{t-1})\,, \quad \lambda_t = d + a\lambda_{t-1} + bY_t\,, \tag{2.5}$$

where $b > 0$, $\{N_t\,,\ t \in \mathbb{N}^*\}$ is an i.i.d. sequence of Poisson processes with rate one and $\{\lambda_t\,,\ t \in \mathbb{N}^*\}$ denotes a mean process. In other words, Y_t is equal to the number of events, say $N_t(\lambda_{t-1})$, of $N_t(\cdot)$ in the time interval $[0, \lambda_{t-1}]$. Obviously, the hidden process X_t is equal to λ_t which, in turn, is related linearly to its past values and Y_{t-1}. It turns out that model (2.5) *cannot accommodate negative correlation among consecutive observations and, perhaps more importantly, cannot include time-dependent covariates in a straightforward manner. Based on these issues, [19] suggested a log-linear model of the form*

$$Y_t = N_t(\exp(\nu_{t-1}))\,, \quad \nu_t = d + a\nu_{t-1} + b\ln(1 + Y_t)\,. \tag{2.6}$$

The transformation of the observed process Y_t to the process $\ln(1 + Y_t)$ avoids the issue of zeroes in the data. Note that for this example, the hidden process X_t is ν_t which is equal to $\ln \lambda_t$, in virtue of (2.5)*. This is an example of a canonical link model because the canonical parameter of the Poisson distribution with mean λ is $\nu = \ln \lambda$. Regardless of which model is applied for data analysis, the same remarks made in Example 2.1 are true. In this case we will need to examine the behavior of Maximum Likelihood Estimator (MLE) when the Poisson assumption is not necessarily true for both of the above models. More generally, [8] suggest the use of mixed Poisson models for modeling count time series data, that is* (2.5) *is replaced by*

$$Y_t = N_t(Z_t\lambda_{t-1})\,, \quad \lambda_t = d + a\lambda_{t-1} + bY_t\,, \tag{2.7}$$

where the notation is as before. The sequence $\{Z_t\,,\ t \in \mathbb{N}^*\}$ consists of i.i.d. positive random variables with mean one and it is assumed to be independent of

$\{N_t\,,\ t \in \mathbb{N}^*\}$. *Similarly, we can define a log-linear model following* (2.6)*; details are omitted. For the case of a mixed Poisson process, a likelihood function is not available in general, except for some special cases. For instance, we can obtain a negative binomial distributed time series when the sequence* $\{Z_t\,,\ t \in \mathbb{N}^*\}$ *consists of i.i.d. Gamma random variables with mean one. However, in general it is not possible to obtain a closed-form expression of the likelihood function. The QMLE approach resolves elegantly this problem by utilizing suitably mean and variance specifications; for more see [35] and [30] in the context of generalized linear models.*

The next model we discuss is not a simplified observation-driven model but it can still be covered by our work; in fact this model is a generalized observation-driven time series; see Section 3, for more.

Example 2.5. *An AR(1) model with GARCH(1,1) errors (see [20], for instance) is specified by the following equations*

$$Y_t = \alpha Y_{t-1} + \epsilon_t\,, \quad \epsilon_t = \sigma_{t-1}\eta_t\,, \quad \sigma_t^2 = d + a\sigma_{t-1}^2 + b\epsilon_t^2\,, \tag{2.8}$$

where $\{\eta_t\,,\ t \in \mathbb{Z}\}$ *is an i.i.d. sequence of standard normal random variables. As in the preceding example, the hidden process* $X_t = \sigma_t^2$ *is a function of the initial value* σ_0^2 *and the past observations*

$$\sigma_t^2 = d\frac{1-a^t}{1-a} + a^t\sigma_0^2 + b\sum_{i=0}^{t-1} a^i \Big(Y_{t-i} - \alpha Y_{t-i-1}\Big)^2\,,$$

for some initial value σ_0^2*; thus* (2.8) *belongs to the general class of observation-driven models. The notable difference between* (2.1) *and* (2.8) *is that for the former, the distribution of* Y_t *given* σ_t^2 *does not depend on any additional parameters other than those appearing in the specification of the GARCH model. In contrast, for model* (2.8)*, the distribution of* Y_t *given* σ_t^2 *and* Y_{t-1} *depends on the parameter* α *through the mean of the assumed Gaussian error distribution. More generally (see [32] among others) consider the following class of models*

$$Y_t = m(Y_{t-1};\alpha) + \epsilon_t\,, \quad \epsilon_t = \sigma_{t-1}\eta_t\,, \quad \sigma_t^2 = f^{\lambda}_{Y_{t-1:t}}(\sigma_{t-1}^2)\,, \tag{2.9}$$

where $m(Y_{t-1};\alpha)$ *represents the conditional mean (which depends on an unknown parameter* α*) and the volatility process is modeled by a non-linear model as discussed above. In this example, the complete model depends on the unknown parameter vector* $\theta = (\alpha, \lambda)$*, whereas the distribution of* Y_t *given* σ_t^2 *and* Y_{t-1} *depends only on the parameter* $\phi(\theta) = \alpha$*. Remarks made for* (2.8) *still hold for the case of* (2.9)*. This contribution covers also this class of observation-driven models (recall Definition 1.1) and examines the consequences of misspecifying the likelihood function.*

The above presentation shows the wide applicability of observation-driven models in various scientific fields. Notably, these models can take into account both qualitative and quantitative data in a unified manner. We proceed to study the asymptotic behavior of the QMLE in the next section.

3. General misspecified models

Two models are under consideration in this work: the generalized and the simplified observation-driven time series (see Definitions 1.1 and 2.1, respectively). The dependence graph between the various random variables that appear in these definitions are shown in Figure 2. The case of the simplified model is obviously a particular case of the general model. However, we specify the assumptions required for studying the QMLE in the simplified models framework to avoid confusions and to compare our results with those obtained in the literature.

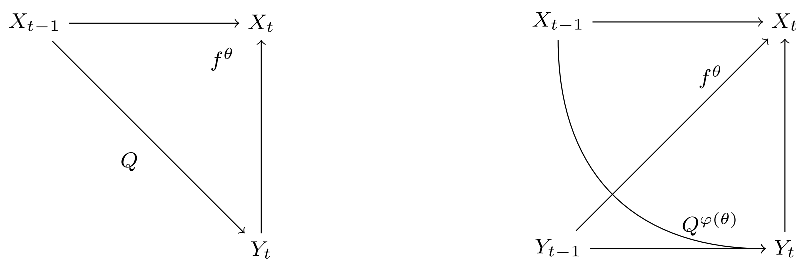

FIG 2. *Simplified observation-driven model (left) and general observation-driven model (right).*

Example 2.5 (Continued). *An example of generalized observation-driven time series model is given by the AR(1)-GARCH model (and its respective non-linear counterpart of (2.9)) discussed in Example 2.5. The conditional distribution of the response Y_t for both cases –for this example it can be assumed to be Gaussian– depends on the autoregressive parameter α and on the hidden process σ_t^2. For the model described in (2.8), it can be easily checked that (1.1) is satisfied with $X_t = \sigma_t^2$ and*

$$q^\alpha(x, y_0; y_1) = \frac{1}{\sqrt{x}} g\left(\frac{y_1 - \alpha y_0}{\sqrt{x}}\right) , \quad f^\theta_{y_{0:1}}(x) = d + ax \ + b(y_1 - \alpha y_0)^2, \quad (3.1)$$

where $g(\cdot)$ is the density of ϵ_t and $\theta = (\alpha, d, a, b)$.

For a generalized observation-driven model, the distribution of $(Y_1, \dots, Y_n)$ given $X_0 = x$ and $Y_0 = y_0$ has a density with respect to the product measure $\mu^{\otimes n}$. It is given by

$$y_{1:n} \mapsto \prod_{t=1}^n q^{\varphi(\theta)}(f^\theta \langle y_{0:t-1} \rangle (x), y_{t-1}; y_t) , \quad (3.2)$$

where we have set for all $t \geq 1$ and all $y_{0:t} \in \mathsf{Y}^{t+1}$,

$$f^\theta \langle y_{0:t} \rangle = f^\theta_{y_{t-1:t}} \circ f^\theta_{y_{t-2:t-1}} \circ \cdots \circ f^\theta_{y_{0:1}} ,$$

with the convention $f^\theta \langle y_0 \rangle (x) = x$. Note that for all $t \geq 0$, X_t is a deterministic function of $Y_{0:t}$ and X_0, i.e.,

$$X_t = f^\theta \langle Y_{0:t} \rangle (X_0) = f^\theta_{Y_{t-1:t}} \circ f^\theta_{Y_{t-2:t-1}} \circ \cdots \circ f^\theta_{Y_{0:1}}(X_0) . \quad (3.3)$$

Now, fix a point x of X. In this section, we focus on the asymptotic properties of $\hat{\theta}_{n,x}$, the conditional Maximum Likelihood Estimator (MLE) of the parameter θ based on the observations $(Y_0, \ldots, Y_n)$ and associated to the parametric family of likelihood functions given in (3.2). In other words, we consider

$$\hat{\theta}_{n,x} \in \operatorname{argmax}_{\theta \in \Theta} \mathsf{L}^{\theta}_{n,x}\langle Y_{0:n}\rangle \;, \tag{3.4}$$

where

$$\mathsf{L}^{\theta}_{n,x}\langle y_{0:n}\rangle := n^{-1} \ln \left(\prod_{t=1}^{n} q^{\varphi(\theta)}(f^{\theta}\langle y_{0:t-1}\rangle(x), y_{t-1}; y_t) \right) \;. \tag{3.5}$$

We are especially interested in the case of *misspecified models*. To be precise, we *do not assume* that the distribution of the observations belongs to the set of distributions where the maximization occurs. In particular, the sequence $\{Y_t\,,\ t \in \mathbb{Z}\}$ does not necessarily correspond to the observation process associated to the recursion (3.3), see [3], [22] and [15]. However, regardless the true data generating process, Theorem 3.1 below shows that the MLE converges to the set of values that minimize the Kullback-Leibler distance between the imposed model and the true model. Before stating the results, some assumptions are needed.

(A1) $\{Y_t,\ t \in \mathbb{Z}\}$ is a strict-sense stationary and ergodic stochastic process.

Under (**A**1), we denote by $\mathbb{P}$ the distribution of $\{Y_t,\ t \in \mathbb{Z}\}$ on $(\mathsf{Y}^{\mathbb{Z}}, \mathcal{Y}^{\mathbb{Z}})$ and by $\mathbb{E}$ the corresponding expectation.

(A2) The functions $(x, \theta) \mapsto q^{\varphi(\theta)}(x, y; y')$ where $(y, y') \in \mathsf{Y}^2$ are fixed and $\theta \mapsto f^{\theta}_{y,y'}(x)$ where $(x, y, y') \in \mathsf{X} \times \mathsf{Y}^2$ are fixed, are continuous.

(A3) There exists a family of $\mathbb{P}$-a.s. finite random variables

$$\left\{ f^{\theta}\langle Y_{-\infty:t}\rangle \;:\; (\theta, t) \in \Theta \times \mathbb{Z} \right\}$$

such that for all $x \in \mathsf{X}$,

(i) the following limit holds $\mathbb{P}$-a.s.,

$$\lim_{m\to\infty} \sup_{\theta\in\Theta} d_{\mathsf{X}}(f^{\theta}\langle Y_{-m:0}\rangle(x), f^{\theta}\langle Y_{-\infty:0}\rangle) = 0 \;,$$

(ii) the following limit holds $\mathbb{P}$-a.s.,

$$\lim_{t\to\infty} \sup_{\theta\in\Theta} |\Delta^{\theta}(f^{\theta}\langle Y_{1:t-1}\rangle(x), f^{\theta}\langle Y_{-\infty:t-1}\rangle, Y_{t-1}, Y_t)| = 0 \;,$$

where $\Delta^{\theta}(x, x', y, y') = \ln q^{\varphi(\theta)}(x, y; y') - \ln q^{\varphi(\theta)}(x', y; y')$.

(iii) $\mathbb{E}\left[\sup_{\theta\in\Theta} \left(\ln q^{\varphi(\theta)}(f^{\theta}\langle Y_{-\infty:-1}\rangle, Y_{-1}; Y_0) \right)_+ \right] < \infty$.

For all $(\theta, t) \in \Theta \times \mathbb{N}$, we set the following notation:

$$\bar{\ell}^{\theta}\langle Y_{-\infty:t}\rangle := \ln q^{\varphi(\theta)}(f^{\theta}\langle Y_{-\infty:t-1}\rangle, Y_{t-1}; Y_t) \; . \tag{3.6}$$

The above assumptions are standard and they are introduced here for facilitating the proof of consistency. Note that under (**A**2), the mapping $\theta \mapsto \mathsf{L}^{\theta}_{n,x}\langle Y_{0:n}\rangle$ is a continuous function on the compact set Θ and thus, the MLE $\hat{\theta}_{n,x}$ obtained by (3.4) is well-defined. Furthermore, under (**A**3)-(i) we obtain, regardless of the initial value of $X_{-m} = x$, that X_0 (and thus X_t) can be approximated by a quantity involving the infinite past of the observations. Assumption (**A**3)-(ii) allows the conditional log-likelihood function to be approximated by a stationary sequence. Furthermore, (**A**3)-(iii) calls for a well-defined maximization problem. Verification of assumption (**A**3) is usually done by introducing the limit, as m tends to infinity, of $f^{\theta}\langle Y_{-m:0}\rangle(x)$ for all fixed $(\theta, x) \in \Theta \times \mathsf{X}$ and by showing that this limit does not depend on x. We can therefore denote it by $f^{\theta}\langle Y_{-\infty:0}\rangle$.

The following theorem establishes the consistency of the sequence of estimators $\{\hat{\theta}_{n,x}\,,\; n \in \mathbb{N}\}$ defined by (3.4) in misspecified models. The proof follows the lines of [11] but the arguments should be adapted to take into account that the kernel density q^{ϕ} here depends on the parameter.

Theorem 3.1. *Suppose that assumptions* (**A***1*)*–*(**A***3*) *hold true. Then, for all* $x \in \mathsf{X}$,

$$\lim_{n\to\infty} d_{\Theta}(\hat{\theta}_{n,x}, \Theta_{\star}) = 0\,, \quad \mathbb{P}\text{-a.s.},$$

where $\Theta_{\star} := \operatorname{argmax}_{\theta\in\Theta} \mathbb{E}[\bar{\ell}^{\theta}\langle Y_{-\infty:1}\rangle]$ *and* $\bar{\ell}^{\theta}\langle Y_{-\infty:1}\rangle$ *is defined by* (3.6)*.*

Proof. The proof directly follows from Theorem A.1 provided that

(a) $\mathbb{E}[\sup_{\theta\in\Theta}(\bar{\ell}^{\theta}\langle Y_{-\infty:0}\rangle)_+] < \infty$,
(b) $\mathbb{P}$-a.s., the function $\theta \mapsto \bar{\ell}^{\theta}\langle Y_{-\infty:0}\rangle$ is upper-semicontinuous,
(c) $\lim_{n\to\infty} \sup_{\theta\in\Theta} |\mathsf{L}^{\theta}_{n,x}\langle Y_{1:n}\rangle - \bar{\mathsf{L}}^{\theta}_{n}\langle Y_{-\infty:n}\rangle| = 0$, $\mathbb{P}$-a.s., where

$$\bar{\mathsf{L}}^{\theta}_{n}\langle Y_{-\infty:n}\rangle = n^{-1} \sum_{k=1}^{n} \bar{\ell}^{\theta}\langle Y_{-\infty:k}\rangle \; .$$

However, (a) follows from (**A**3)-(iii), (b) follows by combining (**A**2) with (**A**3)-(i) since a uniform limit of continuous functions is continuous and (c) is deduced by (**A**3)-(ii) and the definitions of $\mathsf{L}^{\theta}_{n,x}\langle Y_{1:n}\rangle$ and $\bar{\mathsf{L}}^{\theta}_{n}\langle Y_{-\infty:n}\rangle$. The proof is completed. □

Remark 3.1. *Note that, since the model is misspecified, the assumptions of Theorem 3.1 do not imply that for any* $\theta \in \Theta$*, there exists a stationary and ergodic solution to the observation-driven model specified by* $(Q^{\varphi(\theta)}, f^{\theta})$ *(this is not required on our assumptions). Such a property is not always easy to establish but there exists vast literature discussing the conditions under which it holds true. The required condition for* $\{Y_t\}$ *to be stationary and ergodic, depends, in general, upon the model under consideration. For instance, consider* *the GARCH(1,1) model discussed in Example 2.1 with $\{\epsilon_t\,,\ t \in \mathbb{Z}\}$ being an iid sequence of non-degenerate random variables with $\mathbb{E}[\ln^+ \epsilon_t^2] < \infty$ and $d > 0$. If $-\infty \le \gamma := \mathbb{E}[\ln(a + b\epsilon_t^2)] < 0$, then there exists a strict-sense stationary and ergodic process $\{Y_t,\ t \in \mathbb{Z}\}$ obtained from (2.1) with the associated parameter vector $\theta = (d, a, b)$; see [7] and [21, Ch.2] for further references.*

Example 2.1 (Continued). *Assume that the observations $\{Y_t\,,\ t \in \mathbb{Z}\}$ form a strict-sense stationary and ergodic process so that (**A**1) is satisfied. Moreover suppose that $\mathbb{E}[\ln^+ Y_0] < \infty$. We fit to the observations a GARCH(1,1) model of the form (2.1) but with errors $\{\epsilon_t\,,\ t \in \mathbb{Z}\}$ following the Generalized Error Distribution (GED) with density*

$$g_\nu(z) = \frac{\nu \exp\left(-\frac{1}{2}\left|\frac{z}{c}\right|^\nu\right)}{c 2^{(\nu+1)/\nu}\Gamma(\frac{1}{\nu})}\,, \quad z \in \mathbb{R}, \tag{3.7}$$

where $\nu > 0$ and c is a constant which is chosen such that the distribution has zero mean and variance one (see [36]). The parameter ν characterizes the thickness of the tail. When $\nu = 2$, we obtain the standard normal distribution while for $\nu > 2$ (respectively $\nu < 2$) the distribution has thinner (respectively thicker) tails than the normal distribution. The GED distribution is usually employed for GARCH modeling of heavy-tail returns; see the recent work by [15] among others. In this example, we assume that ν is known and the parameter vector is $\theta = (d, a, b) \in \Theta$ which is a compact subset of

$$\left\{(d, a, b)\ :\ d > 0,\ a \ge 0,\ b > 0,\ \mathbb{E}[\ln(a + b\epsilon_t^2)] < 0\right\}\ . \tag{3.8}$$

Note that the above display implies that $a < 1$. Following Remark 3.1, we can also note that under these constraints, there exist stationary and ergodic versions of the process $\{Y_t,\ t \in \mathbb{Z}\}$ in this parametric family.

*We now show that (**A**2) and (**A**3) hold. Set $X_t = \sigma_t^2$. Recall that the recursions given by (2.1) define a simplified and therefore a generalized observation-driven model (1.1) where we have set, with a slight abuse of notation,*

$$q^{\varphi(\theta)}(x, y; y') = q(x, y') = (1/\sqrt{x})g_\nu(y'/\sqrt{x})\,,$$
$$f^\theta_{y,y'}(x) = f^\theta_{y'}(x) = d + ax + b(y')^2\ .$$

*These equations imply clearly that (**A**2) holds. We now turn to (**A**3). Given an initial value of σ_0^2, which will be specified below, the conditional log-likelihood defined in (3.5) may be expressed as*

$$\mathsf{L}^\theta_{n,\sigma_0^2}\langle Y_{1:n}\rangle = n^{-1}\sum_{t=1}^n \Big(-\ln \sigma_{t-1} + \ln g_\nu(Y_t/\sigma_{t-1})\Big)\,, \tag{3.9}$$

where σ_t^2 are computed recursively using (2.1). Note that, since $f_{y,y'}(x) = f_{y'}(x)$ in this particular model, the conditional log-likelihood in (3.9) does not depend on the first observation Y_0 (contrary to the general expression given in (3.5)) and *we therefore write* $\mathsf{L}^{\theta}_{n,\sigma_0^2}\langle Y_{1:n}\rangle$ *instead of* $\mathsf{L}^{\theta}_{n,\sigma_0^2}\langle Y_{0:n}\rangle$. *For a given value of* θ, *the unconditional variance (corresponding to the stationary value of the variance) is usually a sensible choice for the unknown initial value*

$$\sigma_0^2 = \frac{d}{1-a-b} \; .$$

Nevertheless, in what follows, we consider that the initialization is fixed to an arbitrary value $\sigma_0^2 = x$. *For any integer* m, *we have*

$$f^{\theta}\langle Y_{-m:0}\rangle(x) = a^{m+1}x + \sum_{j=0}^{m} a^j (d + bY_{-j}^2) \; .$$

Since Θ *is a compact subset of* (3.8), *there exist* $(a_*, a^*, b_*, b^*) \in (0,1)^2 \times (0,\infty)^2$ *and* $d_*, d^* > 0$ *such that, for all* $\theta = (d,a,b) \in \Theta$,

$$a_* \le a \le a^* \, , \quad b_* \le b \le b^* \, , \quad d_* \le d \le d^* \; .$$

Let $\beta \in (1, (a^*)^{-1/2})$ *and note that, since* $\mathbb{E}[\ln^+(Y_0)] < \infty$,

$$\begin{aligned} \sum_{t=-\infty}^{\infty} \mathbb{P}(|Y_t| > \beta^{|t|}) &\le \sum_{t=-\infty}^{\infty} \mathbb{P}(\ln^+|Y_0|/\ln\beta > |t|) \\ &\le 2(\mathbb{E}[\ln^+|Y_0|/\ln\beta] + 1) < \infty \, , \end{aligned}$$

so that there exists a $\mathbb{P}$-a.s. *finite random variable* M *such that* $|Y_t| \le M\beta^{|t|}$ *for all* $t \in \mathbb{Z}$. *This implies that* $\mathbb{P}$-a.s.,

$$\sum_{j=0}^{\infty} (a^*)^j (d^* + b^* Y_{t-j}^2) \le \frac{d^*}{1-a^*} + \frac{2b^*M^2}{1-a^*\beta^2} < \infty \, ,$$

and we can define

$$f^{\theta}\langle Y_{-\infty:t}\rangle = \sum_{j=0}^{\infty} a^j (d + bY_{t-j}^2) \; .$$

With these definitions, we get

$$\begin{aligned} \sup_{\theta\in\Theta} |f^{\theta}\langle Y_{-m:0}\rangle(x) - f^{\theta}\langle Y_{-\infty:0}\rangle| &\le (a^*)^{m+1}x + \sum_{j=m+1}^{\infty} (a^*)^j (d^* + b^* Y_{-j}^2) \\ &\le (a^*)^{m+1}x + \frac{d^*(a^*)^{m+1}}{1-a^*} + \frac{b^*M^2(a^*\beta^2)^{m+1}}{1-a^*\beta^2} \stackrel{\mathbb{P}\text{-a.s.}}{\longrightarrow} 0 \, , \end{aligned}$$

as $m \to \infty$, *showing therefore* (**A***3*)-(i). *Similarly, we obtain that*

$$\begin{aligned} &\left| \ln q^{\varphi(\theta)}(f^{\theta}\langle Y_{1:t-1}\rangle(x), Y_t) - \ln q^{\varphi(\theta)}(f^{\theta}\langle Y_{-\infty:t-1}\rangle, Y_t) \right| \\ &\qquad \le \frac{1}{2} \left| \ln f^{\theta}\langle Y_{1:t-1}\rangle(x) - \ln f^{\theta}\langle Y_{-\infty:t-1}\rangle \right| \end{aligned}$$

$$+ \frac{|Y_t|^\nu}{2c^\nu} \left| \left(\frac{1}{f^\theta \langle Y_{1:t-1} \rangle (x)} \right)^{\nu/2} - \left(\frac{1}{f^\theta \langle Y_{-\infty:t-1} \rangle} \right)^{\nu/2} \right| .$$

For any $z_1, z_2 > 0$, $|\ln(z_1) - \ln(z_2)| \le |z_1 - z_2| / (z_1 \wedge z_2)$. Moreover, $f^\theta \langle Y_{1:t-1} \rangle (x) \wedge f^\theta \langle Y_{-\infty:t-1} \rangle > d_*$ and the mean value theorem implies that there exists a constant $\gamma > 0$ such that

$$\sup_{\theta \in \Theta} \left| \ln q^{\varphi(\theta)}(f^\theta \langle Y_{1:t-1} \rangle (x), Y_t) - \ln q^{\varphi(\theta)}(f^\theta \langle Y_{-\infty:t-1} \rangle, Y_t) \right|$$
$$\le \gamma(|Y_t|^\nu + 1) \left((a^*)^{t-1} x + \sum_{j=t-1}^{\infty} (a^*)^j (d^* + b^* Y_{t-1-j}^2) \right) .$$

Thus,

$$\sup_{\theta \in \Theta} \left| \ln q^{\varphi(\theta)}(f^\theta \langle Y_{1:t-1} \rangle (x), Y_t) - \ln q^{\varphi(\theta)}(f^\theta \langle Y_{-\infty:t-1} \rangle, Y_t) \right|$$
$$\le \gamma(M^\nu \beta^{\nu t} + 1) \left((a^*)^{t-1} x + \sum_{\ell=0}^{\infty} (a^*)^{t-1+\ell} (d^* + b^* M^2 \beta^{2\ell}) \right) \xrightarrow{\mathbb{P}\text{-}a.s.} 0 ,$$

*as t goes to infinity. This shows (**A**3)-(ii). The proof of (**A**3)-(iii) is along the same lines. Theorem 3.1 then applies and proves the consistency of the estimators $\{\hat{\theta}_{n,x}\}$.*

Example 2.2 (Continued). *Assume, as before, that the observations $\{Y_t\, ,\ t \in \mathbb{Z}\}$ is a strict-sense stationary and ergodic process such that $\mathbb{E}[\ln^+ Y_0] < \infty$. Consider fitting a duration model of the form (2.3) to the observations where the error sequence follows the exponential distribution with mean one, vis.*

$$g(z) = \exp(-z) , \quad z > 0 .$$

Other positive distributions for the sequence $\{\epsilon_t\, ,\ t \in \mathbb{Z}\}$ can be employed; for example the Gamma density suitably normalized to have mean one. However, we discuss the simple case of the exponential distribution for illustrating the verification of the required assumptions. As before, let $\theta = (d, a, b) \in \Theta$ which is assumed to be a compact subset of

$$\{(d, a, b) \,:\, d > 0,\ a > 0,\ b > 0,\ a + b < 1\} .$$

Following Remark 3.1, we can also note that with these constraints, there exists a strictly stationary and ergodic process $\{Y_t\, ,\ t \in \mathbb{Z}\}$ in this parametric family and under some additional assumptions, we can obtain moments of any order for the stationary process (see [31] for more details).

By letting $X_t = \psi_t$, (2.3) defines an observation-driven model with

$$q^{\varphi(\theta)}(x, y; y') = q(x, y') = (1/x) g(y'/x) ,$$

$$f^\theta_{y,y'}(x) = f^\theta_{y'}(x) = d + ax + by' \; .$$

Since $g(\cdot)$ is equal to the exponential density, the conditional log-likelihood may be expressed as

$$\mathsf{L}^\theta_{n,\psi_0}\langle Y_{1:n}\rangle = n^{-1}\sum_{t=1}^n \Big(-\ln\psi_{t-1} - \frac{y_t}{\psi_{t-1}}\Big) \; ,$$

where ψ_t are computed recursively using (2.3) given Y_0 and ψ_0. As before, a typical choice of the initialisation is the stationary mean of the process

$$\psi_0 = \frac{d}{1-a-b} \; ,$$

but in what follows, we consider that the initialization is fixed to an arbitrary value $\psi_0 = x$. Working as before and with the same notation, we obtain that

$$\sup_{\theta\in\Theta} |f^\theta\langle Y_{-m:0}\rangle(x) - f^\theta\langle Y_{-\infty:0}\rangle| \le (a^*)^{m+1}x + \sum_{j=m+1}^\infty (a^*)^j(d^* + b^*Y_{-j}) \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} 0 \; ,$$

*as $m\to\infty$, showing therefore (**A***3*)-(i). Similarly, we obtain that*

$$\begin{aligned}\Big|\ln q^{\varphi(\theta)}(f^\theta\langle Y_{1:t-1}\rangle(x), Y_t) - &\ln q^{\varphi(\theta)}(f^\theta\langle Y_{-\infty:t-1}\rangle, Y_t)\Big| \\ &\le \big|\ln f^\theta\langle Y_{1:t-1}\rangle(x), Y_t) - \ln f^\theta\langle Y_{-\infty:t-1}\rangle, Y_t)\big| \\ &+ Y_t\left|\frac{1}{f^\theta\langle Y_{1:t-1}\rangle(x)} - \frac{1}{f^\theta\langle Y_{-\infty:t-1}\rangle}\right| .\end{aligned}$$

*As in the previous example, we can show that (**A***3*)-(ii) holds true. The proof of (**A***3*)-(iii) is along the same lines and therefore Theorem 3.1 shows that $\{\hat\theta_{n,x}\}$ are consistent.*

We note that Examples 2.3 and 2.4 can be analyzed in a similar way and therefore we omit details.

We now turn to Example 2.5. As previously noted, (2.8) can be put in the framework of observation-driven time series model using (3.1). Hence the assumptions of Theorem 3.1 can be easily checked. Nevertheless, we next focus on the general formulation (2.9) which has been studied by [32] and see how their results are interpreted in our context.

Example 2.5 (Continued). *Consider now (2.9) and suppose again that the observations are realizations of a strict-sense stationary and ergodic process $\{Y_t\,,\ t\in\mathbb{Z}\}$. However, we fit an observation driven model using a Gaussian assumption for the error term $\{\eta_t\,,\ t\in\mathbb{Z}\}$. Suppose that $\theta\in\Theta$ which is assumed to be compact. Given initial values Y_0 and σ_0^2, we obtain the Gaussian quasi-loglikelihood*

$$\mathsf{L}^\theta_{n,\sigma_0^2}\langle Y_{1:n}\rangle = \sum_{t=1}^n \Big(-\frac{1}{2}\ln\sigma^2_{t-1} - \frac{1}{2\sigma^2_{t-1}}(y_t - m(y_{t-1},\alpha))^2\Big) \; .$$

Under some additional assumptions, see C1-C3 in [32, Proposition 1], we obtain that

$$\sup_{\theta\in\Theta} |f^{\theta}\langle Y_{-m:0}\rangle(x) - f^{\theta}\langle Y_{-\infty:0}\rangle| \xrightarrow{\mathbb{P}\text{-}a.s.} 0 ,$$

as $m \to \infty$. Assuming further that σ_t^2 is bounded away from 0, we obtain the consistency of the estimators $\{\hat{\theta}_{n,x}\}$.

4. Quasi-maximum likelihood estimation

When a model has been correctly specified, that is when there exists a parameter $\theta^\star \in \Theta$ such that the data are generated according to this specific process, Theorem 3.1 implies consistency of the MLE to $\theta^\star$ provided that the set $\Theta_\star$ is reduced to the singleton $\{\theta^\star\}$.

An important subclass of misspecified models corresponds to the case where the observation process is assumed to follow the following recursions

$$\begin{aligned} &\mathbb{P}\left[Y_t \in A \,\middle|\, \mathcal{F}_{t-1}\right] = Q^\star(X_{t-1}, Y_{t-1}; A) = \int_A q^\star(X_{t-1}, Y_{t-1}; y)\mu(\mathrm{d}y) ,\\ &X_t = f^{\theta^\star}_{Y_{t-1:t}}(X_{t-1}) , \quad t \in \mathbb{Z} , \end{aligned} \tag{4.1}$$

for any $A \in \mathcal{Y}$, where $\theta^\star$ is supposed to be in Θ^o, the interior of Θ, but

$$q^\star \notin \{q^{\varphi(\theta)} \,:\theta \in \Theta\} .$$

Since $q^\star \neq q^{\varphi(\theta)}$ for any θ, this special case of data generating process falls within the misspecified models framework. Equivalently, we assume that there exists a true parameter $\theta^\star$ such that the second equation of the above display has been correctly specified but the corresponding chosen density $q^\star$ is not equal to the true density associated to the data generating process. This is a standard assumption and has been widely used in practice; for instance parametric inference for GARCH models is most often based on Gaussian log-likelihood.

For this misspecification case, the MLE $\{\hat{\theta}_{n,x}\}$ defined by (3.4) are called QMLE. Note that $\theta^\star$ is not anymore the true value of the parameter in the sense that the distribution of the observation process is not characterized only by $\theta^\star$. Nevertheless, and perhaps surprisingly, it can be shown that, under additional assumptions, the QMLE $\{\hat{\theta}_{n,x}\}$ are consistent and asymptotically normal with respect to the parameter $\theta^\star$. From now on, we assume for simplicity that $\mathsf{X} \subset \mathbb{R}$ and we initially study the consistency property of the QMLE.

4.1. Consistency of the QMLE

Recall that the parameter $\theta^\star$ which appears in the recursion (4.1) satisfies the following assumption:

(A4) The parameter $\theta^\star$ is assumed to be in Θ^o, the interior of Θ.

The main assumption which links the densities $q^\star$ and q^ϕ is the following:

(A5) For all $(x^\star, y) \in \mathsf{X} \times \mathsf{Y}$, the function $\psi_{x^\star,y}$ defined on the set $\Phi \times \mathsf{X}$ by

$$\psi_{x^\star,y}(\phi, x) := \int Q^\star(x^\star, y; \mathrm{d}y') \ln q^\phi(x, y; y') \ , \tag{4.2}$$

has a unique maximum $(\phi, x) = (\varphi(\theta^\star), x^\star)$.

The previous assumption corresponds to an identification condition and is quite analogous to the assumption A3(b) made by [45]. The following theorem shows the consistency of QMLE $\{\hat\theta_{n,x}\}$; its proof is given in the Appendix.

Theorem 4.1. *Assume that assumptions* (**A**1), (**A**2), (**A**3), (**A**4) *and* (**A**5) *hold true. Moreover, assume that* $f^\theta\langle Y_{-\infty:0}\rangle = f^{\theta^\star}\langle Y_{-\infty:0}\rangle$, $\mathbb{P}$-a.s. *implies that* $\theta = \theta^\star$. *Then, for all* $x \in \mathsf{X}$,

$$\lim_{n\to\infty} \hat\theta_{n,x} = \theta^\star \ , \quad \mathbb{P}\text{-a.s.}$$

We now illustrate this result by considering several standard examples of time series models. We first consider the class of *simplified Observation-Driven models* as described in Definition 2.1. This special class of models is characterized by the fact that $q^\phi(x, y; y')$ does not depend on y nor on ϕ and that $f^\theta_{y,y'}(x)$ does not depend on y. Equivalently and with a slight abuse of notation, we assume that

$$\begin{aligned} q^\phi(x, y; y') &= q(x, y') \ , \\ f^\theta_{y,y'}(x) &= f^\theta_{y'}(x) \ . \end{aligned} \tag{4.3}$$

If (4.3) holds then (4.2) reduces to

$$\psi_{x^\star,y}(\phi, x) = \psi_{x^\star}(x) = \int Q^\star(x^\star, \mathrm{d}y') \ln q(x, y') \ ,$$

and assumption (**A**5) is then replaced by the following condition:

(A6) For all $x^\star \in \mathsf{X}$,

$$\mathrm{argmax}_{x\in\mathsf{X}} \int Q^\star(x^\star, \mathrm{d}y) \ln q(x, y) = \{x^\star\} \ .$$

It is worth noting that in the particular case where $Q^{\varphi(\theta)} = Q$ does not depend on θ and is equal to $Q^\star$, we deal with a well-specified model. Since the Kullback-Leibler divergence is nonnegative, we obtain

$$\int Q^\star(x^\star, \mathrm{d}y) \ln q^\star(x, y) \le \int Q^\star(x^\star, \mathrm{d}y) \ln q^\star(x^\star, y) \ ,$$

and, provided that $x \mapsto Q(x, \cdot)$ is a one-to-one mapping, the equality holds if and only if $x = x^\star$. Thus, in well-specified models, (**A**6) is most often satisfied.

Example 2.1 (Continued). *Assume as before that the observations $\{Y_t\,,\ t \in \mathbb{Z}\}$ is a strict-sense stationary ergodic process associated to*

$$\begin{aligned} &\mathbb{P}\left[Y_t \in A \,\middle|\, \mathcal{F}_{t-1}\right] = Q^\star(\sigma^2_{t-1}, A)\,, \quad \text{for any } A \in \mathcal{Y}\,, \\ &\sigma_t^2 = f^{\theta^\star}_{Y_t}(\sigma^2_{t-1})\,, \quad t \in \mathbb{Z}\,, \end{aligned} \tag{4.4}$$

where σ_t^2 is bounded from below. The last display generalizes (2.1) by allowing the volatility process to be a non-linear function of its past values and past values of Y_t. To obtain strict stationarity and ergodicity for model (4.4) it suffices to assume conditions like those reported by [2], for instance. We now assume that (4.4) corresponds to the true data generating process. However, we fit to the observations the following observation-driven model with normal innovations,

$$Y_t = \sigma_{t-1}\epsilon_t\,, \quad \sigma_t^2 = f^{\theta}_{Y_t}(\sigma^2_{t-1})\,, \quad (t,\theta) \in \mathbb{N} \times \Theta\,, \tag{4.5}$$

where $\{\epsilon_t\,,\ t \in \mathbb{N}\})$ is an iid sequence of standard normal random variables. This is a misspecified model; in fact, this approach amounts to using the Gaussian log-likelihood as a quasi log-likelihood function to estimate the parameter θ. By setting $X_t = \sigma_t^2$, we observe that (4.5) corresponds to the case of a simplified observation-driven model given by (4.3) where $q(x,\cdot)$ is the density of a centered normal distribution of variance x. We examine under which conditions assumption ($\mathbf{A}$6) holds true so that a consistent QMLE $\hat\theta_{n,x}$ for θ can be obtained.

For example, when $f^\theta_{Y_t}(\sigma^2_{t-1}) = d + a\sigma^2_{t-1} + bY_t^2$, for $d, b > 0$, this model corresponds to the GARCH(1,1) model with normal innovations discussed in Example 2.1. However, we do not assume any specific expression of $f^\theta_{Y_t}$ in what follows. Now, consider the function

$$\begin{aligned} x \mapsto \int Q^\star(x^\star, \mathrm{d}y) \ln q(x,y) &= \int Q^\star(x^\star, \mathrm{d}y)\left(-\frac{y^2}{2x} - \frac{1}{2}\ln(2\pi x)\right) \\ &= \left(-\frac{\int Q^\star(x^\star, \mathrm{d}y) y^2}{2x} - \frac{1}{2}\ln(2\pi x)\right)\,. \end{aligned}$$

By straightforward algebra, we note that this function is maximized at the point $\int Q^\star(x^\star, \mathrm{d}y)y^2$. We conclude that assumption ($\mathbf{A}$6) is satisfied provided that the condition

$$\int Q^\star(x^\star, \mathrm{d}y) y^2 = x^\star$$

holds true. Plugging this equality into (4.4), we obtain that the observations $\{Y_t\,,\ t \in \mathbb{Z}\}$ is a strict-sense stationary ergodic process associated to

$$\begin{aligned} &Y_t | \mathcal{F}_{t-1} \sim \sigma_{t-1}\epsilon_t^\star \\ &\sigma_t^2 = f^{\theta^\star}_{Y_t}(\sigma^2_{t-1})\,, \quad t \in \mathbb{Z}\,, \end{aligned}$$

where $\{\epsilon^\star_t\,,\ t \in \mathbb{Z}\}$ is an i.i.d. sequence of random variables with potentially any unknown distribution, provided that $\mathbb{E}[(\epsilon_t^\star)^2] = 1$. This is a standard identifiability assumption for GARCH models which implies that Var$[Y_t \mid Y_{0:t-1}] =$

$(1-\mathbb{E}^2[\epsilon_t^\star])\sigma_t^2$. *In particular, we note that* $\mathbb{E}[\epsilon_t^\star]=0$ *is not required for proving consistency.*

For another example within the GARCH models framework, consider again that the true data generating process is given by (4.4), *but we fit model* (4.5) *with errors* $\{\epsilon_t\,,\ t\in\mathbb{N}^*\}$ *following the Laplace distribution with density*

$$g_1(z)=\frac{1}{2}\exp(-|z|)\,,\quad z\in\mathbb{R}\,.$$

Then, working along the previous lines, we derive $\mathbb{E}[|\epsilon_t^\star|]=1$ *as the necessary condition for obtaining consistency of the QMLE. More generally, considering that the error term follows the GED* (3.7), *we obtain that* $\mathbb{E}[|\epsilon_t^\star|^\nu]=1$ *is the necessary condition for proving consistency of the QMLE; see also [21, Sec. 9.2].*

Example 4.1. *This example generalizes Examples 2.2 and 2.4 to the general framework of exponential family models for time series. Let us assume that the observations* $\{Y_t\,,\ t\in\mathbb{Z}\}$ *is a strict-sense stationary ergodic process associated to*

$$\begin{aligned}&\mathbb{P}\left[Y_t\in A\,\middle|\,\mathcal{F}_{t-1}\right]=Q^\star(X_{t-1},A)=\int_A q^\star(X_{t-1},y)\mu(\mathrm{d}y)\,,\quad \text{for any } A\in\mathcal{Y}\,,\\ &X_t=f_{Y_t}^{\theta^\star}(X_{t-1})\,,\quad t\in\mathbb{Z}\,.\end{aligned}$$

We fit to the observations the following observation-driven model

$$\begin{aligned}&\mathbb{P}\left[Y_t\in A\,\middle|\,\mathcal{F}_{t-1}\right]=Q(X_{t-1},A)\,,\quad \text{for any } A\in\mathcal{Y}\,,\\ &X_t=f_{Y_t}^{\theta}(X_{t-1})\,,\quad (t,\theta)\in\mathbb{Z}\times\Theta,\end{aligned}$$

where $Q(x,\cdot)$ *is assumed to belong to the natural exponential family distributions. To be specific, we assume that for all* $(x,y)\in\mathsf{X}\times\mathsf{Y}$,

$$q(x,y)=\exp(yx-\alpha(x))h(y),$$

for some twice differentiable function $\alpha:\mathsf{X}\to\mathbb{R}$ *(which is the cumulant of* Q*) and some measurable function* $h:\mathsf{Y}\to\mathbb{R}^+$. *We investigate conditions under which assumption* (**A**6) *holds true. By noting that*

$$\int Q(x,\mathrm{d}y)\frac{\partial^2\ln q(x,y)}{\partial x^2}\le 0\,,$$

it can be readily checked that $\alpha" \ge 0$ *so that* α *is convex. Therefore, the function*

$$\begin{aligned}x\mapsto\int Q^\star(x^\star,\mathrm{d}y)\ln q(x,y)\quad&=\quad\int Q^\star(x^\star,\mathrm{d}y)\,(yx-\alpha(x)+\ln h(y))\\ &=\quad x\int Q^\star(x^\star,\mathrm{d}y)y-\alpha(x)\\ &+\quad\int Q^\star(x^\star,\mathrm{d}y)\ln h(y)\,,\end{aligned}$$

is concave. The point at which this function achieves its maximum is reduced to a singleton $\{\tilde{x}\}$*, which can be obtained by cancellation of the derivatives with respect to* x*. We obtain that* $\int Q^\star(x^\star, \mathrm{d}y)y - \alpha'(\tilde{x}) = 0$*. Finally,* (**A**6) *is satisfied provided that the condition*

$$\int Q^\star(x^\star, \mathrm{d}y)y = \alpha'(x^\star) \tag{4.6}$$

holds. Because, for the natural exponential family, the function $\alpha(\cdot)$ *corresponds to the cumulant generating function, its first derivative is equal to the mean of* Y_t *given the past* $\mathcal{F}_{t-1}$*. Therefore, the above condition states that the mean function has to be correctly specified regardless of the true data generating process. This fact has been noticed by several authors in the context of longitudinal data analysis (see [47], for example) and in time series modeling; see [48]. However, we show that the right mean specification is a necessary condition for obtaining a consistent QMLE.*

An immediate application of the above fact yields consistency results for duration and count time series models. For instance, recall models (2.5) *and* (2.6)*. Then we obtain*

$$q(x, y) = \exp(yx - \exp(x))/y!,$$

so that $\alpha(x) = \exp(x)$ *which is the mean of* $Q(x, \cdot)$ *under this parametrization. Thus,* (4.6) *yields* $\int Q^\star(x^\star, \mathrm{d}y)y = \exp(x^\star)$ *which implies the following. Suppose that* $\{Y_t\, ,\ t \in \mathbb{Z}\}$ *is any count time series with mean* λ_t *(respectively* $\exp(\nu_t)$*). Then, the QMLE will be consistent for* $\theta^\star$*, provided that the second equations of* (2.5) *(respectively* (2.6)*) has been correctly specified. In particular, recall* (2.7) *for the mixed Poisson count time series models. Then, to obtain a consistent QMLE for* $\theta^\star$*, it suffices to assume that* $\mathbb{E}[Z_t] = 1$ *and the second equation has been correctly specified. Related work on QMLE for count time series models has been recently reported by [1]. These authors established strong consistency of QMLE for count time series models using conditions that imply* (**A**1)–(**A**3) *and* (**A**6) *provided that the mean process* $\lambda_t > d$*. The last condition is trivially satisfied for the case of linear model* (2.5)*. For the case of log-linear model* (2.6)*, this condition can be verified using the results of [19] and [11]. Furthermore [1] establishes asymptotic normality of the QMLE by imposing regularity conditions on the score function and information matrix. Those conditions imply Assumption* (**A**8) *which points to the conclusions of Theorems 4.2 and 4.3.*

In addition, we mention that similar findings are discovered for the simple duration model (2.3)*. In this case, a consistent QMLE for* $\theta^\star$ *is obtained assuming that* $\mathbb{E}[\epsilon_t] = 1$*.*

Example 2.5 (Continued). *Recall the autoregressive model with GARCH noise; for properties of the QMLE for general ARMA-GARCH(p,q) models, see the work by [20] and for the more general model* (2.9) *see [32]. For ease of presentation, we focus on* (2.8)*.*

Assume that the observations $\{Y_t\, ,\ t \in \mathbb{Z}\}$ *is a strict-sense stationary ergodic process associated to*

$$Y_{t+1} = \alpha^\star Y_t + e_t^\star\, , \quad e_t^\star = \sqrt{X_t}\eta_t^\star\, , \quad X_t = d^\star + a^\star X_{t-1} + b^\star (e_{t-1}^\star)^2,$$

where $\{\eta^\star_t\,,\ t\in\mathbb{N}\}$ *is a sequence of i.i.d. random variables with unknown distribution satisfying*

$$\mathbb{E}[\eta^\star_t]=0\,,\quad \mathbb{E}[\{\eta^\star_t\}^2]=1\,.$$

Therefore, $\mathbb{E}\left[Y_{t+1}\,|\,\mathcal{F}_t\right]=\alpha^\star Y_t$ *and* $\mathbb{V}\mathrm{ar}(Y_{t+1}|\mathcal{F}_t)=X_t$*. We fit to the data the following model*

$$Y_{t+1}=\alpha Y_t+e_t\,,\quad e_t=\sqrt{X_t}\eta_t\,,\quad X_t=d+aX_{t-1}+be^2_{t-1},$$

where $b>0$ *and* $\{\eta_t\,,\ t\in\mathbb{N}\}$ *is a sequence of i.i.d. standard normal random variables. As noted in (3.1), this model falls into the class of general observation-driven model because it can be rewritten as*

$$\begin{aligned}
&Y_{t+1}=\alpha Y_t+\sqrt{X_t}\eta_t\sim\mathrm{N}(\alpha Y_t,X_t)=Q^{\varphi(\theta)}(X_t,Y_t;\cdot),\\
&X_{t+1}=d+aX_t+b(Y_{t+1}-\alpha Y_t)^2=f^\theta_{Y_{t:t+1}}(X_t)\,,
\end{aligned}$$

where $\theta=(\alpha,d,a,b)$*,* $\varphi(\theta)=\alpha$ *and* $|\alpha|<1$*. Then, the kernel* Q^a *has a density*

$$q^a(x,y;y')=\exp\left(-\frac{(y'-\alpha y)^2}{2x}-\frac{1}{2}\ln(2\pi x)\right)\,.$$

Now, fix some $y\in\mathsf{Y}$ *and let* ψ_y *be the function*

$$\begin{aligned}
\psi_y(\alpha,x):=&\quad\int Q^\star(x^\star,y;\mathrm{d}y')\ln q^\alpha(x,y;y')\\
=&\quad-\frac{\int Q^\star(x^\star,y;\mathrm{d}y')(y'-\alpha y)^2}{2x}-\frac{1}{2}\ln(2\pi x).
\end{aligned}$$

First note that for all $x\in\mathsf{X}$*,*

$$\mathrm{argmax}_{\alpha\in(0,1)}\psi_y(\alpha,x)=\left\{\int Q^\star(x^\star,y;\mathrm{d}y')y'\Big/y\right\}=\{\alpha^\star\}=\{\varphi(\theta^\star)\},$$

which does not depend on $x\in\mathsf{X}$*. Then, replacing* α *by* $\alpha^\star$ *and maximizing with respect to* x*, we obtain*

$$\mathrm{argmax}_{x\in\mathsf{X}}\psi_y(\alpha^\star,x)=\left\{\int Q^\star(x^\star,y;\mathrm{d}y')(y'-\alpha^\star y)^2\right\}=\{x^\star\}.$$

Because the global maximum of $\psi(\alpha,x)$ *is attained at only one point, namely* $(\varphi(\theta^\star),x^\star)$*, assumption (**A**5) is satisfied.*

4.2. Asymptotic normality of the QMLE for simplified observation-driven models

In this section, we present the asymptotic normality of the QMLE $\hat\theta_{n,x}$ for simplified observation-driven models. We choose to start with this class of models, as defined by (4.3), in order to avoid technicalities and burdensome notation. However, in the next section we develop rigourously all the steps for proving asymptotic normality of the QMLE for general observation-driven models.

We assume that the parameter set Θ is a subset of $\mathbb{R}^{n_\Theta}$. Suppose that for all $y \in \mathsf{Y}$, the function $x \mapsto q(x,y)$ is twice differentiable. For all twice differentiable functions $f : \Theta \to \mathbb{R}$ and all $y \in \mathsf{Y}$, define the following quantities:

$$\chi^\theta(f,y) := \nabla_\theta f(\theta) \frac{\partial \ln q}{\partial x}(f(\theta),y) \;, \tag{4.7}$$

$$\kappa^\theta(f,x,y) := \nabla^2_\theta f(\theta) \frac{\partial \ln q}{\partial x}(f(\theta),y) + \nabla_\theta f(\theta) \nabla_\theta f(\theta)' \frac{\partial^2 \ln q}{\partial x^2}(f(\theta),y) \;. \tag{4.8}$$

These functions appear naturally when differentiating the log-likelihood function $\theta \mapsto \ln q(f(\theta),y)$ with respect to θ. By straightforward algebra we obtain the score function and the Hessian matrix, respectively, as

$$\nabla_\theta \ln q(f(\theta),y) = \chi^\theta(f,y) \;,$$
$$\nabla^2_\theta \ln q(f(\theta),y) = \kappa^\theta(f,y) \;.$$

To prove asymptotic normality, we need the following additional assumptions which are quite standard for maximum likelihood type asymptotics in the framework of time series. More precisely, it is assumed that the score function and the information matrix of the data can be approximated by the infinite past of the process. In addition, all of these quantities are assumed to exist and in particular the Fisher information matrix is not singular. In what follows, the notation $f^\bullet\langle Y_{1:t-1}\rangle(x)$ stands for the function

$$f^\bullet\langle Y_{1:t-1}\rangle(x) : \theta \mapsto f^\theta\langle Y_{1:t-1}\rangle(x) \;.$$

Similarly, $f^\bullet\langle Y_{-\infty:t-1}\rangle$ stands for the function $f^\bullet\langle Y_{-\infty:t-1}\rangle : \theta \mapsto f^\theta\langle Y_{-\infty:t-1}\rangle$.

(A7) For all $y \in \mathsf{Y}$, the function $x \mapsto q(x,y)$ is twice continuously differentiable. Moreover, there exist $\rho > 0$ and a family of $\mathbb{P}$-a.s. finite random variables

$$\left\{ f^\theta\langle Y_{-\infty:t}\rangle \;:\; (\theta,t) \in \Theta \times \mathbb{Z} \right\}$$

such that $f^{\theta^\star}\langle Y_{-\infty:0}\rangle$ is in the interior of X, the function $\theta \mapsto f^\theta\langle Y_{-\infty:0}\rangle$ is, $\mathbb{P}$-a.s., twice continuously differentiable on some ball $\mathsf{B}(\theta^\star,\rho)$ and for all $x \in \mathsf{X}$,

(i) $\mathbb{P}$-a.s. ,

$$\lim_{t\to\infty} \|\chi^{\theta^\star}(f^\bullet\langle Y_{1:t-1}\rangle(x), Y_t) - \chi^{\theta^\star}(f^\bullet\langle Y_{-\infty:t-1}\rangle, Y_t)\| = 0 \;,$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^{n_\Theta}$,

(ii) $\mathbb{P}$-a.s. ,

$$\lim_{t\to\infty} \sup_{\theta\in\mathsf{B}(\theta^\star,\rho)} \|\kappa^\theta(f^\bullet\langle Y_{1:t-1}\rangle(x), Y_t) - \kappa^\theta(f^\bullet\langle Y_{-\infty:t-1}\rangle, Y_t)\| = 0 \;,$$

where by abuse of notation, we use again $\|\cdot\|$ to denote any norm on the set of $n_\Theta \times n_\Theta$-matrices with real entries,

(iii)

$$\mathbb{E}\left[\|\chi^{\theta^\star}(f^\bullet\langle Y_{-\infty:0}\rangle, Y_1)\|^2\right] < \infty\;,$$
$$\mathbb{E}\Big[\sup_{\theta\in\mathsf{B}(\theta^\star,\rho)}\|\kappa^{\theta}(f^\bullet\langle Y_{-\infty:0}\rangle, Y_1)\|\Big] < \infty\;.$$

Moreover, the matrix $\mathcal{J}(\theta^\star)$ defined by

$$\mathcal{J}(\theta^\star) := \mathbb{E}\left[(\nabla_\theta f^{\theta^\star}\langle Y_{-\infty:0}\rangle)(\nabla_\theta f^{\theta^\star}\langle Y_{-\infty:0}\rangle)'\frac{\partial^2 \ln q}{\partial x^2}(f^{\theta^\star}\langle Y_{-\infty:0}\rangle, Y_1)\right] \tag{4.9}$$

is nonsingular.

Theorem 4.2. *Assume* (**A**1), (**A**4), (**A**5) *and* (**A**7) *and suppose that* $\hat\theta_{n,x} \xrightarrow{\mathbb{P}} \theta^\star$. *Then,*

$$\sqrt{n}(\hat\theta_{n,x} - \theta^\star) \overset{D}{\Longrightarrow} \mathrm{N}(0, \mathcal{J}(\theta^\star)^{-1}\mathcal{I}(\theta^\star)\mathcal{J}(\theta^\star)^{-1})\;,$$

where $\mathcal{J}(\theta^\star)$ *is defined in* (4.9) *and* $\mathcal{I}(\theta^\star)$ *is defined by*

$$\mathcal{I}(\theta^\star) := \mathbb{E}\left[(\nabla_\theta f^{\theta^\star}\langle Y_{-\infty:0}\rangle)(\nabla_\theta f^{\theta^\star}\langle Y_{-\infty:0}\rangle)'\left(\frac{\partial \ln q}{\partial x}(\nabla_\theta f^{\theta^\star}\langle Y_{-\infty:0}\rangle, Y_1)\right)^2\right]\;.$$

The proof of Theorem 4.2 follows directly from Theorem 4.3 stated in the next section. We now turn to the case of the asymptotic normality for the QMLE in general observation-driven models.

4.3. Asymptotic normality of the QMLE for general observation-driven models

Obtaining the asymptotic normality of the QMLE for the general observation-driven model proceeds along the previous steps. We will state the main result in this section. For simplicity, assume that $\Phi \subset \mathbb{R}$ and therefore, the function $\theta \mapsto \varphi(\theta)$ takes values on $\mathbb{R}$. If for all $y, y' \in \mathsf{Y}$, $(x,\phi) \mapsto q^\phi(x,y;y')$ is twice continuously differentiable, we can define χ and κ similarly to (4.7) and (4.8). To be specific, for all twice continuously differentiable functions $f, \varphi : \Theta \to \mathbb{R}$ and all $(y,y') \in \mathsf{Y}^2$, define

$$\chi^\theta(f,y,y') := \nabla_\theta f(\theta)\frac{\partial \ln q^{\varphi(\theta)}(f(\theta),y;y')}{\partial x} + \nabla_\theta\varphi(\theta)\frac{\partial \ln q^{\varphi(\theta)}(f(\theta),y;y')}{\partial \phi}\;, \tag{4.10}$$

$$\kappa_1^\theta(f,y,y') := \nabla^2_\theta f(\theta)\frac{\partial \ln q^{\varphi(\theta)}(f(\theta),y;y')}{\partial x} + \nabla^2_\theta\varphi(\theta)\frac{\partial \ln q^{\varphi(\theta)}(f(\theta),y;y')}{\partial \phi}\;. \tag{4.11}$$

$$\kappa_2^\theta(f,y,y') := \nabla_\theta f(\theta)\nabla_\theta f(\theta)'\frac{\partial^2 \ln q^{\varphi(\theta)}(f(\theta),y;y')}{\partial x^2}$$

$$+ (\nabla_\theta f(\theta) \nabla_\theta \varphi(\theta)' + \nabla_\theta \varphi(\theta) \nabla_\theta f(\theta)') \frac{\partial^2 \ln q^{\varphi(\theta)}(f(\theta), y; y')}{\partial \phi \partial x}$$
$$+ \nabla_\theta \varphi(\theta) \nabla_\theta \varphi(\theta)' \frac{\partial^2 \ln q^{\varphi(\theta)}(f(\theta), y; y')}{\partial \phi^2} . \tag{4.12}$$

Moreover, we set

$$\kappa^\theta(f, y, y') = \kappa_1^\theta(f, y, y') + \kappa_2^\theta(f, y, y') . \tag{4.13}$$

As before, these functions correspond to the derivatives of the log–likelihood function

$$\theta \mapsto \ln q^{\varphi(\theta)}(f(\theta), y; y') ,$$

where the function $\theta \mapsto (\varphi(\theta), f(\theta))$ is twice continuously differentiable. It can be readily checked that

$$\nabla_\theta \ln q^{\varphi(\theta)}(f(\theta), y; y') = \chi^\theta(f, y, y') ,$$
$$\nabla_\theta^2 \ln q^{\varphi(\theta)}(f(\theta), y; y') = \kappa^\theta(f, y, y') = \kappa_1^\theta(f, y, y') + \kappa_2^\theta(f, y, y') .$$

For studying the asymptotic normality of the QMLE, assumption (**A**7) is replaced by the following:

(**A8**) For all $y, y' \in \mathsf{Y}$, the functions $(x, \phi) \mapsto q^\phi(x, y; y')$ and $\theta \mapsto \varphi(\theta)$ are twice continuously differentiable and $\varphi(\theta^\star)$ is in the interior of Φ. Moreover, there exist $\rho > 0$ and a family of $\mathbb{P}$-a.s. finite random variables

$$\left\{ f^\theta \langle Y_{-\infty:t} \rangle \,:\, (\theta, t) \in \Theta \times \mathbb{Z} \right\}$$

such that $f^{\theta^\star} \langle Y_{-\infty:0} \rangle$ is in the interior of X, the function $\theta \mapsto f^\theta \langle Y_{-\infty:0} \rangle$ is, $\mathbb{P}$-a.s., twice continuously differentiable on some ball $\mathsf{B}(\theta^\star, \rho)$, and for all $x \in \mathsf{X}$,

(i) $\mathbb{P}$-a.s. ,

$$\lim_{t \to \infty} \|\chi^{\theta^\star}(f^\bullet \langle Y_{1:t-1} \rangle (x), Y_{t-1}, Y_t) - \chi^{\theta^\star}(f^\bullet \langle Y_{-\infty:t-1} \rangle, Y_{t-1}, Y_t)\| = 0,$$

where $\| \cdot \|$ is any norm on $\mathbb{R}^{n_\Theta}$,

(ii) $\mathbb{P}$-a.s. ,

$$\sup_{\theta \in \mathsf{B}(\theta^\star, \rho)} \|\kappa^\theta(f^\bullet \langle Y_{1:t-1} \rangle (x), Y_{t-1}, Y_t) - \kappa^\theta(f^\bullet \langle Y_{-\infty:t-1} \rangle, Y_{t-1}, Y_t)\|,$$

tends to 0, as $t \to \infty$, where we use again $\| \cdot \|$ to denote any norm on the set of $n_\Theta \times n_\Theta$-matrices with real entries,

(iii)

$$\mathbb{E}\left[\|\chi^{\theta^\star}(f^\bullet \langle Y_{-\infty:0} \rangle, Y_0, Y_1)\|^2 \right] < \infty ,$$

$$\mathbb{E}\left[\sup_{\theta\in\mathsf{B}(\theta^\star,\rho)} \|\kappa^\theta(f^\bullet\langle Y_{-\infty:0}\rangle, Y_0, Y_1)\|\right] < \infty \, .$$

Moreover, the matrix $\mathcal{J}(\theta^\star)$ defined by

$$\mathcal{J}(\theta^\star) := \mathbb{E}\left[\kappa_2^{\theta^\star}(f^\bullet\langle Y_{-\infty:0}\rangle, Y_0, Y_1)\right] \tag{4.14}$$

is nonsingular.

Theorem 4.3. *Assume* (**A**1), (**A**4), (**A**5) *and* (**A**8) *and suppose that* $\hat{\theta}_{n,x} \xrightarrow{\mathbb{P}} \theta^\star$. *Then,*

$$\sqrt{n}(\hat{\theta}_{n,x} - \theta^\star) \xRightarrow{D} \mathrm{N}(0, \mathcal{J}(\theta^\star)^{-1}\mathcal{I}(\theta^\star)\mathcal{J}(\theta^\star)^{-1}) \, ,$$

where $\mathcal{J}(\theta^\star)$ *is defined in* (4.14) *and* $\mathcal{I}(\theta^\star)$ *is defined by*

$$\mathcal{I}(\theta^\star) := \mathbb{E}\left[\chi^{\theta^\star}(f^\bullet\langle Y_{-\infty:0}\rangle, Y_0, Y_1)\chi^{\theta^\star}(f^\bullet\langle Y_{-\infty:0}\rangle, Y_0, Y_1)'\right] \, . \tag{4.15}$$

The proof is postponed to the Appendix C.

5. Application

We verify empirically the asymptotic normality of the QMLE. Consider a count time series $\{Y_t\}$ whose true distribution conditional on the past is the geometric distribution with mean process λ_t; in other words set

$$\mathbb{P}[Y_t = y \mid \mathcal{F}_{t-1}] = \frac{1}{\lambda_t + 1}\Big(\frac{\lambda_t}{\lambda_t + 1}\Big)^y, \quad y = 0, 1, 2 \ldots$$

Recall the notation of Equation 2.4 and suppose that the mean process $\{\lambda_t\}$ is defined either by a linear model of the form (2.5) or by a log-linear model of the form (2.6). In this case, the true log–likelihood function is given by

$$\mathsf{L}^{\star\theta}_{n,x}\langle y_{0:n}\rangle = \frac{1}{n}\sum_{t=1}^{n}\Big(y_t \ln\lambda_t(\theta) - (y_t + 1)\ln(1 + \lambda_t(\theta))\Big).$$

However, in practice the true distribution is generally unknown and therefore we choose to use the Poisson distribution as the response distribution. It is easy to check in this case that the "working" likelihood takes on the form

$$\mathsf{L}^{\theta}_{n,x}\langle y_{0:n}\rangle = \frac{1}{n}\sum_{t=1}^{n}\Big(y_t \ln\lambda_t(\theta) - \lambda_t(\theta)\Big).$$

Therefore, the score equations for model (2.5) (equivalently model (2.6)) are given by

$$\chi^\theta(f, y) \quad = \quad \frac{1}{n}\sum_{t=1}^{n}\Big(\frac{y_t}{\lambda_t(\theta)} - 1\Big)\frac{\partial\lambda_t(\theta)}{\partial\theta}$$

$$= \frac{1}{n}\sum_{t=1}^{n}\Big(y_t - \exp(\nu_t(\theta))\Big)\frac{\partial \nu_t(\theta)}{\partial \theta},$$

where the vector of derivatives $\partial\lambda_t(\theta)/\partial\theta$ and $\partial\lambda_t(\theta)/\partial\theta$ can be calculated by recursion.

Table 5 summarizes results of a limited simulation study, where data are generated according to the geometric distribution with linear or log-linear model specification, but with the Poisson distribution is being fitted instead. All results are based on 1000 simulations. Consider the upper panel of the table which corresponds to results obtained after fitting the linear model. In fact, the Table reports the estimates of the parameters obtained by averaging out the results from all simulations. The first two rows correspond to the mean and standard error of the simulated QMLE. In all cases, we see that these estimators approach the true values quite satisfactory. The next three rows show some summary statistics of the sampling distribution of the standardized MLE. In particular, the row p-values, correspond to the p-values obtained from a Kolmogorov-Smirnov test statistic (for testing normality) for the standardized MLE obtained by the simulation. In all cases, we note that the asserted asymptotic normality is quite adequate. The second panel of Table 5 reports results for the log-linear model. We note again that we have quite satisfactory approximation to the true value of the parameter and the normality of the estimators is achieved.

Linear Model						
	$d = 0.50,\ a = 0.20,\ b = 0.40$			$d = 0.20,\ a = 0.20,\ b = 0.50$		
Estimator	0.5140	0.1966	0.4907	0.2051	0.1949	0.4883
SE	0.0790	0.0660	0.0559	0.0321	0.0667	0.0601
Skewness	0.3625	-0.0409	0.2545	0.3327	0.2004	0.0955
Kurtosis	3.3638	3.0360	3.3467	3.4318	2.9884	3.4862
p-value	0.1262	0.9900	0.8785	0.3862	0.2985	0.1802
Loglinear Model						
	$d = 0.50,\ a = 0.20,\ b = 0.40$			$d = 0.10,\ a = -0.20,\ b = 0.40$		
Estimator	0.5196	0.1852	0.3972	0.1011	-0.2121	0.3991
SE	0.1149	0.1007	0.0467	0.0818	0.1660	0.0645
Skewness	0.2909	-0.0345	0.0705	0.1634	0.3704	0.0047
Kurtosis	3.1565	2.9257	2.9621	2.9676	3.5348	3.1178
p-value	0.3684	0.3359	0.9451	0.8016	0.5215	0.8512

Table 1

Results of 1000 simulations obtained after fitting a linear (2.5) and log-linear model (2.6) to a count time series of 1000 observations. Data are generated according to the geometric distribution with mean λ_t but with the Poisson model being fitted instead.

6. Outlook

We have studied a rich class of time series models that have been found quite useful in diverse applications. As the list of references shows, several studies addressed the problem of estimation and inference for observation-driven models by QMLE methodology. This work unifies the existing literature in a coherent and simple way. Furthermore, the methodology can be extended to the case of multivariate data. For instance, consider the so-called vector GARCH model [21, Sec. 11.2.2] which is given by

$$Y_t = \Sigma_{t-1}^{1/2}\epsilon_t\,, \quad \mathrm{vech}(\Sigma_t) = d + A\mathrm{vech}(\Sigma_{t-1}) + B\mathrm{vech}(Y_{t-1}Y_{t-1}^{'})\,,$$

where ϵ_t is an i.i.d. sequence of m-dimensional standard normal random variables, Σ_t is a $m \times m$ positive definite matrix and the notation vech denotes the half-vectorization of an $m \times m$ square matrix C; in other words if $C = (c_{ij})$, then $\mathrm{vech}(C) = (c_{11}, c_{21}, ..., c_{m1}, c_{22}, ..., c_{m2}, ..., c_{mm})^{'}$. Additionally, the vector d is $m(m+1)/2$-dimensional and the matrices A and B are of dimension $\frac{m(m+1)}{2} \times \frac{m(m+1)}{2}$. Comparing the last display with (2.1) we note that the hidden process X_t is equal to Σ_t and the conditional density of Y_t given $X_{t-1} = x$ is given by $y \mapsto q(x, y) = (2\pi)^{-m/2}|x|^{-1/2}\exp(-\frac{1}{2}y^{'}x^{-1}y)$. Similar models can be developed for other classes of processes. The proposed framework advances the theoretical background for both univariate and multivariate observation-driven models and lists easily verifiable conditions for studying the QMLE.

Appendix A: Consistency of max-estimators using stationary approximations

Let X be a Polish space equipped with its Borel sigma-field $\mathcal{X}$. Assume that $(\mathsf{X}^{\mathbb{Z}}, \tilde{\mathcal{B}}, \mathbb{P}, \mathrm{S})$ is a measure-preserving ergodic dynamical system, where $\mathrm{S} : \mathsf{X}^{\mathbb{Z}} \to \mathsf{X}^{\mathbb{Z}}$ denotes the shift operator defined by: for all $x = (x_t)_{t\in\mathbb{Z}}$, $\mathrm{S}(x) = y$ where $y = (y_t)_{t\in\mathbb{Z}}$ and $y_t = x_{t+1}$. Denote by $\mathbb{E}$ the expectation operator associated to $\mathbb{P}$.

Let $(\bar{\ell}^\theta\,,\, \theta \in \Theta)$ be a family of measurable functions $\bar{\ell}^\theta : \mathsf{X}^{\mathbb{Z}} \to \mathbb{R}$, indexed by $\theta \in \Theta$ where (Θ, d) is a compact metric space and denote $\bar{\mathsf{L}}_n^\theta := n^{-1}\sum_{k=0}^{n-1} \bar{\ell}^\theta \circ \mathrm{S}^k$. Moreover, consider $(\mathsf{L}_n^\theta\,,\, n \in \mathbb{N}^*,\, \theta \in \Theta)$ a family of upper-semicontinuous functions $\mathsf{L}_n^\theta : \mathsf{X}^{\mathbb{Z}} \to \mathbb{R}$ indexed by $n \in \mathbb{N}^*$ and $\theta \in \Theta$. Consider the following assumptions:

(C1) $\mathbb{E}\left(\sup_{\theta\in\Theta} \bar{\ell}_+^\theta\right) < \infty$,

(C2) $\mathbb{P}$-a.s., the function $\theta \mapsto \bar{\ell}^\theta$ is upper-semicontinuous,

(C3) $\lim_{n\to\infty} \sup_{\theta\in\Theta} |\mathsf{L}_n^\theta - \bar{\mathsf{L}}_n^\theta| = 0\,, \quad \mathbb{P}$-a.s.

Let $\left\{\bar{\theta}_n \,:\, n \in \mathbb{N}^*\right\} \subset \Theta$ and $\left\{\hat{\theta}_n \,:\, n \in \mathbb{N}^*\right\} \subset \Theta$ such that for all $n \geq 1$,

$$\bar{\theta}_n \in \mathrm{argmax}_{\theta\in\Theta}\bar{\mathsf{L}}_n^\theta\,, \quad \hat{\theta}_n \in \mathrm{argmax}_{\theta\in\Theta}\mathsf{L}_n^\theta\,.$$

Assumptions (**C**1-2) are quite standard and can be adapted directly from [37] (which treated the case of independent sequence $\{X_n\,,\, n \in \mathbb{N}\}$). The statement of the following theorem and the associated proof can be found in [11, Theorem 33].

Theorem A.1. *Assume* (**C**1-2).

(i) Then, $\lim_{n\to\infty} \mathsf{d}(\bar{\theta}_n, \Theta_\star) = 0$, $\mathbb{P}$-a.s., *where* $\Theta_\star := \operatorname{argmax}_{\theta\in\Theta} \mathbb{E}(\bar{\ell}^\theta)$.

(ii) Assume in addition that (**C**3) *holds. Then,* $\lim_{n\to\infty} \mathsf{d}(\hat{\theta}_n, \Theta_\star) = 0$, $\mathbb{P}$-a.s. *Moreover,*

$$\lim_{n\to\infty} \mathsf{L}_n^{\hat{\theta}_n} = \sup_{\theta\in\Theta} \mathbb{E}(\bar{\ell}^\theta)\,, \quad \mathbb{P}\text{-a.s.}$$
$$\textit{for all } \theta \in \Theta, \quad \lim_{n\to\infty} \mathsf{L}_n^{\theta} = \mathbb{E}(\bar{\ell}^\theta)\,, \quad \mathbb{P}\text{-a.s.}$$

Appendix B: Proof of Theorem 4.1

According to Theorem 3.1, we only need to prove that $\Theta_\star$ is reduced to the singleton $\{\theta^\star\}$. By definition of $\bar{\ell}^\theta\langle Y_{-\infty:1}\rangle$ given in (3.6), we have under (**A**5) that for all $\theta \in \Theta$,

$$\begin{aligned}
&\mathbb{E}[\bar{\ell}^\theta\langle Y_{-\infty:1}\rangle] \\
&\quad = \mathbb{E}\Big[\mathbb{E}\left[\ln q^{\varphi(\theta)}(f^\theta\langle Y_{-\infty:0}\rangle, Y_0; Y_1) \,\Big|\, Y_s\,,\ s \le 0\right]\Big] \\
&\quad = \mathbb{E}\left[\int Q^\star(f^{\theta^\star}\langle Y_{-\infty:0}\rangle, Y_0; \mathrm{d}y) \ln q^{\varphi(\theta)}(f^\theta\langle Y_{-\infty:0}\rangle, Y_0; y)\right] \\
&\quad \le \mathbb{E}\left[\int Q^\star(f^{\theta^\star}\langle Y_{-\infty:0}\rangle, Y_0; \mathrm{d}y) \ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{-\infty:0}\rangle, Y_0; y)\right] \\
&\quad = \mathbb{E}[\bar{\ell}^{\theta^\star}\langle Y_{-\infty:1}\rangle]\,.
\end{aligned} \tag{B-1}$$

Moreover, (4.2) also implies that if the equality holds in (B-1), then $f^\theta\langle Y_{-\infty:0}\rangle = f^{\theta^\star}\langle Y_{-\infty:0}\rangle$, $\mathbb{P}$-a.s. which in turn implies that $\theta = \theta^\star$. Thus, $\Theta_\star = \{\theta^\star\}$ and the proof follows.

Appendix C: Proof of Theorem 4.3

For proving Theorem 4.3 we will use the following technical lemmas which establish the asymptotic behavior of the score function and the Hessian matrix.

Lemma C.1. *Assume* (**A**1), (**A**4), (**A**5) *and* (**A**8). *Then,*

$$n^{-1/2} \sum_{t=1}^{n} \nabla_\theta \ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{-\infty:t-1}\rangle, Y_{t-1}; Y_t) \overset{D}{\Longrightarrow} \mathrm{N}(0, \mathcal{I}(\theta^\star))\,,$$

where $\mathcal{I}(\theta^\star)$ *is defined in* (4.15).

Proof. Let $\mathcal{F}$ be the filtration $\mathcal{F} = (\mathcal{F}_n)_{n\in\mathbb{N}}$ where $\mathcal{F}_n = \sigma(Y_s\,,\ s \le n)$ and let

$$M_n := \sum_{t=1}^{n} \nabla_\theta \ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{-\infty:t-1}\rangle, Y_{t-1}; Y_t)$$

$$= \sum_{t=1}^{n} \chi^{\theta^\star}(f^\bullet\langle Y_{-\infty:t-1}\rangle, Y_{t-1}, Y_t) .$$

Note that according to (**A**8)-(iii), $\mathbb{E}(\|M_n\|^2) < \infty$. Moreover, note that under (**A**5), the function $\psi_{x^\star,y}$ defined in (4.2) attains its maximum on $(\varphi(\theta^\star), x^\star)$ and thus, for all $(x^\star, y) \in \mathsf{X} \times \mathsf{Y}$,

$$\frac{\partial \psi_{x^\star,y}}{\partial \phi}(\varphi(\theta^\star), x^\star) = 0 = \frac{\partial \psi_{x^\star,y}}{\partial x}(\varphi(\theta^\star), x^\star) .$$

This implies

$$\begin{aligned}
&\mathbb{E}^{\theta^\star}\left[\nabla_\theta \ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{-\infty:t-1}\rangle, Y_{t-1}; Y_t) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= \nabla_\theta f^{\theta^\star}\langle Y_{-\infty:t-1}\rangle \\
&\qquad \int Q^\star(f^{\theta^\star}\langle Y_{-\infty:t-1}\rangle, Y_{t-1}; \mathrm{d}y) \frac{\partial \ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{-\infty:t-1}\rangle, Y_{t-1}; y)}{\partial x} \\
&\quad + \nabla_\theta \varphi(\theta^\star) \\
&\qquad\quad \int Q^\star(f^{\theta^\star}\langle Y_{-\infty:t-1}\rangle, Y_{t-1}; \mathrm{d}y) \frac{\partial \ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{-\infty:t-1}\rangle, Y_{t-1}; y)}{\partial \phi} \\
&= 0 ,
\end{aligned}$$

where the last equality follows from (4.2). Finally, $\{M_t\,, t \geq 1\}$ is an ergodic (see (**A**1)) and square integrable $\mathcal{F}$-martingale with stationary increments. The proof follows by applying the results of [5]. □

Lemma C.2. *Assume* (**A***4*), (**A***5*) *and* (**A***8*). *Let* $\{\theta_n\,,\ n \in \mathbb{N}\}$ *be a sequence of random vectors such that* $\theta_n \xrightarrow{\mathbb{P}} \theta^\star$. *Then, for all* $i, j \in \{1, \ldots, d\}$,

$$\begin{aligned}
&n^{-1} \sum_{t=1}^{n} \frac{\partial^2 \ln q^{\varphi(\theta_n)}(f^{\theta_n}\langle Y_{-\infty:t-1}\rangle, Y_{t-1}; Y_t)}{\partial \theta_i \partial \theta_j} \\
&\qquad\qquad \xrightarrow{\mathbb{P}} \mathbb{E}\left[\frac{\partial^2 \ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{-\infty:0}\rangle, Y_0; Y_1)}{\partial \theta_i \partial \theta_j}\right] .
\end{aligned}$$

Moreover,

$$\mathcal{J}(\theta^\star) = \left(\mathbb{E}\left[\frac{\partial^2 \ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{-\infty:0}\rangle, Y_0; Y_1)}{\partial \theta_i \partial \theta_j}\right]\right)_{1 \leq i,j \leq d} . \tag{C-1}$$

Proof. Denote

$$A_t(\theta) = \frac{\partial^2 \ln q^{\varphi(\theta)}(f^{\theta}\langle Y_{-\infty:t-1}\rangle, Y_t)}{\partial \theta_i \partial \theta_j} .$$

By the Birkhoff ergodic theorem we have that

$$n^{-1} \sum_{t=1}^{n} A_t(\theta^\star) \xrightarrow{\mathbb{P}} \mathbb{E}[A_1(\theta^\star)] .$$

Hence, we only need to show that

$$n^{-1}\sum_{t=1}^{n}|A_t(\theta^\star)-A_t(\theta_n)|\xrightarrow{\mathbb{P}}0\;. \tag{C-2}$$

Let $\epsilon>0$ and choose $0<\eta<\rho$ such that

$$\mathbb{E}\left(\sup_{\theta\in\mathsf{B}(\theta^\star,\eta)}|A_1(\theta^\star)-A_1(\theta)|\right)<\epsilon\;. \tag{C-3}$$

The existence of such η follows from the $\mathbb{P}$-a.s. continuity of $\theta\mapsto A_t(\theta)$ under (**A**8) and by the Lebesgue convergence theorem under (**A**8)-(iii). We then have

$$\limsup_{n\to\infty}\mathbb{P}\left(n^{-1}\sum_{t=1}^{n}|A_t(\theta^\star)-A_t(\theta_n)|\geq\epsilon\,,\ \theta_n\in\mathsf{B}(\theta^\star,\eta)\right)$$
$$\leq\limsup_{n\to\infty}\mathbb{P}\left(n^{-1}\sum_{t=1}^{n}\sup_{\theta\in\mathsf{B}(\theta^\star,\eta)}|A_t(\theta^\star)-A_t(\theta)|\geq\epsilon\right)=0\;,$$

where the last equality follows from (C-3) and the Birkhoff ergodic theorem. Moreover, since $\hat\theta_n\xrightarrow{\mathbb{P}}\theta^\star$, $\lim_{n\to\infty}\mathbb{P}(\theta_n\notin\mathsf{B}(\theta^\star,\eta))=0$ so that finally,

$$\lim_{n\to\infty}\mathbb{P}\left(n^{-1}\sum_{t=1}^{n}|A_t(\theta^\star)-A_t(\theta_n)|\geq\epsilon\right)=0\;.$$

Thus, equation (C-2) holds and the proof follows. It remains to show (C-1). Since under (**A**5),

$$\begin{aligned}
&\mathbb{E}\left[\kappa_1^{\theta^\star}(f^\bullet\langle Y_{-\infty:0}\rangle,Y_0,Y_1)\right]\\
&\qquad=\mathbb{E}\left[\nabla^2_\theta f^{\theta^\star}\langle Y_{-\infty:0}\rangle\frac{\partial\ln q^{\varphi(\theta^\star)}}{\partial x}(\nabla_\theta f^{\theta^\star}\langle Y_{-\infty:0}\rangle,Y_0;Y_1)\right]\\
&\qquad+\mathbb{E}\left[\nabla^2_\theta\varphi(\theta^\star)\frac{\partial\ln q^{\varphi(\theta^\star)}}{\partial\phi}(\nabla_\theta f^{\theta^\star}\langle Y_{-\infty:0}\rangle,Y_0;Y_1)\right]\\
&\qquad=0,
\end{aligned}$$

we have, using (4.13),

$$\begin{aligned}
\mathcal{J}(\theta^\star)&=\mathbb{E}\left[\kappa^{\theta^\star}(f^\bullet\langle Y_{-\infty:0}\rangle,Y_0,Y_1)\right]\\
&=\left[\mathbb{E}\left(\frac{\partial^2\ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{-\infty:0}\rangle,Y_0;Y_1)}{\partial\theta_i\partial\theta_j}\right)\right]_{1\leq i,j\leq d}\;.
\end{aligned}$$

□

Proof of Theorem 4.3. A Taylor expansion of the score function at $\theta=\theta^\star$ with an integral form of the remainder yields

$$n^{-1/2}\sum_{t=1}^{n}\nabla_\theta\ln q^{\varphi(\hat\theta_{n,x})}(f^{\hat\theta_{n,x}}\langle Y_{1:t-1}\rangle(x),Y_{t-1};Y_t)=0$$

$$
\begin{aligned}
&= n^{-1/2} \sum_{t=1}^{n} \nabla_\theta \ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{1:t-1}\rangle(x), Y_{t-1}; Y_t) \\
&+ n^{-1} \sum_{t=1}^{n} \left(\int_0^1 \nabla_\theta^2 \ln q^{\varphi(\theta_{n,s})}(f^{\theta_{n,s}}\langle Y_{1:t-1}\rangle(x), Y_{t-1}; Y_t) \mathrm{d}s \right) \sqrt{n}(\hat{\theta}_n - \theta^\star) \ ,
\end{aligned}
\tag{C-4}
$$

where $\theta_{n,s} = s\hat{\theta}_{n,x} + (1-s)\theta^\star$, for $s \in (0,1)$. The proof of Theorem 4.2 then follows from (C-4) and the Slutsky Lemma, provided that for all $\theta_n \xrightarrow{\mathbb{P}} \theta^\star$,

$$
\begin{aligned}
&n^{-1/2} \sum_{t=1}^{n} \nabla_\theta \ln q^{\varphi(\theta^\star)}(f^{\theta^\star}\langle Y_{1:t-1}\rangle(x), Y_{t-1}; Y_t) \overset{D}{\Longrightarrow} \mathrm{N}(0, \mathcal{I}(\theta^\star)) \ , \\
&n^{-1} \sum_{t=1}^{n} \frac{\partial^2 \ln q^{\varphi(\theta_n)}(f^{\theta_n}\langle Y_{1:t-1}\rangle(x), Y_{t-1}; Y_t)}{\partial\theta_i \partial\theta_j} \xrightarrow{\mathbb{P}} \mathcal{J}(\theta^\star) \ .
\end{aligned}
$$

However this follows from Lemma C.1 and Lemma C.2. The proof is completed. □

References

[1] Ahmad, A. and Franq, C. (2016). Poisson QMLE of count time series models. *J. Time Ser. Anal.*, 291–314. MR3512959

[2] Bardet, J.-M. and Wintenberger, O. (2010). Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes. *Ann. Statist. 37*, 2730–2759. MR2541445

[3] Berkes, I. and Horváth, L. (2004). The efficiency of the estimators of the parameters in GARCH processes. *Ann. Statist. 32*, 633–655. MR2060172

[4] Berkes, I., Horváth, L., and Kokoszka, P. (2003). GARCH processes: structure and estimation. *Bernoulli 9*, 201–227. MR1997027

[5] Billingsley, P. (1961). The Lindeberg-Levy theorem for martingales. *Proceedings of the American Mathematical Society* **12**, 5, 788–792. MR0126871

[6] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics 31*, 307–327. MR0853051

[7] Bougerol, P. and Picard, N. (1992). Strict stationarity of generalized autoregressive processes. *Ann. Prob.* 20, 1714–1729. MR1188039

[8] Christou, V. and Fokianos, K. (2014). Quasi–likelihood inference for negative binomial time series models. *J. Time Ser. Anal. 35*, 55–78. MR3148248

[9] Cox, D. (1981). Statistical analysis of time-series: some recent developments. *Scand. J. Statist. 8*, 93–115. MR0623586

[10] Davis, R., Dunsmuir, W., and Streett, S. (2003). Observation-driven models for Poisson counts. *Biometrika 90*, 777–790. MR2024757

[11] Douc, R., Doukhan, P., and Moulines, E. (2013). Ergodicity of observation-driven time series models and consistency of the maximum likelihood estimator. *Stochastic Process. Appl. 123*, 2620–2647. MR3054539

[12] Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. CRC Press. MR3289095

[13] Engle, R. F. (2000). The econometrics of ultra-high-frequency data. *Econometrica 68*, 1–22.

[14] Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica 66*, 1127–1162. MR1639411

[15] Fan, J., Qi, L., and Xiu, D. (2014). Quasi-Maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. *J. Bus. Econ. Statist. 32*, 178–191. with discussion. MR3207827

[16] Ferland, R., Latour, A., and Oraichi, D. (2006). Integer-valued GARCH process. *J. Time Ser. Anal. 27*, 923–942. MR2328548

[17] Fokianos, K. (2012). Count time series models. In *Handbook of Statistics: Time Series Analysis–Methods and Applications*, T. S. Rao, S. S. Rao, and C. R. Rao, Eds. Vol. **30**. Elsevier B. V., Amsterdam, 315–347. MR3295420

[18] Fokianos, K., Rahbek, A., and Tjøstheim, D. (2009). Poisson autoregression. *J. Am. Statist. Assoc. 104*, 1430–1439. MR2596998

[19] Fokianos, K. and Tjøstheim, D. (2011). Log-linear poisson autoregression. *J. of Multivariate Analysis 102*, 563–578. MR2755016

[20] Francq, C. and Zakoian, J.-M. (2004). Maximum likelihood estimation of pure garch and arma-garch processes. *Bernoulli 10*, 605–637. MR2076065

[21] Francq, C. and Zakoïan, J.-M. (2010). *GARCH models: Structure, Statistical Inference and Financial Applications*. Wiley, United Kingdom. MR3186556

[22] Francq, C. and Zakoïan, J.-M. (2015). Risk-parameter estimation in volatility models. *Journal of Econometrics 184*, 158–173. MR3277559

[23] Godambe, V. P. (1991). *Estimating Functions*. Oxford Science Publications, Oxford. MR1163993

[24] Hautsch, N. (2012). Univariate multiplicative error models. In *Econometrics of Financial High-Frequency Data*. Springer Berlin Heidelberg, 99–142. MR2848661

[25] Heyde, C. C. (1997). *Quasi-Likelihood and its Applications: A General Approach to Optimal Parameter Estimation*. Springer, New York. MR1461808

[26] Jensen, S. T. and Rahbek, A. (2004). Asymptotic normality of the QMLE estimator of ARCH in the nonstationary case. *Econometrica 72*, 641–646. MR2036735

[27] Kauppi, H. and Saikkonen, P. (2008). Predicting US recessions with dynamic binary response models. *The Review of Economics and Statistics 90*, 777–791.

[28] Koopman, S. J. (1993). Disturbance smoother for state space models. *Biometrika 80*, 117–126. MR1225219

[29] Lee, S. and Hansen, B. (1994). Asymptotic theory for the GARCH(1,1) quasi–maximum likelihood estimator. *Econometric Theory 10*, 29–52. MR1279689

[30] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London. MR3223057

[31] Meitz, M. and Saikkonen, P. (2008). Ergodicity, mixing, and existence of moments of a class of Markov models with application to GARCH and ACD models. *Econometric Theory 24*, 1291–1320. MR2440741

[32] Meitz, M. and Saikkonen, P. (2011). Parameter estimation in nonlinear AR GARCH models. *Econometric Theory 27*, 1236–1278. MR2868839

[33] Mikosch, T. and Straumann, D. (2006). Stable limits of martingale transforms with application to the estimation of GARCH parameters. *Ann. Statist. 34*, 493–522. MR2275251

[34] Moysiadis, T. and Fokianos, K. (2014). On binary and categorical time series models with feedback. *Journal of Multivariate Analysis 131*, 209–228. MR3252645

[35] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A 135*, 370–384.

[36] Nelson, D. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica 59*, 347–370. MR1097532

[37] Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika 14*, 249–272.

[38] Russell, J. R. and Engle, R. F. (2005). A discrete-state continuous-time model of financial transactions prices and times. *Journal of Business and Economic Statistics 23*, 166–180. MR2157268

[39] Rydberg, T. H. and Shephard, N. (2003). Dynamics of trade-by-trade price movements: decomposition and models. *Journal of Financial Econometrics 1*, 2–25.

[40] Streett, S. (2000). Some observation driven models for time series of counts. Ph.D. thesis, Colorado State University, Department of Statistics. MR2701787

[41] Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association 89*, 208–218.

[42] Teräsvirta, T., Tjøstheim, D., and Granger, C. W. J. (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press, Oxford. MR3185399

[43] Wedderburn, R. W. M. (1974). Quasi–likelihood functions, generalized linear models and the Gaussian method. *Biometrika 61*, 439–447. MR0375592

[44] West, M. and Harrison, J. (1989). *Bayesian Forecasting and Dynamic Models*. Springer. MR1020301

[45] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1–25. MR0640163

[46] Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika 75*, 621. MR0995107

[47] Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics 42*, 121–130.

[48] Zeger, S. L. and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics 44*, 1019–1031. MR0980997