# A New Class of Robust Observation-Driven Models*

F. Blasques†, Christian Francq‡and Sébastien Laurent§

September 18, 2020

**Abstract**

This paper introduces a new class of observation-driven models, including score models as a special case. This new class inherits and extends the basic ideas behind the development of score models and addresses a number of unsolved issues in the score literature. In particular, the new class of models (i) allows QML estimation of static parameters, (ii) allows production of leverage effects in the presence of negative outliers, (iii) allows update asymmetry and asymmetric forecast loss functions in the presence of symmetric or skewed innovations, and (iii) achieves out-of-sample outlier robustness in the presence of sub-exponential tails. We establish the asymptotic properties of the QLE, QMLE and MLE as well as likelihood ratio and Lagrange multiplier test statistics. The finite sample properties are studied by means of an extensive Monte Carlo study. Finally, we show the empirical relevance of this new class of models on real data.

*The authors gratefully acknowledge Kris Boudt, Christophe Croux, Paul Embrechts, Patrick Gagliardini, Andrew Harvey, Elvezio Ronchetti, Olivier Scaillet and Bilel Sanhaji for helpful discussions.

†VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. E-mail: f.blasques@vu.nl. Francisco Blasques acknowledges the financial support of the Dutch Science Foundation (NWO) under grant Vidi.195.099.

‡CREST, Institut Polytechnique de Paris and University of Lille E-mail: christian.francq@univ-lille3.fr

§Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS, Aix-Marseille Graduate School of Management – IAE, France. E-mail: sebastien.laurent@univ-amu.fr. Sébastien acknowledges research support by the French National Research Agency Grant ANR-17-EURE-0020.

# 1 Introduction

Generalized autoregressive score ($GAS$) models, also known as dynamic conditional score ($DCS$) models, have been proposed independently by Creal, Koopman and Lucas (2012) and Harvey and Chakravarty (2008).

$GAS$ models provide a general modelling strategy for time series data. Consider a time-series $\{y_t\}_{t\in\mathbb{Z}}$ with conditional density indexed by a time-varying parameter $\{f_t\}_{t\in\mathbb{Z}}$,

$$p_t(y_t, \theta) = p(y_t|f_t, \theta) \quad \forall\ t \in \mathbb{Z}.$$

A $GAS(1,1)$ model for $f_t$ is a model of the form

$$f_{t+1} = \omega + \alpha S(f_t)\frac{\partial \log p(y_t|f_t,\theta)}{\partial f_t} + \beta f_t, \tag{1}$$

where $S(f_t)$ is a scaling function for the score.

There are more than 220 papers referenced on the `gasmodel.com` website that build upon this modelling strategy and have applications in various areas such as default and credit risk modelling, stock volatility and correlation modelling, modelling time-varying dependence structures, CDS spread modelling, systemic risk, and high-frequency data.

The success of $GAS$ models is because (i) these models nest and extend existing observation-driven models such as the GARCH, (ii) their estimation does not require sophisticated techniques (maximum likelihood is the rule), (iii) they constitute a natural way to achieve robustness in the presence of fat-tailed innovations, (iv) statistical inference is standard, and (v) the models usually fit the data quite well.

$GAS$ models have, for instance, received considerable attention in the literature on volatility modelling because when $p_t(y_t, \theta)$ has fatter tails than the Gaussian distribution (e.g., a Student's t-distribution with a finite degree of freedom), $\frac{\partial \log p_t(y_t|f_t,\theta)}{\partial f_t}$ downweights and even bounds the effect of large shocks on the conditional variance; in contrast, in $GARCH$ models, the squared shock is the main driver of the dynamics, irrespective of the choice of the density. This is in line with the empirical literature that suggests that $GARCH$ models may overestimate the conditional volatility for several days or even weeks following very large unexpected shocks (see Lecourt, Laurent and Palm, 2016, among others).

However, it is also clear from (1) that $GAS$ models impose a strong link between the conditional distribution of $y_t$, i.e., $p_t(y_t|f_t, \theta)$, and the updating equation of $f_t$, which is not always desirable. Testing the relevance of these restrictions and eventually relaxing them if they are rejected by the data can therefore be advantageous. In this paper, we keep the downweighting mechanism of the above $GAS$ model but allow the updating equation of $f_t$ to be disconnected from the den-

sity of the innovations if needed. Our family of models, called $\psi GAS$, therefore encompasses $GAS$ models.

We study the statistical properties of $\psi GAS$ models in the case where the conditional moment of interest depends on some covariates. We also study three estimation methods for this model. Since $\psi GAS$ models disconnect the dynamics in $f_t$ and the density of the innovations, unlike $GAS$ models, they allow considering the estimation of the parameters by QML or by ML with a score function in (1) taken with respect to a conditional density other than the one of $y_t$. We also study the estimation of $\psi GAS$ models using the estimating functions approach, which encompasses the QMLE.

In addition to the fact that these estimators are consistent and asymptotically normal, we also show that likelihood ratio and Lagrange multiplier tests of linear restrictions have the usual $\chi^2$ distribution that offers a strategy to test some restrictions implied by standard $GAS$ models.

We present several examples of $\psi GAS$ models throughout this paper but study in detail the $\psi_T GAS - T$ model, a volatility model extending the $\beta_T GAS$ model of Harvey and Chakravarty (2008). This model relies on a standardized Student's t-density for the innovations and the score of a standardized Student's t-density in the updating equation of the conditional variance but does not restrict the degrees of freedom to be the same. The additional flexibility of this model (over the $\beta_T GAS$) is found to be significant at the 5% significance level in more than one-third of the cases out of more than 400 US stocks.

The rest of the paper is structured as follows. Section 2 presents the $\psi$-filtering equation and the properties of this model. The estimation of this model is studied in Section 3. Section 4 studies in more detail the $\psi_T GAS - T$ model. The small sample properties of the $\psi_T GAS - T$ model as well as some empirical results are reported in Section 5. Finally, Section 6 concludes. All proofs are given in the appendix.

# 2 The $\psi$-filtering equation

## 2.1 Score models and robust estimation

In the robust statistics literature, the criterion function of an M-estimator is usually called the *"$\rho$ function"*. The shape of the criterion $\rho$ defines the robustness properties of the estimator. Well-known examples include the *quadratic* $\rho$, the *absolute-error* $\rho$, the *Winsorizing* $\rho$, the *censorizing* $\rho$, and the *biweight* $\rho$. When $\rho$ is the conditional log-likelihood of the data, we obtain the MLE. Additionally, the derivative of $\rho$ with respect to the parameter of interest is usually known in the robust statistics literature as the *"$\psi$ function"*. The $\psi$ function defines implicitly the z-estimator counterpart of the M-estimator. When $\rho$ is a log-likelihood, the $\psi$

function is the *score*.

We propose a new class of observation-driven models with an updating equation given by

$$f_{t+1} = \omega + \alpha\psi(y_t, f_t, \theta)S(f_t) + \beta f_t, \quad \text{where} \quad \psi(y_t, f_t, \theta) := \frac{\partial\rho(y_t, f_t, \theta)}{\partial f_t}.$$

Note that when $\rho(y_t, f_t, \theta)$ is the log-likelihood $\log p(y_t|f_t, \theta)$, we obtain the class of score models. When $p(y_t|f_t, \theta)$ is a Student's t-density and $\rho(y_t, f_t, \theta)$ is a Gaussian density, we obtain the $GARCH - T$ model, which is not nested by score models. In this formulation, the QMLE also becomes a naturally viable alternative to MLE. This is in contrast to score models where there is a strong link between the innovation density and the updating equation, which makes it unnatural to use QMLE. We note also that, in our new model formulation, we can formulate updating equations that Winsorize or censorize outliers, regardless of the conditional distribution $p(y_t|f_t, \theta)$. More generally, $\psi GAS$ models allow us to obtain filtering equations that employ many popular loss functions used in the robust statistics. These include the Cauchy–Lorentzian, the Geman–McClure and the Welsch–Leclerc criteria, as well as the generalized Charbonnier and pseudo Huber–Charbonnier loss functions. Additionally, in empirical applications, we can define updating equations for volatility models that incorporate leverage effects even if the conditional density of $y_t$ is symmetric or left-skewed. This stands in sharp contrast to "pure" score models that are unable to deliver an updating equation with a leverage effect when the innovation density is left-skewed. One can also have an asymmetric updating equation that gives greater penalty to over-prediction of conditional means or under-prediction of conditional volatilities (as is common in macro and financial policy) regardless of the conditional distributions of $y_t$. This is impossible in the more restrictive class of score models since $\rho(y_t, f_t, \theta)$ must be equal to $\log p(y_t|f_t, \theta)$.

Examples 1-3 cover examples of location and volatility filtering involving nonlinear asymmetric criteria as well as fat-tailed and skewed innovations.

EXAMPLE 1. (Leverage effect with left-skewed innovations) *Stock returns are typically heavy tailed and left-skewed. As such, score models of the conditional volatility $y_t = f_t\epsilon_t$ employing asymmetric distributions such as the asymmetric Gaussian or asymmetric Student's t-distribution, may define an updating equation*

$$f_{t+1} = \omega + \alpha s(y_t, f_t, \theta) + \beta f_t,$$

*where the score $s(y_t, f_t, \theta)$ is an asymmetric function of the returns $y_t$ that produces higher volatility for positive returns (i.e., $y_t > 0$) and is more conservative for negative returns (i.e., $y_t < 0$). Unfortunately, this is contrary to the empirical*

*evidence for the 'leverage effect' that predicts higher volatility as a result of negative returns. Depending on the asymmetric Student's t-distribution that is adopted, score models may thus be unable to capture the leverage effect. This issue does not affect the larger class of $\psi GAS$ models*

$$f_{t+1} = \omega + \alpha\psi(y_t, f_t, \theta) + \beta f_t$$

*since the $\psi$ function can adopt nonlinear functional forms independently of the density of the innovations $\epsilon_t$.*

EXAMPLE 2. (Robust count $\psi GAS$ models) *Consider a model for a time series of counts for which the conditional distribution of $y_t$ is Poisson with parameter $f_t$ (Davis et al., 2003, Fokianos et al., 2009). The updating equation for the intensity parameter $f_t$ obtained in the score model framework with conditional variance scaling is linear (Blasques et al, 2014),*

$$f_{t+1} = \omega + \alpha(y_t - f_t) + \beta f_t.$$

*This formulation can be sensitive to large outliers in $y_t$. A strong form of robustness can, however, be achieved in a $\psi GAS$ model with the updating equation determined by an appropriate $\rho$ function. For example, the negative Cauchy–Lorentzian loss function $\rho(y_t, f_t, \theta) = \exp\big(-\frac{1}{2}\frac{(y_t - f_t)^2}{\delta^2}\big) - 1$ delivers*

$$f_{t+1} = \omega + \alpha\delta^{-2}(y_t - f_t)\exp\Big(-\frac{(y_t - f_t)^2}{2\delta^2}\Big) + \beta f_t.$$

*Note that this updating equation is approximately linear at the origin but uniformly bounded in $y_t$ and $f_t$. We recover the linear score update by taking $\delta \to \infty$ and $\alpha \to 0$ such that $\alpha\delta^{-2} \to \alpha^*$.*

EXAMPLE 3. (Asymmetric forecast with symmetric innovations) *Consider a location model where $y_t = f_t + \epsilon_t$ and $\epsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$. The score is symmetric in $\epsilon_t$, and the resulting score updating equation is given by*

$$f_{t+1} = \omega + \alpha\frac{\epsilon_t}{\sigma^2} + \beta f_t.$$

*Asymmetric loss functions are often employed in practical problems that involve forecasting economic time series. The reasoning is simple and practical: as pointed out by Granger (1999), the cost of arriving 10 min early in the airport is quite different from arriving 10 min late. Similarly, Zellner (1986), points out that, in dam construction, an underestimate of the peak water is usually much more serious*

*than an overestimate. Consider the inverse linex forecast loss function introduced in Varian (1975),*

$$\rho(y_t, f_t, \theta) = 1 + \delta\epsilon_t - \exp(\delta\epsilon_t).$$

*This loss function suggests a $\psi$-filtering equation given by*

$$f_{t+1} = \omega + \alpha\delta(\exp(\delta\epsilon_t) - 1) + \beta f_t.$$

## 2.2 Local $\rho$-improvement of the $\psi$-filter

Compared to score models, the additional flexibility of the class of $\psi$-filters may come at some cost. In fact, the results in Blasques et al (2015) provide a reasoning for imposing the restriction of score models that

$$\psi(y_t, f_t, \theta) = \frac{\partial \log p(y_t|f_t, \theta)}{\partial f_t}.$$

Blasques et al (2015) show that only the score filter guarantees that the parameter update from $f_t$ to $f_{t+1}$ produces a local improvement in the log-likelihood of the model and, under appropriate conditions, an improvement in the Kullback-Leibler distance to the true conditional distribution of the data. In particular, Blasques et al (2015) explore the fact that, in regions of high probability, the conditional log-likelihood is improved (i.e., $\log p(y_t, f_t) \leq \log p(y_t, f_{t+1})$) when the update is small $f_{t+1} \approx f_t$ if and only if the parameter update is *score equivalent*. This happens because, under appropriate conditions, the score can be seen as a derivative of a local Kullback-Leibler divergence between the true unknown conditional density $p_t^0$ of $y_t$ given its past $y^{t-1}$, and the conditional density $p(\cdot|f_t)$ implied by the model; i.e., the score term takes the form

$$s_t = \frac{\partial \log p(y_t|f_t)}{\partial f_t} = \lim_{\delta \to 0} \frac{\partial}{\partial f_t} KL_{(y_t,\delta)}\Big( p_t^0,\ p(\cdot|f_t) \Big),$$

where $KL_{(y_t,\delta)}$ is a local Kullback-Leibler divergence that places its mass on a $\delta$-neighbourhood of $y_t$. The $\psi GAS$ model allows for a generalization of this idea whereby $\psi_t$ is a derivative of some local distance function,

$$\psi_t = \lim_{\delta \to 0} \frac{\partial}{\partial f_t} D_{(y_t,\delta)}\Big( p_t^0,\ p(\cdot|f_t) \Big).$$

Proposition 1 highlights the trivial but relevant notion that the $\psi$-update can be used as a Newton-type algorithm when the $\rho$-function is adopted as a filtering objective criterion and the parameter update is smooth. For simplicity, we focus

on updates that resemble a *Newton step* by setting $(\omega, \beta) \approx (0, 1)$. For completeness, a short justification for Proposition 1 is given in the appendix. Definition 1 introduces the notion of *$\psi$-equivalent update* as being an update that always steps in the same direction as the $\psi$-update.

DEFINITION 1. ($\psi$-equivalent update) *A parameter update of the form*

$$f_{t+1} = \omega + \alpha\xi(y_t, f_t, \theta) + \beta f_t,$$

*is said to be $\psi$-equivalent if* $\operatorname{sign}\big(\xi(y, f, \theta)\big) = \operatorname{sign}\big(\psi(y, f, \theta)\big) \;\; \forall \; (y, f, \theta)$.

PROPOSITION 1. (local $\rho$-improvement of $\psi$-updates) *Let $\rho$ be continuously differentiable in $f_t$ and suppose that $(\omega, \beta) \approx (0, 1)$. Then,*

$$\rho(y_t, f_{t+1}, \theta) - \rho(y_t, f_t, \theta) \geq 0 \quad \textit{for every } y_t \in \mathbb{R} \textit{ and } f_{t+1} \approx f_t$$

*if and only if $f_t$ is $\psi$-equivalent. Additionally, let $\rho_\eta$ and $\eta$ be such that*

$$\rho_\eta(\eta(y), f, \theta) = \rho(y, f, \theta) \; \forall \; (f, y, \theta)$$

*with $\rho_\eta$ continuously differentiable in $\eta(y)$ and $\eta(y_{t+1}) \approx \eta(y_t)$. Then,*

$$\rho(y_{t+1}, f_{t+1}, \theta) - \rho(y_t, f_t, \theta) \geq 0 \quad \textit{for every } f_{t+1} \approx f_t$$

*if and only if $f_t$ is $\psi$-equivalent.*

The following two examples illustrate the reasoning behind Proposition 1 on conditional location and scale examples.

EXAMPLE 4. (Location model) *For the location model $y_t = f_t + \epsilon_t$ with the inverse linex forecast loss function, $\rho(y_t, f_t, \theta) = 1 + \delta\epsilon_t - \exp(\delta\epsilon_t)$, Proposition 1 tells us that the $\psi$-update with $\psi(y_t, f_t, \theta) = \delta \exp(\delta\epsilon_t) - \delta$ delivers one-step-ahead local improvements of the inverse linex criterion (i.e., $\rho(y_t, f_{t+1}, \theta) > \rho(y_t, f_t, \theta)$). Furthermore, in this case, we can set $\eta(y_t) = y_t$ and hence conclude that we also improve relative to $y_{t+1}$ (i.e., $\rho(y_{t+1}, f_{t+1}, \theta) > \rho(y_t, f_t, \theta)$) if the data evolve smoothly.*

EXAMPLE 5. (Volatility model) *The same reasoning applies to a volatility model. Here, one might set $\eta(y_t) = y_t^2$ so that the $\psi$-update is ensured to deliver*

$$\rho(y_{t+1}, f_{t+1}, \theta) > \rho(y_t, f_t, \theta)$$

*when both $f_t$ and $y_t^2$ evolve smoothly.*

## 2.3 Stationarity and invertibility of $\psi GAS$ models

The examples of $\psi GAS$ models previously given explain $y_t$ by means of their past values only. It is often the case that some extra information is available, under the form of exogenous covariates, such as transaction volumes, or realized measures based on high frequency intraday data, or even series of other returns. To exploit the extra information conveyed by such covariates, let us investigate the following extension

$$y_t = g(f_t, \epsilon_t), \tag{2}$$
$$f_{t+1} = \omega + \alpha\psi(y_t, X_t, f_t, \theta) + \beta f_t, \tag{3}$$

where $\epsilon_t$ is a random variable that can be interpreted as an error term, $\omega$, $\alpha$, $\beta$ are real parameters and $\theta$ is an element of a parameter space $\Theta \subset \mathbb{R}^p$, $g$ and $\psi$ are differentiable measurable functions and $X_t$ is a vector of exogenous random variables. When there are no exogenous variables, with some abuse of notation, we simply write

$$f_{t+1} = \omega + \alpha\psi(y_t, f_t, \theta) + \beta f_t, \tag{4}$$

instead of (3).

We now give general conditions for stationarity and invertibility of Model (2)-(3). These general conditions will be illustrated on specific examples and on the $\psi_T GAS - T$ model presented in Section 4.

Let $z_t = (\epsilon_t, X_t^\top)^\top \in \mathbb{R}^d$. Note that the time-varying parameter $f_t$ satisfies a Stochastic Recurrence Equation (SRE) of the form

$$f_{t+1} = \varphi(z_t, f_t), \tag{5}$$

where $\varphi : E \times F \to F$ is measurable, and we assume that $E$ is a convex subspace of $\mathbb{R}^d$ and $F$ is an interval.

Lemma 1 details conditions for the $\psi GAS$ model to generate stationary sequences as a data generating process.

LEMMA 1. (Existence of a DGP) *Assume that $(z_t)$ is stationary and ergodic. Suppose that*

*(i)* $\mathbb{E}\log^+ |\psi(g(f^0, \epsilon_t), X_t, f^0, \theta)| < \infty$ *for some constant* $f^0 \in F \subset \mathbb{R}$*;*

*(ii)* $\mathbb{E}\log\sup_f \left|\alpha\frac{\partial\psi(g(f,\epsilon_t),X_t,f,\theta)}{\partial f} + \beta\right| < 0.$

*Then there exist unique strictly stationary and ergodic solutions $\{f_t\}_{t\in\mathbb{Z}}$ and $\{y_t\}_{t\in\mathbb{Z}}$ to Equations (2)-(3).*

The following example gives explicit conditions derived from Lemma 1 for a nonlinear conditional volatility model.

EXAMPLE 6. (Robust asymmetric volatility $\psi GAS$ models) *Consider a volatility model with $y_t = f_t\epsilon_t$, with $\epsilon_t$ i.i.d. and an updating equation defined by an asymmetric Charbonnier loss function,*

$$f_{t+1} = \omega + \alpha\sqrt{(y_t - \delta)^2 + \iota^2} + \beta f_t.$$

*Lemma 1 tells us that both $\{f_t\}_{t\in\mathbb{Z}}$ and $\{y_t\}_{t\in\mathbb{Z}}$ generated by this $\psi GAS$ model are stationary and ergodic when*

$$\mathbb{E}\log\sup_f \left|\alpha\frac{\epsilon_t^2 f - \delta\epsilon_t}{\sqrt{(f\epsilon_t - \delta)^2 + \iota^2}} + \beta\right| < 0.$$

*If, for instance, the distribution of $\epsilon_t$ is Student's t, then the latter condition is implied by the familiar GARCH contraction condition*

$$\mathbb{E}\log|\alpha\epsilon_t + \beta| < 0.$$

*Note that the stationarity condition is unchanged when the updating equation is augmented by an exogenous term of the form $\pi^\top X_t$. The model takes the form* (3) *by setting $\psi(y_t, X_t, f_t, \theta) = \sqrt{(y_t - \delta)^2 + \iota^2} + \pi^\top X_t/\alpha$.*

Lemma 2 states sufficient conditions for data generated by the $\psi GAS$ model to generate data with bounded unconditional moments.

LEMMA 2. (Existence of a marginal moment) *Under the assumptions of Lemma 1, if the sequence $(z_t)$ is i.i.d.,*

$$\mathbb{E}\left|\psi(g(f^0, \epsilon_t), X_t, f^0, \theta)\right|^r < \infty \quad \textit{and} \quad \mathbb{E}\sup_f \left|\frac{\partial\psi(g(f, \epsilon_t), X_t, f, \theta)}{\partial f}\right|^r < \infty$$

*for some $r > 0$, then the stationary solution to Equations (2)-(3) satisfies $\mathbb{E}|f_t|^s < \infty$ for some $s > 0$.*

Assume that, for some $\theta = \theta_0$ satisfying the assumptions of Lemma 1, $(y_t)$ is the stationary solution to (2)-(3) and recall that the time-varying parameter $f_t$ depends on the true but unknown parameter $\theta_0$. For all $\theta$, let us investigate the solutions of the filter

$$f_{t+1}(\theta) = \omega + \alpha\psi(y_t, X_t, f_t(\theta), \theta) + \beta f_t(\theta), \quad t \in \mathbb{Z}, \tag{6}$$

so that $f_t(\theta_0) = f_t$. Note, however, that $f_t(\theta)$ cannot be computed from a finite number of past observations $y_1, \ldots, y_{t-1}$ and $X_1, \ldots, X_{t-1}$. We thus approximate $f_t(\theta)$ by the statistics

$$\widehat{f}_{t+1}(\theta) = \omega + \alpha\psi(y_t, X_t, \widehat{f}_t(\theta), \theta) + \beta\widehat{f}_t(\theta), \quad t \geq 1, \tag{7}$$

with a starting value $\widehat{f}_1(\theta) \in \mathbb{C}(\Theta, F)$, where $\mathbb{C}(\Theta, F)$ denotes the space of the continuous functions from $\Theta$ to $F$.

Lemma 3 gives sufficient conditions for the invertibility of the $\psi GAS$ filter.

LEMMA 3. (Properties of the filter) *Let $\{y_t, X_t\}_{t\in\mathbb{Z}}$ be stationary and ergodic, and suppose that*

*(i) for all $\theta \in \Theta$ there exists $f^0 \in F$ such that $\mathbb{E}\log^+|\psi(y_t, X_t, f^0, \theta)| < \infty$;*

*(ii) $\mathbb{E}\log\sup_{f\in\mathbb{R}}\sup_{\theta\in\Theta}\left|\alpha\frac{\partial\psi(y_t,X_t,f,\theta)}{\partial f} + \beta\right| < 0$.*

*Then, for all $\theta \in \Theta$, there exists a unique strictly stationary and ergodic solution $\{f_t(\theta)\}_{t\in\mathbb{Z}}$ to (6). Furthermore, for all starting functions $\widehat{f}_1(\cdot) \in \mathbb{C}(\Theta, F)$, there exists $\varrho \in (0, 1)$ such that*

$$\varrho^{-t}\sup_{\theta\in\Theta}\left|\widehat{f}_t(\theta) - f_t(\theta)\right| \to 0 \quad \text{a.s. as} \quad t \to \infty. \tag{8}$$

When (8) holds, the model is said to be uniformly invertible. This property will be essential to find a consistent estimator of $\theta_0$ and to approximate the time-varying parameter $f_t$.

The following example verifies the invertibility conditions stated in Lemma 3 in the context of a location model.

EXAMPLE 7. (Location $\psi GAS$ models) *Consider the location model with an updating equation obtained from the 'negative pseudo Huber loss function' of Charbonnier et al (1997) and Hartley and Zisserman (2003), i.e.,*

$$f_{t+1}(\theta) = \omega + \alpha\frac{y_t - f_t(\theta)}{\sqrt{(y_t - f_t(\theta))^2/\delta^2 + 1}} + \beta f_t(\theta).$$

*Lemma 3 tells us that this model is uniformly invertible if*

$$\mathbb{E}\log\sup_{\theta\in\Theta}\sup_{f}|-\alpha(1 + (y_t - f)^2/\delta^2)^{-3/2} + \beta| < 0 \;\Leftarrow\; \sup_{\theta\in\Theta}(|\alpha| + |\beta|) < 1.$$

*Note that the invertibility condition for the filter stands in contrast to the stationarity contraction condition of Lemma 1, which takes the form $|\beta| < 1$ for this model.*

It is necessary to study the first and second derivatives of the filter (6):

$$f'_{t+1}(\theta) := \frac{\partial f_{t+1}(\theta)}{\partial \theta} = A_t + b_t f'_t(\theta), \tag{9}$$

$$f''_{t+1}(\theta) := \text{vec}\left(\frac{\partial^2 f_{t+1}(\theta)}{\partial \theta \partial \theta^\top}\right) = C_t + b_t f''_t(\theta), \tag{10}$$

where

$$\begin{aligned}
A_t &= \frac{\partial \omega}{\partial \theta} + \psi_t \frac{\partial \alpha}{\partial \theta} + \alpha \frac{\partial \psi_t}{\partial \theta} + f_t(\theta)\frac{\partial \beta}{\partial \theta}, \qquad b_t = \alpha \frac{\partial \psi_t}{\partial f} + \beta, \\
C_t &= \text{vec}\left(\frac{\partial^2 \omega}{\partial \theta \partial \theta^\top} + \psi_t \frac{\partial^2 \alpha}{\partial \theta \partial \theta^\top} + \frac{\partial \alpha}{\partial \theta}\frac{\partial \psi_t}{\partial \theta^\top} + \frac{\partial \psi_t}{\partial f}\frac{\partial \alpha}{\partial \theta}(f'_t)^\top \right. \\
&\quad + \frac{\partial \psi_t}{\partial \theta}\frac{\partial \alpha}{\partial \theta^\top} + \alpha \frac{\partial^2 \psi_t}{\partial f \partial \theta}(f'_t)^\top + \alpha \frac{\partial^2 \psi_t}{\partial \theta \partial \theta^\top} + f_t \frac{\partial^2 \beta}{\partial \theta \partial \theta^\top} + \frac{\partial \beta}{\partial \theta}(f'_t)^\top \\
&\quad \left. + \frac{\partial \psi_t}{\partial f} f'_t \frac{\partial \alpha}{\partial \theta^\top} + \alpha f'_t \frac{\partial^2 \psi_t}{\partial f \partial \theta^\top} + f_t \frac{\partial \beta}{\partial \theta^\top} + \alpha \frac{\partial^2 \psi_t}{\partial f^2} f'_t (f'_t)^\top \right),
\end{aligned}$$

with $\psi_t = \psi(y_t, X_t, f_t(\theta), \theta)$ and, using Leibniz's notation,

$$\frac{\partial \psi_t}{\partial \theta} = \left.\frac{\partial \psi(y, X, f, \theta)}{\partial \theta}\right|_{(y,X,f,\theta)=(y_t,X_t,f_t(\theta),\theta)},$$

$$\frac{\partial \psi_t}{\partial f} = \left.\frac{\partial \psi(y, X, f, \theta)}{\partial f}\right|_{(y,X,f,\theta)=(y_t,X_t,f_t(\theta),\theta)}$$

and similar notations for the other derivatives. Assume that $\Theta$ is a compact subspace of $\mathbb{R}^p$ with $p \geq 3$. Without loss of generality, assume that $\theta = (\theta_1, \ldots, \theta_p)'$ with $\theta_1 = \omega$, $\theta_2 = \alpha$ and $\theta_3 = \beta$. Note that the expressions of $A_t$ and $C_t$ then become more explicit because, for instance, $\partial \omega / \partial \theta = (1, 0, \ldots, 0)$. As in (7), we approximate $f'_t(\theta)$ by

$$\widehat{f}'_{t+1}(\theta) = \widehat{A}_t + \widehat{b}_t \widehat{f}'_t(\theta), \quad t \geq 1, \tag{11}$$

with a starting value $\widehat{f}'_1(\theta) \in \mathbb{C}(\Theta, \mathbb{R}^p)$, and where $\widehat{A}_t$ and $\widehat{b}_t$ are obtained by substituting $\widehat{f}_t(\theta)$ for $f_t(\theta)$ in $A_t$ and $b_t$. With similar notations and assumptions, let

$$\widehat{f}''_{t+1}(\theta) = \widehat{C}_t + \widehat{b}_t \widehat{f}''_t(\theta), \quad t \geq 1. \tag{12}$$

Lemma 4 establishes stationarity and invertibility properties for the derivatives of the filter.

LEMMA 4. (Derivatives of the filter) *Let the conditions of Lemma 3 hold, assume that $\psi$ admits continuous second-order derivatives with respect to its last two components, and suppose that*

*(i) for all* $\theta \in \Theta$, $\quad \mathbb{E}\left\{\log^+|\psi_t| + \log^+\left\|\frac{\partial \psi_t}{\partial \theta}\right\| + \log^+\left|\frac{\partial \psi_t}{\partial f}\right| + \log^+|f_t(\theta)|\right\} < \infty.$

*Then, for all* $\theta \in \Theta$, *there exists a unique strictly stationary and ergodic solution* $\{f_t'(\theta)\}_{t\in\mathbb{Z}}$ *to* (9). *If in addition*

*(ii)* $\mathbb{E}\left\{\log^+\left(\sup_f\left|\frac{\partial \psi_t}{\partial f}\right| + \sup_{f,\theta}\left\|\frac{\partial^2 \psi_t}{\partial \theta \partial f}\right\| + \sup_f\left|\frac{\partial^2 \psi_t}{\partial f^2}\right| + \sup_\theta\|f_t'(\theta)\|\right)\right\} < \infty,$

*then, for all starting functions* $\widehat{f}_1(\cdot) \in \mathbb{C}(\Theta, F)$ *and* $\widehat{f}_1'(\cdot) \in \mathbb{C}(\Theta, \mathbb{R}^p)$, *there exists* $\varrho \in (0,1)$ *such that*

$$\varrho^{-t}\sup_{\theta\in\Theta}\left\|\widehat{f}_t'(\theta) - f_t'(\theta)\right\| \to 0 \text{ a.s. as } t \to \infty.$$

*If we further assume*

*(iii) for all* $\theta \in \Theta$, $\quad \mathbb{E}\left\{\log^+\left\|\frac{\partial^2 \psi_t}{\partial \theta \partial \theta^\top}\right\| + \log^+\left\|\frac{\partial^2 \psi_t}{\partial \theta \partial f}\right\| + \log^+\left|\frac{\partial^2 \psi_t}{\partial f^2}\right|\right\} < \infty,$

*then, for all* $\theta \in \Theta$ *there exists a unique strictly stationary and ergodic solution* $\{f_t''(\theta)\}_{t\in\mathbb{Z}}$ *to* (10). *Under the additional assumption*

*(iv)* $\mathbb{E}\left\{\log^+\left(\sup_{f,\theta_i,\theta_j}\left|\frac{\partial^3 \psi_t}{\partial \theta_i \partial \theta_j \partial f}\right| + \sup_{f,\theta_i}\left\|\frac{\partial^3 \psi_t}{\partial \theta_i \partial f^2}\right\| + \sup_f\left|\frac{\partial^3 \psi_t}{\partial f^3}\right|\right)\right\} < \infty,$

*then, for all starting functions* $\widehat{f}_1(\cdot) \in \mathbb{C}(\Theta, F)$, $\widehat{f}_1'(\cdot) \in \mathbb{C}(\Theta, \mathbb{R}^p)$ *and* $\widehat{f}_1''(\cdot) \in \mathbb{C}(\Theta, \mathbb{R}^{p^2})$, *there exists* $\varrho \in (0,1)$ *such that*

$$\varrho^{-t}\sup_{\theta\in\Theta}\left\|\widehat{f}_t''(\theta) - f_t''(\theta)\right\| \to 0 \text{ a.s. as } t \to \infty.$$

The following example considers the derivatives of the filter in the context of a robust $\psi GAS$ conditional volatility model.

EXAMPLE 8. (Robust asymmetric volatility $\psi GAS$ models) *Consider the volatility model introduced in Example 6 with* $y_t = f_t\epsilon_t$ *and* $f_{t+1} = \omega + \alpha\sqrt{(y_t-\delta)^2+\iota^2} + \beta f_t$. *As shown above, the conditions of Lemma 3 are satisfied for this model when* $|\beta| < 1$. *Note that the derivatives of* $\psi(y_t, f, \theta)$ *w.r.t.* $f$ *are zero and the derivatives w.r.t.* $\delta$ *and* $\iota$ *are given by*

$$\frac{\partial \psi_t}{\partial \delta} = -\frac{y_t - \delta}{\sqrt{(\delta - y_t)^2 + \iota^2}}, \quad \frac{\partial^2 \psi_t}{\partial \delta^2} = -\frac{\iota^2}{(\delta^2 - 2\delta y_t + \iota^2 + y_t^2)^{3/2}},$$

$$\frac{\partial^2 \psi_t}{\partial \delta \partial \iota} = -\frac{\iota(y_t-\delta)}{((y_t-\delta)^2+\iota^2)^{3/2}}, \quad \frac{\partial \psi_t}{\partial \iota} = \frac{\iota}{\sqrt{(y_t-\delta)^2+\iota^2}},$$
$$\frac{\partial^2 \psi_t}{\partial \iota^2} = \frac{(y_t-\delta)^2}{(\delta^2-2\delta y_t+\iota^2+y_t^2)^{3/2}}.$$

*Additionally, note that we have already shown that* $\mathbb{E}\log^+\sup_{\theta\in\Theta}|f_t(\theta)|<\infty$ *holds when* $\sup_{\theta\in\Theta}|\beta|<1$ *and* $\mathbb{E}|y_t|^n<\infty$ *for some* $n>0$. *Finally, we have that* $\mathbb{E}\log^+\sup_{\theta\in\Theta}|f_t'(\theta)|<\infty$ *holds since*

$$\sup_t \mathbb{E}\sup_{\theta\in\Theta}\|\widehat{f}'_{t+1}(\theta)\|^n \le \sup_t\sum_{j=0}^{t-1}\sup_{\theta\in\Theta}|\beta|^{nj}\mathbb{E}\sup_{\theta\in\Theta}\|A_{t-j}\|^n + \sup_t\sup_{\theta\in\Theta}|\beta^t|^n\sup_{\theta\in\Theta}\|\widehat{f}'_1(\theta)\|^n$$
$$\le (1-\sup_{\theta\in\Theta}|\beta|^n)^{-1}\mathbb{E}\sup_{\theta\in\Theta}\|A_{t-j}\|^n + \sup_{\theta\in\Theta}|\beta|^n\sup_{\theta\in\Theta}\|\widehat{f}_1(\theta)'\|^n<\infty,$$

*with* $\mathbb{E}\sup_{\theta\in\Theta}\|A_{t-j}\|^n<\infty$ *implied by* $\mathbb{E}|y_t|^n<\infty$ *and* $\mathbb{E}\sup_{\theta\in\Theta}|f_t|^n$ *(shown above) since*

$$A_t = \begin{bmatrix} 1 & \psi_t & f_t & \alpha\dfrac{\partial\psi_t}{\partial\delta} & \alpha\dfrac{\partial\psi_t}{\partial\iota}\end{bmatrix}^\top.$$

*Finally, conditions (i), (ii) and (iii) of Lemma 4 are satisfied. We thus conclude that* $\{\widehat{f}'_t\}_{t\in\mathbb{N}}$ *and* $\{\widehat{f}''_t\}_{t\in\mathbb{N}}$ *converge e.a.s. to unique strictly stationary and ergodic sequences as* $t\to\infty$.

# 3 Estimating the $\psi GAS$ models

In contrast to score models, the $\psi GAS$ models disentangle the parameters involved in $f_t$ of those involved in the conditional distribution $p(\cdot \mid f,\theta)$. We first consider the case where the time-varying parameter of interest is $f_t = f_t(\theta_0)$. It makes sense to estimate $\theta_0$, trying to be as agnostic as possible on the $p$ distribution. This is generally achieved by using QML estimators. Since $\psi GAS$ models are closely related to score functions and, more generally, to z-estimators, it seems natural to estimate the static parameters of a $\psi GAS$ model by means of a z-estimator. In the next section, we consider a class of z-estimators that encompasses the QMLE.

Assume that $(y_t)$ is a stationary process satisfying the $\psi GAS$ models (2)-(3) for some unknown parameter value $\theta=\theta_0$ and $f_t = f_t(\theta_0)$.

## 3.1 The estimating functions approach

To estimate $\theta_0$ using very weak assumptions, the estimating functions theory can be used. This is a general estimation method that has been introduced in the seminal papers of Durbin (1960) and Godambe (1960) and that encompasses moment, likelihood and quasi-likelihood-based techniques (see Chandra and Taniguchi, 2001,

Bera and Bilias, 2002, Heyde, 2008 and the references therein). By extending the Gauss-Markov theorem, Godambe (1960, 1985) developed a concept of optimal estimating function that applies in finite i.i.d. samples, as well as for stochastic processes.

In score models, there generally exist "unbiased estimating functions" $h_t = h_t(\theta_0) \in \mathbb{R}^p$, depending on $y_t$ and $f_t = f_t(\theta_0)$, such that

$$E_{t-1}(h_t) = 0_p,$$

where $E_{t-1}$ denotes the conditional expectation given the sigma-field $\mathcal{F}_{t-1}$ generated by $\{y_s, X_s; s < t\}$. For a location model of the form $y_t = f_t + \epsilon_t$ where $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is i.i.d. with $E\epsilon_t = 0$, or for a duration model of the form $y_t = f_t\epsilon_t$ where $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is i.i.d. positive with $E\epsilon_t = 1$, one can take $h_t(\theta) = y_t - f_t(\theta)$. For a volatility model $y_t = \sqrt{f_t}\epsilon_t$, with standard notation, we can set $h_t(\theta) = y_t^2 - f_t(\theta)$. Obviously, under standard regularity conditions, the score $\partial \log p_t(y_t, \theta_0)/\partial\theta$ is also an unbiased estimating function. An estimator of $\theta_0$ can be obtained by solving an "estimating equation" of the form

$$\sum_{t=1}^{T} a_{t-1} h_t(\widehat{\theta}) = 0_r, \tag{13}$$

where the $r \times p$ matrices $a_t = a_t(\theta) \in \mathcal{F}_t$. Godambe (1985) shows that, within the class of the estimating functions of this form and under mild assumptions, the optimal choice of the $a_t$s is

$$a_{t-1} = E_{t-1}\left(\frac{\partial h_t^\top}{\partial\theta}\right)\left(E_{t-1}h_t h_t^\top\right)^{-1}. \tag{14}$$

According to the terminology of the estimating functions theory, a solution to (13)–(14) is called quasi-likelihood estimator (QLE).

### 3.1.1 Conditional moment estimation

We consider the case where $f_t = E_{t-1}y_t^k$ for some $k > 0$. Location and duration models correspond to $k = 1$, and volatility models to $k = 2$. We thus set $h_t(\theta) = y_t^k - f_t(\theta)$. Of course, in our framework, the estimating function $h_t(\theta)$ is generally not computable because it depends on the unknown values $\{y_t, X_t; t \leq 0\}$. We thus approximate $f_t(\theta)$ by $\widehat{f}_t(\theta)$ in (7) and $\partial f_t(\theta)/\partial\theta$ by $\partial\widehat{f}_t(\theta)/\partial\theta = \widehat{f}_t'(\theta)$ in (11). Let $\widehat{h}_t(\theta) = y_t^k - \widehat{f}_t(\theta)$. Under the assumptions of Lemma 4, we have seen that there exists $\varrho \in (0,1)$ such that

$$\varrho^{-t}\sup_{\theta\in\Theta}\left\{\left|\widehat{f}_t(\theta) - f_t(\theta)\right| + \left\|\frac{\partial\widehat{f}_t(\theta)}{\partial\theta} - \frac{\partial f_t(\theta)}{\partial\theta}\right\|\right\} \to 0 \text{ a.s. as } t\to\infty. \tag{15}$$

Let $\sigma_t^2(\theta) = E_{t-1}h_t^2(\theta)$ be the assumed conditional variance of $y_t^k$ (possibly multiplied by an unimportant non-zero constant). In general, $\sigma_t^2(\theta)$ also depends on the unknown values $\{y_t, X_t; t \leq 0\}$, but we assume that there exists a sequence $\{\widehat{\sigma}_t^2(\theta)\}_{t\in\mathbb{N}}$ computable from $y_1, \ldots, y_t$ and $X_1, \ldots, X_t$ such that

$$\varrho^{-t} \sup_{\theta\in\Theta} \left|\widehat{\sigma}_t^2(\theta) - \sigma_t^2(\theta)\right| \to 0 \text{ a.s. as } t \to \infty. \tag{16}$$

Moreover, assume that there exists a constant $\underline{\sigma}^2 > 0$ such that

$$\inf_{\theta\in\Theta} \left|\sigma_t^2(\theta)\right| > \underline{\sigma}^2 \text{ a.s.} \tag{17}$$

As approximations of (13)–(14), it seems natural to consider the solutions of

$$\widehat{G}_T(\widehat{\theta}) = 0_p, \qquad \widehat{G}_T(\theta) = \frac{1}{T}\sum_{t=t_0+1}^{T} \frac{\widehat{h}_t(\theta)}{\widehat{\sigma}_t^2(\theta)} \frac{\partial \widehat{f}_t(\theta)}{\partial\theta}. \tag{18}$$

The integer $t_0$ is fixed and does not matter for the asymptotic behaviour of the estimator but is expected to attenuate the effect of the (arbitrary) choice of the initial values $\widehat{f}_1(\theta)$ and $\partial\widehat{f}_1(\theta)/\partial\theta$. In practice, one could take $t_0 = 5$ (one week for most daily series), $\widehat{f}_1(\theta) = \sum_{t=1}^{t_0} y_t^k/t_0$ and $\partial\widehat{f}_1(\theta)/\partial\theta = 0_p$.

### 3.1.2 Existence of the estimator

Note that the existence of a solution $\widehat{\theta} \in \Theta$ to (18) is not guaranteed. For instance, consider a location model $y_t = f_t + \epsilon_t$ where $(\epsilon_t)$ is i.i.d. with a mean of 0 and variance $\sigma_\epsilon^2$. If $f_t(\theta) = \omega + \alpha y_{t-1}$ with $\theta_0 = (\omega_0, 0)$ and $\Theta = [\underline{\omega}, \overline{\omega}] \times [0, \overline{\alpha}]$, then with non-zero probability, we have

$$\widehat{G}_T(\theta) = \frac{1}{T}\sum_{t=t_0+1}^{T} \frac{y_t - \omega - \alpha y_{t-1}}{\sigma_\epsilon^2} \begin{pmatrix} 1 \\ y_{t-1} \end{pmatrix} \neq 0_2, \quad \forall\theta \in \Theta.$$

More precisely, when the first component of $\widehat{G}_T(\theta)$ is null and $\sum_t (y_t - \overline{y})(y_{t-1} - \overline{y}) < 0$ (which should be the case with probability of approximately $1/2$ when $\alpha_0 = 0$), the second component of $\widehat{G}_T(\theta)$ is strictly negative for any value of $\alpha \geq 0$. Instead of (18), we thus define a QLE as a measurable solution of

$$\widehat{\theta}_T = \arg\min_{\theta\in\Theta} \left\|\widehat{G}_T(\theta)\right\| \tag{19}$$

with

$$\widehat{G}_T(\theta) = \frac{1}{T}\sum_{t=t_0+1}^{T} \widehat{g}_t(\theta), \qquad \widehat{g}_t(\theta) = \frac{\widehat{h}_t(\theta)}{\widehat{\sigma}_t^2(\theta)} \frac{\partial \widehat{f}_t(\theta)}{\partial\theta}.$$

Since $\Theta$ is a compact set and $f_t$ is assumed to be of class $C^1$, a solution of (19) always exists, but may not be unique. We will see that the asymptotic value of $\widehat{\theta}_T$ does not depend on the norm taken in (19).

### 3.1.3 Moment and identifiability assumptions

Let the rescaled innovations $\eta_t(\theta) = h_t(\theta)/\sigma_t(\theta)$. Consider the moment conditions

$$E \sup_{\theta\in\Theta} |\eta_t(\theta)|^r < \infty, \qquad E \sup_{\theta\in\Theta} \left\| \frac{1}{\sigma_t(\theta)} \frac{\partial f_t(\theta)}{\partial \theta} \right\|^r < \infty. \tag{20}$$

Let us illustrate these conditions on a duration model.

EXAMPLE 9. (Autoregressive Conditional Duration (ACD) model) *Let $y_t = f_t\epsilon_t$, where $(\epsilon_t)$ is i.i.d. positive with a mean of 1, $\theta = (\omega, \alpha_1, \ldots, \alpha_{q_0}, \beta_1, \ldots, \beta_{p_0}) \in \Theta$ and $f_t = f_t(\theta) = \omega + \sum_{i=1}^{q_0} \alpha_i y_{t-i} + \sum_{j=1}^{p_0} \beta_j f_{t-j}$. We have $\sigma_t^2(\theta) = f_t^2(\theta)\sigma_\epsilon^2$, where $\sigma_\epsilon^2 > 0$ is the variance of $\epsilon_t$. If $\Theta$ is a compact subset of $(0,\infty)^{p_0+q_0+1}$, by reproducing standard arguments used to show the CAN of the QMLE of GARCH models (see, in particular, (7.54) in Francq and Zakoian, 2019) it can be seen that the last condition of* (20) *holds for any $r$ and that the first holds whenever $E|\epsilon_t|^{r+\kappa} < \infty$ for some $\kappa > 0$. To show the latter result, we note that*

$$\left\| \frac{y_t}{f_t(\theta)} \right\|_r \leq \left\| \frac{f_t(\theta_0)}{f_t(\theta)} \right\|_{\frac{r(r+\kappa)}{\kappa}} \|\epsilon_t\|_{r+\kappa}$$

*by the Holder inequality.*

Under (20) with $r = 2$, let

$$G(\theta) = E g_1(\theta), \qquad g_t(\theta) = \frac{h_t(\theta)}{\sigma_t^2(\theta)} \frac{\partial f_t(\theta)}{\partial \theta}.$$

Since $E_{t-1}(y_t^k) = f_t(\theta_0)$, we obviously have $G(\theta_0) = 0$. Assume that the equality holds at only $\theta = \theta_0$:

$$\theta_0 \in \Theta \quad \text{and} \quad G(\theta) = 0 \text{ for } \theta \in \Theta \text{ if and only if } \theta = \theta_0. \tag{21}$$

Let us discuss the identifiability condition on the previous example.

EXAMPLE 10. (Identifiability of the ACD model) *Let us come back to Example 9. Let*

$$\mathcal{A}_\theta(z) = \sum_{i=1}^{q_0} \alpha_i z^i \quad \textit{and} \quad \mathcal{B}_\theta(z) = 1 - \sum_{j=1}^{p_0} \beta_j z^j.$$

*For any fixed $\theta^*$, the function*

$$\theta \mapsto E\frac{(y_t - f_t(\theta))^2}{\sigma_t^2(\theta^*)} = E\frac{(y_t - f_t(\theta_0))^2}{\sigma_t^2(\theta^*)} + E\frac{(f_t(\theta_0) - f_t(\theta))^2}{\sigma_t^2(\theta^*)}$$

*admits a minimum at $\theta$ iff $f_t(\theta) = f_t(\theta_0)$ almost surely. At such a point, and for any $\theta^*$, we have*

$$E\frac{y_t - f_t(\theta)}{\sigma_t^2(\theta^*)}\frac{\partial f_t(\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} = 0_{p_0+q_0+1}.$$

*Taking $\theta^* = \theta$, we thus see that $G(\theta) = 0$ if and only if $\theta$ is such that $f_t(\theta) = f_t(\theta_0)$ almost surely. By standard arguments (see,* e.g.*, step (b) in the proof of Theorem 7.1 of Francq and Zakoian, 2019 for an analogous result in GARCH models), it can be checked that*

$$f_t(\theta) = f_t(\theta_0) \quad a.s. \quad \Rightarrow \quad \theta = \theta_0$$

*under the following assumptions:*

1. *The distribution of $\epsilon_t$ is not degenerate;*
2. *If $p_0 > 0$, $\mathcal{A}_{\theta_0}(z)$ and $\mathcal{B}_{\theta_0}(z)$ have no common roots, $\mathcal{A}_{\theta_0}(1) \neq 0$, and $(\alpha_{0q_0}, \beta_{0p_0}) \neq (0, 0)$.*

To show the asymptotic normality of the QLEs, we need to consider the extra moment conditions

$$E\sup_{\theta\in\Theta}\left\|\frac{1}{\sigma_t^2(\theta)}\frac{\partial\sigma_t^2(\theta)}{\partial\theta\partial\theta^\top}\right\|^r < \infty, \quad E\sup_{\theta\in\Theta}\left\|\frac{1}{\sigma_t(\theta)}\frac{\partial^2 f_t(\theta)}{\partial\theta\partial\theta^\top}\right\|^r < \infty. \tag{22}$$

To deal with the effect of the initial values, we also need

$$\varrho^{-t}\sup_{\theta\in\Theta}\left\|\frac{\partial\widehat{\sigma}_t^2(\theta)}{\partial\theta} - \frac{\partial\sigma_t^2(\theta)}{\partial\theta}\right\| \to 0 \text{ a.s. as } t \to \infty. \tag{23}$$

### 3.1.4 Asymptotic behaviour of the QLE

Under (20) with $r = 2$, let us define the information matrices as follows:

$$\mathcal{I} = E\frac{h_t^2(\theta_0)}{\sigma_t^4(\theta_0)}\frac{\partial f_t(\theta_0)}{\partial\theta}\frac{\partial f_t(\theta_0)}{\partial\theta^\top}, \qquad \mathcal{J} = E\frac{1}{\sigma_t^2(\theta_0)}\frac{\partial f_t(\theta_0)}{\partial\theta}\frac{\partial f_t(\theta_0)}{\partial\theta^\top}.$$

Assume that

$$\mathcal{J} \text{ is invertible.} \tag{24}$$

Theorem 1 establishes the consistency and asymptotic normality of the QLE for conditional moment models, when the sample size diverges, i.e., when $T \to \infty$.

THEOREM 1. (CAN of the QLE for conditional moment models) *Let $\{y_t\}_{t\in Z}$ be generated by (2)-(3) with $\theta$ replaced by $\theta_0$ and $E_{t-1}(y_t^k) = f_t(\theta_0)$ for some $\theta_0 \in \Theta$ and $k > 0$. Let the conditions of Lemma 1 hold at $\theta = \theta_0$ and the conditions of Lemma 4 hold. Assume that $E\log^+|y_t|^k < \infty$, $E\log^+\sup_{\theta\in\Theta}|f_t(\theta)| < \infty$ and*

$E \log^+ \sup_{\theta\in\Theta} \|\partial f_t(\theta)/\partial\theta\| < \infty$. *Suppose further* (16), (17), (20) *with* $r = 2$, (21), $\Theta$ *is a compact subset of* $\mathbb{R}^p$ *and* $\theta_0 \in \Theta$. *Then, for any sequence* $\widehat{\theta}_T$ *satisfying* (19) *for* $T$ *large enough, we have* $\widehat{\theta}_T \overset{as}{\to} \theta_0$ *as* $T \to \infty$.

*Moreover, if* $\theta_0$ *belongs to the interior of* $\Theta$, (22) *holds with* $r = 2$, (23) *and* (24), *then*

$$\sqrt{T}(\widehat{\theta}_T - \theta_0) = \mathcal{J}^1 \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{\eta_t(\theta_0)}{\sigma_t(\theta_0)} \frac{\partial f_t(\theta_0)}{\partial\theta} + o_P(1) \overset{d}{\to} N(0, \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1})$$

*with the usual notation.*

The following remark discusses the link between the QLE in Theorem 1 and the QMLE.

REMARK 1. (Link with the QMLEs) *Since the works of Wedderburn (1974) and Gouriéroux, Monfort and Trognon (1984), it is known that in some location models of the form* $y_t = f_t(\theta) + \epsilon_t$, *the parameter* $\theta$ *can be estimated consistently by a quasi-maximum likelihood estimator (QMLE) that does not assume a particular distribution for* $\epsilon_t$ *but coincides with the MLE when the distribution of* $\epsilon_t$ *belongs to the linear exponential family, i.e., when, with respect to some* $\sigma$*-finite measure,* $\epsilon_t$ *admits a density of the form*

$$p_{f_t}(x) = \exp\{A(f_t) + B(x) + C(f_t)x\}.$$

*Since* $A'(f_t) + C'(f_t)f_t = 0$ *and* $C'(f_t(\theta)) = 1/s_t^2(\theta)$, *where* $s_t^2(\theta)$ *is the variance of the density* $p_{f_t(\theta)}$, *the quasi-score of this QMLE is*

$$s(\theta) = \frac{1}{T} \sum_{t=1}^{T} \frac{y_t - f_t(\theta)}{s_t^2(\theta)} \frac{\partial f_t(\theta)}{\partial\theta}.$$

*For instance, the Poisson QMLE*

$$\widehat{\theta}_T = \arg\max_{\theta\in\Theta} \sum_{t=1}^{T} \{-f_t(\theta) + y_t \log f_t(\theta)\}$$

*is the QLE obtained by assuming* $\sigma_t^2 = f_t(\theta)$ *in* $g(\theta)$. *The only–but essential–difference between QLE and QMLE is that the QMLE is based on a quasi-score with a variance constrained to be that of a linear exponential distribution. When the true density of* $\epsilon_t$ *does not belong to that family, the QLE and QMLE are generally consistent, but the QLE may be more efficient.*

Examples 11-14 below address the properties of the QLE and QMLE for duration and volatility models.

EXAMPLE 11. (Duration model) *Let the duration model* $y_t = f_t\epsilon_t$ *where* $\{\epsilon_t\}_{t\in\mathbb{Z}}$ *is i.i.d. positive with a mean of 1. For this model, take the QLE with* $\widehat{\sigma}_t^2 = \widehat{f}_t^2$ *(up to an unimportant multiplicative constant). Then*

$$\mathcal{I} = (E\epsilon_t^4 - 1)\mathcal{J}^{-1}, \qquad \mathcal{J} = E\frac{1}{f_t^2(\theta_0)}\frac{\partial f_t(\theta_0)}{\partial\theta}\frac{\partial f_t(\theta_0)}{\partial\theta^\top}.$$

*Note that the asymptotic variance of the QMLE is* $\mathcal{J}_0^{-1}\mathcal{I}_0\mathcal{J}_0^{-1}$, *where*

$$\mathcal{I}_0 = (E\epsilon_t^4 - 1)E\frac{f_t^2(\theta_0)}{s_t^4(\theta_0)}\frac{\partial f_t(\theta_0)}{\partial\theta}\frac{\partial f_t(\theta_0)}{\partial\theta^\top}, \qquad \mathcal{J}_0 = E\frac{1}{s_t^2(\theta_0)}\frac{\partial f_t(\theta_0)}{\partial\theta}\frac{\partial f_t(\theta_0)}{\partial\theta^\top}.$$

*Setting* $D_t = \mathcal{J}_0^{-1}\frac{(\epsilon_t^2-1)f_t(\theta_0)}{s_t^2(\theta_0)}\frac{\partial f_t(\theta_0)}{\partial\theta} - \mathcal{J}^{-1}\frac{\epsilon_t^2-1}{f_t(\theta_0)}\frac{\partial f_t(\theta_0)}{\partial\theta}$, *we have*

$$ED_tD_t' = \mathcal{J}_0^{-1}\mathcal{I}_0\mathcal{J}_0^{-1} - (E\epsilon_t^4 - 1)\mathcal{J}^{-1},$$

*showing that the QLE is asymptotically more, or equally, efficient than any QMLE, with equality when* $s_t^2(\theta_0) = f_t^2(\theta_0)$. *This corresponds to the QMLE based on the exponential distribution* $f(x) = \exp(-x)1_{x>0}$ *for* $\epsilon_t$.

EXAMPLE 12. (Standard volatility models and link with the Gaussian QMLE) *Consider the case where* (2) *is of the form* $y_t = \sqrt{f_t}\epsilon_t$ *with* $\epsilon_t$ *i.i.d. with a mean of 0 and variance of 1. The usual QMLE of the volatility parameter* $\theta_0$ *is*

$$\widehat{\theta}_T = \arg\min_{\theta\in\Theta}\sum_{t=t_0+1}^{T}\frac{y_t^2}{\widehat{f}_t(\theta)} + \log\widehat{f}_t(\theta).$$

*Writing the first-order conditions and noting that* $\sigma_t^2 = f_t^2\,(E\epsilon_t^4 - 1)$, *it is easy to see that, in the case, the QMLE coincides with the optimal QLE.*

EXAMPLE 13. (A non-multiplicative volatility model) *Let* $Z_t$ *be a random variable whose distribution, conditional on* $\mathcal{F}_{t-1}$, *is a Gamma law of shape parameter* $k_t = f_t^2/\sigma_t^2$ *and rate parameter* $\theta_t = \sigma_t^2/f_t$ *(so that* $E_{t-1}(Z_t) = f_t$ *and* $var_{t-1}(Z_t) = \sigma_t^2$*). Let* $y_t = s_t\sqrt{Z_t}$, *where* $s_t$ *is uniformly distributed on* $\{-1, 1\}$. *We thus have* $E_{t-1}(y_t^2) = f_t$ *and* $var_{t-1}(y_t^2) = \sigma_t^2$. *When* $\sigma_t^2$ *is not proportional to* $f_t^2$, *the sequence* $(y_t)$ *does not follow the standard volatility model of Example 12 and the QMLE is not the optimal QLE of* $\theta_0$ *involved in* $f_t = f_t(\theta_0)$. *For the sake of illustration, we run a Monte Carlo simulation in which we simulated* $T = 4,000$ *observations using the above model in which* $f_t$ *is specified as a GARCH*$(1,1)$ *model, i.e.,* $f_t = \omega + \alpha Z_t + \beta f_{t-1}$ *with* $\omega = 0.03, \alpha = 0.13$ *and* $\beta = 0.84$ *and* $\sigma_t^2 = 2$ *so that* $\sigma_t^2$ *is not proportional to* $f_t^2$. *A GARCH*$(1,1)$ *model is then estimated by Gaussian QMLE (i.e.,* $\sigma_t^2 = 2f_t^2$*) and optimal QLE (i.e.,* $\sigma_t^2 = 2$*). The biases (over 1,000 simulations) are found to be marginal for both methods. The RMSE of the three parameters, i.e.,* $\omega, \alpha$ *and* $\beta$, *are 0.0117, 0.0217, 0.0291 and 0.0073, 0.0093, 0.0128, respectively, for the Gaussian QMLE and optimal QLE, so that, on average, the optimal QLE is two times more efficient than the Gaussian QMLE.*

## 3.2 The MLE approach

When the conditional distribution $p(y_t \mid f_t, \theta)$ of the observations is entirely specified, the MLE is the benchmark estimator. It results in simultaneous estimation of the parameters involved in the time-varying parameter $f_t$ and the extra parameters involved in $p(\cdot \mid f, \theta)$. To estimate the parameters of the model, the MLE is often much more efficient than the QML and QL estimators when the conditional distribution $p(y_t \mid f_t, \theta)$ is well specified but is likely to be inconsistent when this distribution is misspecified. We will therefore study the asymptotic behaviour of the MLE in both situations. For the sake of simplicity, and in order to be able to apply existing ML estimation results, we focus on the case where there are no exogenous variables.

### 3.2.1 Strong consistency

Theorem 2 establishes the consistency of the MLE $\widehat{\theta}_T$ for $\psi GAS$ models satisfying the stationarity and invertibility conditions stated in Section 2.3. This theorem allows model misspecification and ensures only the convergence of the MLE $\widehat{\theta}_T$ to the pseudo-true parameter $\theta_0^*$ that maximizes the limit log-likelihood and minimizes the limit Kullback-Leilber divergence between the true conditional density of the data and the model-implied conditional density; see, e.g., White (1994, Chapter 3) for details. Below, $\ell(y_t, \widehat{f}_t(\theta), \theta)$ denotes the logarithm of the conditional density of $y_t$ given $\widehat{f}_t$, i.e., $\ell(y_t, \widehat{f}_t(\theta), \theta) = \log p(y_t|\widehat{f}_t, \theta)$, and $\widehat{\theta}_T$ is the MLE defined as

$$\widehat{\theta}_T = \arg\max_{\theta\in\Theta} \widehat{\boldsymbol{\ell}}_T(\theta), \quad \widehat{\boldsymbol{\ell}}_T(\theta) = \frac{1}{T}\sum_{t=2}^{T} \ell(y_t, \widehat{f}_t(\theta), \theta).$$

THEOREM 2. (Consistency of MLE under misspecification) *Let the conditions of Lemma 3 hold. Suppose further that $\ell$ is continuous, the parameter space $\Theta$ is compact, $\mathbb{E}\sup_{\theta\in\Theta}|\ell(y_t, \widehat{f}_t(\theta), \theta)| < \infty$ and there exists $\theta_0^* \in \Theta$ such that $\mathbb{E}\ell(y_t, f_t(\theta), \theta) < \mathbb{E}\ell(y_t, f_t(\theta_0^*), \theta_0^*)$ for every $\theta \neq \theta_0^*$, $\theta \in \Theta$. Then, $\widehat{\theta}_T \xrightarrow{as} \theta_0^* \in \Theta$ for every $\widehat{f}_1 \in \mathbb{C}(\Theta, \mathbb{R})$, as $T \to \infty$, and*

$$\theta_0^* := \arg\min \mathbb{E}\ \mathrm{KL}\Big(p_t^0(y_t),\ p(y_t|f_t(\theta), \theta)\Big).$$

When the $\psi GAS$ model is misspecified, the assumption of a unique maximizer of the limit log-likelihood $\theta_0^* \in \Theta$ may be too restrictive. Freedman and Diaconis (1982) show that uniqueness fails in a simple location problem with i.i.d. data. Kabaila (1983) provides similar results for ARMA models. Lemma 5 below follows Postcher and Prucha (1997, Lemma 4.2) and highlights that when the

uniqueness fails, the estimator can still be consistent to the argmax set of the limit log-likelihood as long as the *level sets* of the limit log-likelihood function are *regular* (see Definition 4.1 in Postcher and Prucha, 1997).

LEMMA 5. (Set consistency of MLE under possible misspecification) *Let the conditions of Lemma 3 hold. Suppose further that $\ell$ is continuous, $\Theta$ is compact, and $\mathbb{E}\sup_{\theta\in\Theta}|\ell(y_t, f_t(\theta), \theta)| < \infty$. Then, $\widehat{\theta}_T \overset{as}{\to} \theta_0^* \in \Theta$ for every $\widehat{f}_1 \in \mathbb{C}(\Theta, \mathbb{R})$, as $T \to \infty$, and*

$$\Theta_0^* := \arg\min \mathbb{E}\ \mathrm{KL}\Big(p_t^0(y_t),\ p(y_t|f_t(\theta), \theta)\Big).$$

The following example takes the MLE consistency and set consistency results in the context of a robust asymmetric $\psi GAS$ model for conditional volatilities.

EXAMPLE 14. (Robust asymmetric volatility $\psi GAS$ models) *Consider again the volatility model with $y_t = \sqrt{f_t}\epsilon_t$ with $f_t$ derived from an asymmetric Charbonnier loss function,*

$$f_{t+1}(\theta) = \omega + \alpha\sqrt{(y_t - \delta)^2 + \iota^2} + \beta f_t(\theta),$$

*where $\epsilon_t$ follows a standardized Student law $\mathrm{St}_\nu$ (with $\nu > 2$). The conditions of Lemma 3 are satisfied when $|\beta| < 1$. Furthermore, since the conditional log-likelihood is given by*

$$\ell(y_t, f_t(\theta), \theta) \propto -\frac{1}{2}\log(f_t(\theta)) - \frac{\lambda+1}{2}\log\Big(1 + \frac{y_t^2}{\lambda^2 f_t(\theta)}\Big),$$

*the moment condition $\mathbb{E}\sup_{\theta\in\Theta}|\ell(y_t, \widehat{f}_t(\theta), \theta)| < \infty$ holds easily as long as $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$ and the data have a small $n > 0$ bounded moment $\mathbb{E}|y_t|^n < \infty$,*

$$\mathbb{E}\sup_{\theta\in\Theta}|\ell(y_t, \widehat{f}_t(\theta), \theta)| \leq \mathbb{E}\sup_{\theta\in\Theta}|\log(\widehat{f}_t(\theta))| + \frac{\lambda+1}{2}\mathbb{E}\sup_{\theta\in\Theta}\Big|\log\Big(1 + \frac{y_t^2}{\lambda^2\omega^2}\Big)\Big| < \infty.$$

*Note that the small bounded moment for the data $y_t$, together with the contraction $|\beta| < 1$ and the compactness of $\Theta$, implies that the filter $\widehat{f}_t$ has $n$ moments uniformly in $t \in \mathbb{N}$,*

$$\begin{aligned}\sup_t \mathbb{E}\sup_{\theta\in\Theta}|\widehat{f}_{t+1}(\theta)|^n &\leq \sup_t \sum_{j=0}^{t-1}\sup_{\theta\in\Theta}|\beta|^{nj}\mathbb{E}\sup_{\theta\in\Theta}|u_{t-j}|^n + \sup_t\sup_{\theta\in\Theta}|\beta^t|^n\sup_{\theta\in\Theta}|\widehat{f}_1(\theta)|^n \\ &\leq (1 - \sup_{\theta\in\Theta}|\beta|^n)^{-1}\mathbb{E}\sup_{\theta\in\Theta}|u_{t-j}|^n + \sup_{\theta\in\Theta}|\beta|^n\sup_{\theta\in\Theta}|\widehat{f}_1(\theta)|^n \ < \ \infty,\end{aligned}$$

*where $u_t := \omega + \alpha\sqrt{(y_t - \delta)^2 + \iota^2}$ and $\mathbb{E}\sup_{\theta\in\Theta}|u_{t-j}|^n < \infty \Leftarrow \mathbb{E}|y_t|^n < \infty$. As a result, we can conclude by Theorem 2 that the MLE is strongly consistent for the pseudo-true parameter $\widehat{\theta}_T \overset{as}{\to} \theta_0^*$ (when $\theta_0^*$ is unique), or set-consistent $\widehat{\theta}_T \overset{as}{\to} \Theta_0^*$ (when uniqueness fails), as the sample size diverges, i.e., $T \to \infty$.*

In Theorem 2, we imposed high-level conditions on the data $\{y_t\}_{t\in\mathbb{Z}}$ since the data generating process was left unspecified. Corollary 1 highlights that if the $\psi GAS$ model is assumed to be well specified, then we can derive the properties of the data and the convergence of the MLE $\widehat{\theta}_T$ to the vector of true parameters $\theta_0$ can be obtained under the additional conditions of Lemma 1 (ensuring that the data are well behaved).

COROLLARY 1. (Consistency of MLE under correct specification) *Let $\{y_t\}_{t\in Z}$ be generated by (2) and (4) under some $\theta_0 \in \Theta$ and let the conditions of Lemma 1 hold at $\theta_0 \in \Theta$, and the conditions of Lemma 3 hold on $\Theta$. Suppose further that $\Theta$ is compact and $\mathbb{E}\sup_{\theta\in\Theta}|\ell(y_t, f_t(\theta), \theta)| < \infty$. Finally, let $\mathbb{E}\log^+|f_t| < \infty$ and $\mathbb{E}\ell_t(\theta_0) > \mathbb{E}\ell_t(\theta)\ \forall\ \theta \neq \theta_0$. Then, $\widehat{\theta}_T \xrightarrow{as} \theta_0 \in \Theta$.*

The following example revisits the robust location $\psi GAS$ model and discusses MLE consistency under correct specification. Additionally, we also consider the robust asymmetric conditional volatility model.

EXAMPLE 15. (Robust location $\psi GAS$ models) *We revisit the location model $y_t = f_t + \epsilon_t$, with standardized Student's t-distributed innovations $\epsilon_t$ (i.e., $\epsilon_t \sim \mathrm{St}_\nu$, with $\nu > 2$), and the negative pseudo Huber loss function,*

$$f_{t+1}(\theta) = \omega + \alpha\frac{(y_t - f_t(\theta))}{\sqrt{(y_t - f_t(\theta))^2/\delta^2 + 1}} + \beta f_t(\theta).$$

*As we have already seen, the contraction condition of Lemma 1 holds if $|\beta_0| < 1$, and the conditions of Lemma 3 hold if $\sup_{\theta\in\Theta}(|\alpha| + |\beta|) < 1$. Note that since $|\psi|$ is uniformly bounded by $|\delta| > 0$, the moment conditions of Lemmas 1 and 3 are immediately satisfied, and furthermore, $f_t(\theta)$ and $\widehat{f}_t(\theta)$ are both uniformly bounded,*

$$|f_t(\theta)| \leq |\omega| + |\alpha\delta| + |\beta||f_{t-1}(\theta)| \leq \frac{|\omega| + |\alpha\delta|}{1 - |\beta|} < \infty, \quad \textit{and}$$

$$\sup_t \sup_{\theta\in\Theta} |\widehat{f}_{t+1}(\theta)| \leq (1 - \sup_{\theta\in\Theta}|\beta|)^{-1} \sup_{\theta\in\Theta}(|\omega| + |\alpha\delta|) + \sup_{\theta\in\Theta}|\beta| \sup_{\theta\in\Theta}|\widehat{f}_1(\theta)| \ < \ \infty.$$

*As a result, the data generated by the model have a logarithmic moment*

$$\mathbb{E}\log^+|y_t| = \mathbb{E}\log^+|f_t\epsilon_t| \leq \log^+\frac{|\omega_0| + |\alpha_0\delta_0|}{1 - |\beta_0|} + \mathbb{E}\log^+|\epsilon_t| < \infty$$

*as long as $\lambda_0 > 0$. Furthermore, the moment bound $\mathbb{E}\sup_{\theta\in\Theta}|\ell(y_t, f_t(\theta), \theta)| < \infty$ holds easily since*

$$\ell(y_t, f_t(\theta), \theta) \propto -\frac{1}{2}\log(\sigma^2) - \frac{\lambda + 1}{2}\log\Big(1 + \frac{(y_t - f_t(\theta))^2}{\lambda^2\sigma^2}\Big), \quad \textit{and hence,}$$

$$\mathbb{E}\sup_{\theta\in\Theta}|\ell(y_t, f_t(\theta), \theta)| \leq \sup_{\theta\in\Theta}|\log(\sigma^2)| + \frac{\lambda+1}{2}\mathbb{E}\sup_{\theta\in\Theta}\left|\log\left(1+\frac{(y_t - f_t(\theta))^2}{\lambda^2\sigma^2}\right)\right| < \infty.$$

*As a result, we can conclude by Theorem 2 that if $\theta_0$ is identifiable, then the MLE is strongly consistent for the true parameter $\widehat{\theta}_T \xrightarrow{as} \theta_0$ as $T \to \infty$.*

EXAMPLE 16. (Robust asymmetric volatility $\psi GAS$ models) *Revisit once more the volatility $\psi GAS$ models with $y_t = f_t\epsilon_t$ and asymmetric Charbonnier loss function $f_{t+1} = \omega + \alpha\sqrt{(y_t - \delta)^2 + \iota^2} + \beta f_t$. We already know that conditions of Lemmas 1 and 3 hold. Since $\psi(y, f, \theta) \geq 0 \; \forall \; (y, f, \theta)$ holds trivially, we obtain the consistency of the MLE as long as the true conditional volatility has a logarithmic moment.*

### 3.2.2 Asymptotic normality

We now turn to the asymptotic normality of the ML and QML estimators for the static parameters of $\psi GAS$ models.

When the $\psi GAS$ model is correctly specified, we can use the martingale difference sequence property of the score at $\theta_0$ to obtain a central limit theorem. However, when the model is misspecified, the score will generally fail to be a martingale difference sequence; see White (1994). Lemma 6 ensures that the MLE's score is near epoch dependent (NED) on an underlying (strong mixing) sequence; see e.g., Davidson (1994) and Potscher and Prucha (1997, Definition 6.3). This lemma is written for robust $\psi GAS$ models with bounded updates delivered by a uniformly bounded $\psi$ function $\sup_{f,y}|\psi(y, f, \theta)| < \infty$ with uniformly bounded derivatives. The NED property gives us sufficient fading memory for establishing the asymptotic normality of the score when the model is misspecified and the score fails to be a martingale difference sequence (Potscher and Prucha, Chapter 10).

Let $\widehat{\ell}'_t(\theta_0)$ denote the score evaluated at $\theta_0$ and defined as follows:

$$\widehat{\ell}'_t(\theta_0) = \frac{\partial\ell(y_t, \widehat{f}_t(\theta_0), \theta_0)}{\partial\theta} + \frac{\partial\ell(y_t, \widehat{f}_t(\theta_0), \theta_0)}{\partial f}\widehat{f}_t(\theta_0)'.$$

Notice that a hat is used in the notation $\widehat{\ell}'_t$ to highlight the fact that the score depends on the filtered values $(\widehat{f}_t, \widehat{f}'_t)$.

LEMMA 6. (Near epoch dependent score) *Let $\{y_t\}$ have two bounded moments $\sup_t \mathbb{E}|y_t|^2 < \infty$ and be NED of size $-q$ on some process $\{e_t\}_{t\in\mathbb{Z}}$ and suppose that*

*(i)* $\sup_{y,f}\left|\frac{\partial\psi(y,f,\theta_0)}{\partial y}\right| < \infty$; *(ii)* $\sup_{y,f}\left|\alpha_0\frac{\partial\psi(y,f,\theta_0)}{\partial f} + \beta_0\right| < 1.$

*Then,* $\{\widehat{f}_t(\theta_0)\}_{t\in\mathbb{N}}$ *is NED of size* $-q$ *on* $\{e_t\}_{t\in\mathbb{Z}}$*. Additionally, if we have that*

*(iii)* $\sup_{y,f}\left|\psi(y,f,\theta_0)\right|<\infty$*;* *(iv)* $\sup_{y,f}\left|\frac{\partial\psi(y,f,\theta_0)}{\partial\theta}\right|<\infty$*;*

*(v)* $\sup_{y,f}\left|\frac{\partial^2\psi(y,f,\theta_0)}{\partial\theta\partial y}\right|<\infty$*;* *(vi)* $\sup_{y,f}\left|\frac{\partial^2\psi(y,f,\theta_0)}{\partial\theta\partial f}\right|<\infty$

*then the derivative process* $\{\widehat{f}'_t(\theta_0)\}_{t\in\mathbb{N}}$ *is NED of size* $-q$ *on* $\{e_t\}_{t\in\mathbb{Z}}$*. Finally, if the score* $\widehat{\ell}'_t(\theta_0)$ *is Lipschitz on* $(y_t,\widehat{f}_t,\widehat{f}'_t)$*,*

*(vii)* $\sup_{y,f,\widehat{f}_\theta}\left|\frac{\partial^2\ell(y_t,\widehat{f}_t,\theta_0)}{\partial\theta\partial y}\right|<\infty$*;* *(viii)* $\sup_{y,f,\widehat{f}_\theta}\left|\frac{\partial^2\ell(y_t,\widehat{f}_t,\theta_0)}{\partial\theta\partial f}\right|<\infty$*;*

*(ix)* $\sup_{y,f,\widehat{f}_\theta}\left|\frac{\partial^2\ell(y_t,\widehat{f}_t,\theta_0)}{\partial f^2}\right|<\infty$*;* *(x)* $\sup_{y,f,\widehat{f}_\theta}\left|\frac{\partial^2\ell(y_t,\widehat{f}_t,\theta_0)}{\partial f\partial y}\right|<\infty$*;*

*(xi)* $\sup_{y,f,\widehat{f}_\theta}\left|\frac{\partial\ell(y_t,\widehat{f}_t,\theta_0)}{\partial f}\right|<\infty$*.*

*Then,* $\{\widehat{\ell}'_t(\theta_0)\}_{t\in\mathbb{N}}$ *is also NED of size* $-q$ *on* $\{e_t\}_{t\in\mathbb{Z}}$*.*

EXAMPLE 17. (Robust location $\psi GAS$ models) *Consider the location model* $y_t=f_t+\epsilon_t$ *with Student's t-innovations and the negative pseudo Huber loss function,*

$$f_{t+1}=\omega+\alpha\frac{(y_t-f_t)}{\sqrt{(y_t-f_t)^2/\delta^2+1}}+\beta f_t.$$

*Conditions (i)–(vi) hold since the* $\psi$ *function and its derivatives are uniformly bounded. Furthermore, the score* $\widehat{\ell}$ *is Lipschitz on* $(y,\widehat{f}_t,\widehat{f}'_t)$ *since the uniform bounds (vii)–(xi) hold for the Student's t and log-likelihood scores. We thus conclude that if* $\{y_t\}$ *is NED of size* $-q$ *on some sequence* $\{e_t\}_{t\in\mathbb{Z}}$*, the score* $\{\widehat{\ell}'_t(\theta_0)\}_{t\in\mathbb{N}}$ *is also NED of size* $-q$ *on* $\{e_t\}_{t\in\mathbb{Z}}$*.*

Theorem 3 uses the stochastic properties discussed in Lemmas 4 and 6 to obtain the asymptotic normality of the MLE in a setting where the model is allowed to be misspecified; see White (1982), Domowitz and White (1982), White (1994), and Potscher and Prucha (1997). In this theorem, we assume that the data are near epoch dependent on an underlying $\phi$-mixing sequence of size $-r/(r-1)$. The same result can, however, be obtained for $\alpha$-mixing sequences of size $-2r/(r-2)$.

THEOREM 3. (Asymptotic normality of MLE under possible misspecification) *Assume that the conditions in Theorem 2 and Lemmas 4 and 6 are satisfied. Suppose*

*further that* $\theta_0^* \in int(\Theta)$ *and* $\{y_t\}_{t\in\mathbb{Z}}$ *is near epoch dependent of size* $-1$ *on a* $\phi$*-mixing sequence of size* $-r/(r-1)$ *for some* $r > 2$*, and that*

$$\mathbb{E}|\ell'(y_t, f_t, \theta_0)|^r < \infty, \quad \mathbb{E}\sup_{\theta\in\Theta}|\ell'(y_t, f_t, \theta)| < \infty \quad \textit{and} \quad \mathbb{E}\sup_{\theta\in\Theta}|\ell''(y_t, f_t, \theta)| < \infty.$$

*Suppose further that* $\mathbb{E}\widehat{\ell}_t''(\theta_0^*)$ *is invertible. Then* $\sqrt{T}(\widehat{\theta}_T - \theta_0^*) \xrightarrow{d} N\big(0, \Sigma(\theta_0^*)\big)$ *as* $T \to \infty$*, where*

$$\Sigma(\theta_0^*) = \Big(\mathbb{E}\widehat{\ell}_t''(\theta_0^*)\Big)^{-1}\Big(\mathbb{E}\widehat{\ell}_t'(\theta_0)\mathbb{E}\widehat{\ell}_t'(\theta_0)^\top\Big)\Big(\mathbb{E}\widehat{\ell}_t''(\theta_0^*)\Big)^{-1}.$$

EXAMPLE 18. (Robust location $\psi GAS$ models) *For the location model* $y_t = f_t + \epsilon_t$*, with the negative pseudo Huber loss function, we have already seen that*

$$\ell(y_t, f_t, \theta) \propto -\frac{1}{2}\log(\sigma^2) - \frac{\lambda+1}{2}\log\Big(1 + \frac{(y_t - f_t)^2}{\lambda^2\sigma^2}\Big).$$

*The conditions of Theorem 3 can be easily verified for this model since the filter and its derivatives are smooth in the parameters and uniformly bounded with uniformly bounded contractions.*

COROLLARY 2. (Asymptotic normality of MLE under correct specification) *Let* $\{y_t\}_{t\in Z}$ *be generated by (2) and (4) under some* $\theta_0 \in \Theta$*, and let the conditions of Corollary 1 hold and Lemma 4 hold. Suppose further that*

$$\mathbb{E}|\ell'(y_t, f_t, \theta_0)|^2 < \infty \quad \mathbb{E}\sup_{\theta\in\Theta}|\ell'(y_t, f_t, \theta)| < \infty \quad \textit{and} \quad \mathbb{E}\sup_{\theta\in\Theta}|\ell''(y_t, f_t, \theta)| < \infty.$$

*Then,* $\sqrt{T}(\widehat{\theta}_T - \theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0)^{-1})$*, where* $\mathcal{I}(\theta_0)$ *denotes the Fisher information matrix.*

EXAMPLE 19. (Robust location $\psi GAS$ models) *As seen above, the conditions in Theorem 3 hold for this model when the data* $\{y_t\}_{t\in\mathbb{Z}}$ *are strictly stationary and ergodic and satisfy* $\mathbb{E}|y_t|^n < \infty$ *for* $n > 2$*. If the model is correctly specified, then we can ensure the stationarity and ergodicity of* $y_t$ *by imposing that* $\beta_0 < 1$ *since*

$$f_{t+1}(\theta_0) = \omega_0 + \alpha_0\frac{\epsilon_t}{\sqrt{(\epsilon_t)^2/\delta_0^2 + 1}} + \beta_0 f_t(\theta_0),$$

*and hence* $\mathbb{E}\log sup_f|\partial f_{t+1}/\partial f_t| = \log(\beta) < 0$*. Additionally, we obtain* $\mathbb{E}|y_t|^n < \infty$ *by imposing* $\lambda > n$ *since by the* $c_n$*-inequality* $\exists 0 < c < \infty$ *such that*

$$\mathbb{E}|y_t|^n \le c\mathbb{E}|f_t(\theta_0)|^n + \mathbb{E}|\epsilon_t|^n \le cM^n + \mathbb{E}|\epsilon_t|^n < \infty$$

*as* $|f_t| < M$ *a.s. and* $\epsilon_t \sim t(\lambda)$ *and hence* $\mathbb{E}|\epsilon_t|^n < \infty \;\Leftarrow\; \lambda > n$*.*

As a final result, we obtain the asymptotic distribution of the score (also called Lagrange Multiplier) and log-likelihood ratio tests for testing $r$ linear restrictions on the $p > r$ dimensional parameter vector $\theta_0$. The null hypothesis of interest is $\mathrm{H}_0 : R\theta_0 = \boldsymbol{r}$, where $R$ is a given full rank $r \times (p-r)$ matrix and $\boldsymbol{r}$ is a given $r$-dimensional vector. Below, $\widehat{\theta}_T^r$ denotes the MLE of $\theta_0$ in the model constrained by the null.

THEOREM 4. (Score and Likelihood Ratio tests) *Let the conditions of Theorem 2 hold. Then, under $H_0 : R\theta_0 = \boldsymbol{r}$, we have that*

$$\begin{aligned} \mathrm{LM}_T &= T \frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta^\top} \widehat{\mathcal{I}}^{-1} \frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta} \xrightarrow{d} \chi^2_r, \\ \mathrm{LR}_T &= 2T\Big(\widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T) - \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)\Big) \xrightarrow{d} \chi^2_r \quad \textit{as } T \to \infty, \end{aligned}$$

*where $\widehat{\mathcal{I}}$ is a weakly consistent estimator of $\mathcal{I}(\theta_0)$. One can take, for instance,*

$$\widehat{\mathcal{I}} = -\frac{\partial^2 \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta \partial \theta^\top} \qquad \textit{or} \qquad \widehat{\mathcal{I}} = \frac{1}{T} \sum_{t=1}^{T} \widehat{\ell}_t'(\widehat{\theta}_T^r) \widehat{\ell}_t'^\top(\widehat{\theta}_T^r). \tag{25}$$

*In the latter case, we have $\mathrm{LM}_T = \mathbf{1}^\top \widehat{L}^\top \left(\widehat{L}^\top \widehat{L}\right)^{-1} \widehat{L}\mathbf{1}$ where $\widehat{L}$ is a $p \times T$ matrix whose row $t$ is $\widehat{\ell}_t'(\widehat{\theta}_T^r)$ and $\mathbf{1}^\top = (1, \ldots, 1) \in \mathbb{R}^T$. Note that $\mathrm{LM}_T = T \times R^2$, where $R^2$ denotes the coefficient of determination in the regression of 1 on $\widehat{\ell}_t'(\widehat{\theta}_T^r)$.*

# 4 The $\psi_T GAS - T$ example

In this section, we illustrate our general results on an extension of one of the most popular score-driven volatility models.

## 4.1 An extension of the $\beta_T GAS$

Assume the volatility model $y_t = \sqrt{f_t}\epsilon_t$, where

$$f_{t+1} = \omega + \alpha \frac{\nu+1}{\nu - 2 + \epsilon_t^2} \epsilon_t^2 f_t + \beta f_t, \tag{26}$$

where $\omega > 0$, $\alpha > 0$ and $\beta \geq 0$ to ensure positivity and avoid triviality. Harvey and Chakravarty (2008) show that (26) is the updating equation of a $GAS$ model when the i.i.d. sequence $(\epsilon_t)$ follows a standardized Student law $\mathrm{St}_\nu$ with $\nu > 2$ degrees

of freedom (and variance 1). The model is often called $Beta-t\ GARCH(1,1)$ but we call it $\beta_T GAS(1,1)$ in this paper. Note that $\nu$ plays two roles in this model: it determines the shape of the density of the innovations $\epsilon_t$ and bounds the effects of large shocks $\epsilon_t$ on future values of the conditional variance (i.e., $f_{t+1}$). Note also that it is common to reparameterize (26) in terms of $\xi = 1/\nu$, i.e.,

$$f_{t+1} = \omega + \alpha \frac{1+\xi}{1-2\xi+\xi\epsilon_t^2}\epsilon_t^2 f_t + \beta f_t \tag{27}$$

with $0 \leq \xi < 1/2$ so that the $GARCH(1,1)$ appears as a special case of (27) when $\xi = 0$.

In the sequel, we keep the downweighting mechanism of the above $\beta_T GAS(1,1)$ model but disconnect the updating equation of the conditional variance and the density of the innovations. To do so, we assume $\epsilon_t \sim \mathrm{St}_\nu$ as for the $\beta_T GAS$, but we introduce an additional parameter $\zeta$ in the updating equation that is not related to $\nu$ (or its inverse). The model, called $\psi_T GAS(1,1) - T$, is parameterized as follows:[1]

$$f_{t+1} = \omega + \alpha \frac{1+\zeta}{1-2\zeta+\zeta\epsilon_t^2}\epsilon_t^2 f_t + \beta f_t, \tag{28}$$

where $\epsilon_t \overset{i.i.d.}{\sim} \mathrm{St}_{1/\xi}$ with $0 \leq \xi < 1/2$. This model is identical to a $\beta_T GAS(1,1)$ when $\xi = \zeta$. When $\zeta = 0$, Equation (28) corresponds to a standard $GARCH(1,1)$ model $f_{t+1} = \omega + \alpha y_t^2 + \beta f_t$ with standardized Student's t-innovations.

When $\zeta < 0$ or $\zeta > 1/2$, Equation (28) does not define a proper volatility model because when $\epsilon_t^2 \simeq 2 - 1/\zeta$, $f_{t+1}$ in (28) can be infinite or negative.

Note also that when $\zeta = 1/2$, the volatility model is degenerated since $f_{t+1} = \omega + 3\alpha f_t + \beta f_t$ is then constant. The same remark holds when $\zeta = -1$. To ensure positivity and non-degeneracy of the volatility equation, one can impose $0 \leq \zeta < 1/2$. However, to avoid $\zeta$ to be on the boundary of the parameter space when testing the null hypothesis $\zeta = 0$ (i.e., that the true model is a $GARCH(1,1)$ model), we also consider the alternative specification

$$f_{t+1} = \omega + \alpha\Psi\left(\frac{1+\zeta}{1-2\zeta+\zeta\epsilon_t^2}\right)\epsilon_t^2 f_t + \beta f_t \tag{29}$$

with $-1 < \zeta < 1/2$ and $\Psi : \mathbb{R} \to [0,\infty)$ of class $C^2$. To approximately recover (28) when $\zeta \geq 0$, one can chose for $\Psi$ a smooth approximation of the absolute value function. For instance, one can set $\Psi(x) = \sqrt{x^2 + c}$ for some small $c > 0$ or

$$\Psi(x) = x\frac{1-e^{-cx}}{1+e^{-cx}} \tag{30}$$

[1] $\psi_T$ in the name of the model refers to the score used in the updating equation, while $-T$ refers to the density of the innovations. However, in this model both the score and the density of the innovations are taken from a standardized Student's t-density, and the degrees of freedom are not assumed to be the same as for the $\beta_T GAS$ model.

for some large $c > 0$. The latter function is equivalent to $|x|$ when $|x|$ or $c$ is large and is equivalent to $cx^2/2$ when $|x|$ is small. More generally, assume that

$$\Psi(x) \leq c_1(|x| + 1), \qquad \Psi(x) \geq c_2|x|^{c_3}, \qquad |\Psi'(x)| \leq c_4 \tag{31}$$

for some positive constants $c_i$, $i = 1, \ldots, 4$. In the simulations and the empirical application, we rely on (30) with $c = 1,000$.

## 4.2 Stationarity and positivity conditions

Let us consider the stationarity of the general $\psi_T GAS$ model (28) without assuming a particular distribution for $(\epsilon_t)$. For the moment, we just assume that $(\epsilon_t)$ is stationary and ergodic with $E\epsilon_t^2 = 1$. By the Cauchy root test, it is easy to show that there exists a stationary (ergodic) solution to this $\psi_T GAS(1,1)$ model, explicitly given by

$$f_t = \omega \left\{ 1 + \sum_{i=1}^{\infty} a(\epsilon_{t-1}) \cdots a(\epsilon_{t-i}) \right\}, \quad a(z) = \alpha \Psi \left( \frac{1+\zeta}{1-2\zeta+\zeta z^2} \right) z^2 + \beta,$$

when

$$E \log \left( \alpha \Psi_t \epsilon_t^2 + \beta \right) < 0, \qquad \Psi_t = \Psi \left( \frac{1+\zeta}{1-2\zeta+\zeta\epsilon_t^2} \right). \tag{32}$$

Note that (28) corresponds to (4) with

$$\psi \left( g(f, \epsilon_t), f, \theta \right) = \Psi_t \epsilon_t^2 f, \quad \frac{\partial \psi(g(f, \epsilon_t), f, \theta)}{\partial f} = \Psi_t \epsilon_t^2.$$

Using the first two inequalities of (31), it can be seen that condition (i) of Lemma 1 is satisfied when

$$E \log^- |1 - 2\zeta + \zeta\epsilon_t^2| < \infty \tag{33}$$

and that (ii) is equivalent to (32). Note that the moment condition (33) is very mild and is always satisfied when $\zeta \geq 0$, or when the distribution of $\epsilon_t^2$ has a bounded density. On the other hand, (33) for $\zeta < 0$ precludes a distribution of $\epsilon_t^2$ with a mass at $2 - 1/\zeta$. Note also that (32) is also a necessary condition for stationarity when $(\epsilon_t)$ is i.i.d., which shows that Lemma 1 provides sharp stationarity conditions, at least in this framework.

## 4.3 Invertibility of the filter

We now assume that $(\epsilon_t)$ is an i.i.d. sequence and that conditions (31)-(32) hold true. The conditions of Lemma 2 are satisfied, which shows that the stationary

solution of the $\psi_T GAS(1,1)-T$ volatility model is such that $E|y_t|^s < \infty$ for some $s > 0$. Since we have (6) and (7) with

$$\begin{aligned}\psi(y_t, f_t, \theta) &= \Psi\left(\frac{1+\zeta}{1-2\zeta+\zeta\frac{y_t^2}{f_t}}\right) y_t^2, \\ \frac{\partial\psi(y_t, f_t, \theta)}{\partial f} &= \Psi'\left(\frac{1+\zeta}{1-2\zeta+\zeta\frac{y_t^2}{f_t}}\right) \frac{1+\zeta}{\left(1-2\zeta+\zeta\frac{y_t^2}{f_t}\right)^2}\zeta\frac{y_t^4}{f_t^2},\end{aligned}$$

condition (i) of Lemma 3 is satisfied (using (31) and Lemma 2, which entails the existence of a small moment for $y_t$). Assume that $\Theta$ is such that, for all $\theta = (\omega, \alpha, \beta, \zeta)^\top \in \Theta$ we have $0 \leq \zeta \leq \overline{\zeta} < 1/2$. The uniform invertibility condition (ii) is then satisfied when

$$E\log\left(c_4\overline{\alpha}\frac{1+\overline{\zeta}}{\left(1-2\overline{\zeta}+\overline{\zeta}\frac{y_t^2}{\underline{f}}\right)^2}\overline{\zeta}\frac{y_t^4}{\underline{f}} + \overline{\beta}\right) < 0, \tag{34}$$

where $\underline{\omega}$, $\underline{\alpha}$, $\underline{\beta}$ and $\overline{\omega}$, $\overline{\alpha}$, $\overline{\beta}$ are, respectively, upper and lower bounds for $\omega$, $\alpha$ and $\beta$ over $\Theta$, and $\underline{f} = \underline{\omega}/(1-\underline{\beta})$ is a lower bound for the time-varying volatility. Note that the expectation of the left-hand side of (34) cannot be computed exactly because the stationary distribution of $(y_t)$ is generally unknown, but it can be easily evaluated by means of simulations. To relax the constraint $\zeta \geq 0$ or to obtain a more stringent identifiability condition (in particular, to account for a non-cubic parameter space $\Theta$), the supremum involved in condition (ii) of Lemma 3 can be computed numerically.

## 4.4 Derivatives of the filter

Setting $\theta = (\omega, \alpha, \beta, \nu)'$, (9) holds with

$$A_t = \begin{pmatrix} 1 \\ \psi_t \\ f_t \\ \alpha\frac{\partial\psi_t}{\partial\zeta} \end{pmatrix}, \quad \frac{\partial\psi_t}{\partial\zeta} = y_t^2\Psi'\left(\frac{1+\zeta}{1-2\zeta+\zeta\frac{y_t^2}{f_t}}\right)\frac{3-\frac{y_t^2}{f_t}}{\left(1-2\zeta+\zeta\frac{y_t^2}{f_t}\right)^2}.$$

Assume $0 \leq \zeta \leq \overline{\zeta} < 1/2$. We thus have $1-2\zeta+\zeta\frac{y_t^2}{f_t} \geq 1-2\overline{\zeta} > 0$. Since $f_t \geq \omega > 0$, Lemma 3 entails that $E\|A_t\|^s < \infty$ for some $s > 0$. Therefore, $E\log^+\|A_t\| < \infty$ and (i) of Lemma 4 is satisfied. Similarly, it can be seen that the other conditions of that lemma hold true.

## 4.5 Estimating the parameters

We now consider the estimation of the $\psi_T GAS(1,1) - T$ model. We thus assume the standardized Student's conditional distribution

$$p(y \mid f, \theta) = \frac{1}{\sqrt{f}\sqrt{\frac{\nu-2}{\nu}}} p_\nu \left( \frac{y}{\sqrt{f}\sqrt{\frac{\nu-2}{\nu}}} \right), \tag{35}$$

$$p_\nu(y) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

with $\nu > 2$. To allow the Gaussian distribution, set $\xi = 1/\nu$ and impose $0 \leq \xi < 1/2$, the case $\xi = 0$ corresponding to the $\mathcal{N}(0,1)$ conditional distribution. Let $\theta = (\omega, \alpha, \beta, \zeta, \xi)'$, $\Theta$ a compact subset of $(0,\infty)^2 \times [0,1] \times (-1, 1/2) \times [0, 1/2)$ and the MLE

$$\widehat{\theta}_T = \arg\max_{\theta \in \Theta} \frac{1}{T} \sum_{t=t_0+1}^{T} \ell(y_t, \widehat{f}_t(\theta), \theta),$$

where $\ell(y, f, \theta) = \log p(y \mid f, \theta)$ and

$$\widehat{f}_{t+1}(\theta) = \omega + \alpha \Psi \left( \frac{1+\zeta}{1 - 2\zeta + \zeta \frac{y_t^2}{\widehat{f}_t(\theta)}} \right) + \beta \widehat{f}_t(\theta),$$

with the initial value $\widehat{f}_1(\theta) = \sum_{t=1}^{t_0} y_t^2 / t_0$ and $t_0 = 5$, for instance.

### 4.5.1 Testing the $\beta_T GAS$

The standard $\beta_T GAS$ is obtained when $0 < \zeta = \xi < 1/2$ and $\Psi(x) = x$. It is thus of interest to test the null $H_0 : \xi_0 = \zeta_0$. This hypothesis can be written as $H_0 : K\theta_0 = 0$ with $K = (0, 0, 0, 1, -1)$. Let the Wald test statistic be

$$W_T^{\zeta,\xi} = T \widehat{\theta}_T^\top K^\top \left( K \widehat{\Sigma} K^\top \right)^{-1} K \widehat{\theta}_T,$$

where $\widehat{\Sigma}$ is a consistent estimator of the matrix $\mathcal{I}^{-1}(\theta_0)$ defined in Corollary 2. A direct consequence of that corollary is that $W_T^{\zeta,\xi}$ asymptotically follows a $\chi_1^2$ under $H_0$. The test of critical region $\{W_T^{\zeta,\xi} > \chi_1^2(1-\alpha)\}$ thus has the asymptotic level $\alpha$.

Alternatively, one can use Theorem 4 and replace the Wald statistic by the score and likelihood ratio (LR) test statistics

$$LM_T^{\zeta,\xi} = T \frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta^\top} \widehat{\mathcal{I}}^{-1} \frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta}, \quad LR_T^{\zeta,\xi} = 2T \Big( \widehat{\ell}_T(\widehat{\theta}_T) - \widehat{\ell}_T(\widehat{\theta}_T^r) \Big).$$

For the Wald statistics, it is natural to take $\widehat{\Sigma} = \widehat{\mathcal{I}}^{-1}$ where $\widehat{\mathcal{I}}$ is defined by (25), replacing $\widehat{\theta}_T^r$ by $\widehat{\theta}_T$.

### 4.5.2 Testing the $GARCH-T$

The standard $GARCH(1,1)$ volatility model with Student's t-innovations is obtained when $\zeta = 0$ and $\Psi(x) = x$. It is thus of interest to test the null $H_0 : \zeta_0 = 0$ in the $\psi_T GAS(1,1) - T$ model defined by (35) and (29), with $-1 < \zeta < 1/2$ and $\Psi$ satisfying (31). Another possibility would be to test $\zeta_0 = 0$ in the model defined by (35) and (29) constrained by $0 \leq \zeta < 1/2$. The drawback of the latter test is that, because the parameter stands at the boundary of the parameter space under the null, the asymptotic distribution of the Wald statistic is non-standard (see Pedersen and Rahbek, 2019 and the reference therein).

By considering model (29), we afford to have $-1 < \zeta < 1/2$, and thus the parameter belongs to the interior of $\Theta$ under $H_0 : \zeta_0 = 0$. Corollary 2 then entails that the Wald test of critical region $\{W_T^\zeta > \chi_1^2(1-\alpha)\}$ with

$$W_T^\zeta = T\widehat{\theta}_T^\top \boldsymbol{e}_4 \left(\boldsymbol{e}_4^\top \widehat{\Sigma} b e_4\right)^{-1} \boldsymbol{e}_4^\top \widehat{\theta}_T, \quad \boldsymbol{e}_4^\top = (0,0,0,1,0),$$

has asymptotic level $\alpha$.

### 4.5.3 Testing the standard $GARCH$

The parameter of main interest is often the volatility $f_t = f(\theta_0)$ with $\theta_0 = (\omega_0, \alpha_0, \beta_0, \zeta_0)'$ and $\Theta$ changed accordingly. It is then desirable to estimate $\theta_0$ without assuming (35) or any other particular conditional distribution.

The benchmark estimator in this framework is the QMLE

$$\widehat{\theta}_{QMLE} = \arg\min_{\theta\in\Theta} \frac{1}{T} \sum_{t=t_0+1}^{T} \frac{y_t^2}{\widehat{f}_t(\theta)} + \log \widehat{f}_t(\theta). \tag{36}$$

As discussed in Section 3.1, one can also use an alternative QLE based on the estimating functions theory:

$$\widehat{\theta}_T = \arg\min_{\theta\in\Theta} \left\| \frac{1}{T} \sum_{t=t_0+1}^{T} \frac{y_t^2 - \widehat{f}_t(\theta)}{\widehat{\sigma}_t^2(\theta)} \frac{\partial \widehat{f}_t(\theta)}{\partial \theta} \right\| \tag{37}$$

for some function $\widehat{\sigma}_t^2(\theta) > 0 \in \mathcal{F}_{t-1}$. If $\widehat{\sigma}_t^2(\theta)$ is chosen proportional to $\widehat{f}_t^2(\theta)$, then the two estimators (36) and (37) are equivalent, but they are not if, for instance, one takes $\widehat{\sigma}_t^2(\theta) = \widehat{f}_t(\theta)$.

# 5 Small Sample Properties and empirical application

In this section, we present a Monte Carlo experiment that studies the finite-sample properties of the $\psi_T GAS - T$ model as well as an application on real data.

## 5.1 Monte Carlo Simulation

In the simulation study, we consider three data generating processes (DGPs) corresponding to particular cases of the following $\psi_T GAS(1,1) - T$ model:

$$y_t = \mu + \sqrt{f_t}\epsilon_t \tag{38}$$

$$\epsilon_t \sim T(0,1,1/\xi) \tag{39}$$

$$f_{t+1} = \omega + \alpha \frac{1+\zeta}{1-2\zeta+\zeta\epsilon_t^2}\epsilon_t^2 f_t + \beta f_t \tag{40}$$

with $0 < \xi < 1/2$ and $-1 < \zeta < 1/2$. Recall that this model is a $\beta_T GAS(1,1)$ when $\xi = \zeta$ and a $GARCH(1,1) - T$ model when $\zeta = 0$.

In all simulations, we set $\mu = 0, \omega = 0.03, \alpha = 0.13$ and $\beta = 0.84$. In the first simulation (i.e., Table 1), we set $\xi = \zeta = 0.2$ so that the model is a $\beta_T GAS(1,1)$ model with a degree of freedom of $\xi^{-1} = 5$. In the second simulation (i.e., Table 2), we set $\xi = 0.2$ and $\zeta = 0.1$ so that the model is a $\psi_T GAS - T$ with a higher degree of freedom in the conditional variance equation than for the density of the innovations. Finally, in the third simulation (i.e., Table 3), $\xi = 0.2$ while $\zeta = 0$ so that the true model is a $GARCH(1,1) - T$ model.

In all cases, four models are estimated. Three models (i.e., $\psi_T GAS(1,1) - T$, $\beta_T GAS(1,1)$ and $GARCH(1,1) - T$) are estimated by ML. The fourth model is the $\psi_T GAS(1,1)$ estimated by Gaussian QML (and therefore $\xi$ is not estimated). Note that during the optimization, the positivity of the conditional variance of the $\psi_T GAS$ models is imposed by replacing (40) by (29), as discussed in Section 4.

In all the cases considered in this section, $\sigma_t^2$ is proportional to $f_t^2$ so that the optimal QLE corresponds to the Gaussian QMLE, which is the reason why specific results for the QLE are not reported below.

Each of the three tables is divided in two major parts. The top panels correspond to the results for a sample size of 3,000 observations, while the bottom panels are for 4,000 observations. Each panel is again divided in two parts. The first one contains summary statistics on the estimated parameters, while the second reports rejection frequencies of two LR tests (LRT). Figures at the right of the name of the models are the empirical biases over 1,000 replications. Figures in parenthesis correspond to RMSEs, while those in squared brackets are the 95% coverage probabilities (i.e., percentage of 95% confidence intervals drawn from the asymptotic distribution containing the true parameter). The second part contains rejection frequencies of two LR tests computed from the ML estimates. The first one is for the null hypothesis that the true model is a $\beta_T GAS(1,1)$, i.e., $\xi = \zeta$, while the second test is for the null hypothesis that the model is a $GARCH(1,1) - T$, i.e., $\xi = 0$. Note that some of the figures reported in this part correspond to empirical sizes or powers depending on the DGP.

Some comments are in order.

- The most important result is that the bias of the MLEs of the $\psi_T GAS(1,1)-T$ is negligible for the two considered sample sizes and the three DGPs.

- When the true model is a $\beta_T GAS(1,1)$ (see Table 1), the (inverse of) the degree of freedom of innovations $\xi$ is slightly more precisely estimated with the $\beta_T GAS(1,1)$ model than with the $\psi_T GAS(1,1)-T$, but the difference is marginal. Indeed, the RMSE is only 0.001 higher for the latter. Furthermore, while the biases of $\xi$ and $\zeta$ are small, the RMSE of $\zeta$ is between three and four times higher than for $\xi$. This is a consequence of the fact that the identification of $\zeta$ is only possible from the observations for which the shocks are truncated, whereas all observations can be used to identify $\xi$. Testing the null hypothesis that $\xi = \zeta$ is therefore desirable to gain efficiency by imposing this restriction when the null hypothesis is not rejected.

- As expected, some of the parameters of the $GARCH(1,1)-T$ model (especially $\alpha$) are biased when wrongly imposing the assumption that $\zeta = 0$, as shown in Tables 1 and 2.

- Similarly, some of the parameters of the $\beta_T GAS(1,1)$ model are biased when the true model is a $\psi_T GAS(1,1)-T$ with $\xi \neq \zeta$, as shown in Tables 2 and 3.

- As expected again, the QML of the $\psi_T GAS(1,1)-T$ is less precise than its ML version. The bias is higher than for the ML, while the RMSE is approximately 20-25% higher.

- The coverage probabilities of the parameters of the $\psi_T GAS(1,1)-T$ are satisfactory except for $\zeta$. For a sample size of 3,000 observations, the true value of $\zeta$ belongs to the 95% confidence interval drawn from the asymptotic distribution in approximately 85 to 89% of the cases either for the MLE or the QMLE estimators. The results are slightly better for a sample size of 4,000 observations. Unreported simulation results suggest that a sample size of at least 15,000 observations is needed to perform correct statistical inference on $\zeta$ on the basis of t-tests and confidence intervals relying on the asymptotic distribution. For the sample sizes considered in Tables 1 to 3, standard errors of $\zeta$ are on average too small compared to the RMSE of the estimated $\zeta$ parameter.

- While statistical inference on $\zeta$ relying on its standard error (e.g., t-tests and Wald tests) requires a very large sample, the LRT on $\zeta$ has good finite sample properties. Indeed, when the sample size is 4,000, the rejection frequencies of the null hypothesis $H_0 : \xi = \zeta$ in Table 1 (where $\xi = \zeta = 0.2$ in the DGP) and of the null hypothesis $H_0 : \zeta = 0$ in Table 3 are close their nominal

sizes. The rejection frequencies for the other tests correspond to empirical powers. Interestingly, the LRT of the null hypothesis $H_0 : \zeta = 0$ has very high power to reject the GARCH for which the squared shocks drive the dynamic of the conditional variance (see Tables 1 and 2), while the LRT of the null hypothesis $H_0 : \xi = \zeta$ has very high power when the true model is a $GARCH(1,1)$ (see Table 3) and decent power when the true model is a $\psi_T GAS(1,1) - T$ with $\xi = 0.2$ and $\zeta = 0.1$ (approximately 43% for a nominal size of 5% and a sample size of 4,000 observations). The power of course increases with the distance between $\xi$ and $\zeta$.

Finally, to help visualize the impact of a misspecification of the conditional variance, Figure 1 plots 50 observations around a large shock. The DGP is a $\psi_T GAS(1,1) - T$ with $\xi = 0.2$ and $\zeta = 0.1$ (as in Table 2), and the models are estimated by ML on 4,000 observations. This figure plots the absolute value of the simulated log-returns (thin solid red line) as well as the estimated conditional volatilities of the $\psi_T GAS - T$ (thick solid pink line), $\beta_T GAS$ (thin green dashed line), GARCH-$T$ (thin blue line with long dashes) and the true conditional volatility (black solid line). It is clear from this graph that unlike the $\psi_T GAS - T$, the GARCH model overestimates the volatility during approximately two weeks (i.e., approximately 15 observations) following the large shock (occurring at observation 1475), while the $\beta_T GAS$ underestimates the volatility during the same period.

## 5.2 Empirical Application

In the empirical application, we consider all stocks belonging to the S&P500 index for the period spanning from 03-01-1995 (or later) to 28-02-2019. All stocks for which less than 4,000 observations are available have been discarded so that we are left with 408 stocks. The four volatility models used in the previous section are considered in this empirical application, i.e., the $\psi_T GAS - T$ estimated by ML and QML and the $\beta_T GAS$ and $GARCH - T$ estimated by ML.

The stationarity and invertibility conditions seem to be satisfied for all series according to conditions $(ii)$ of Lemma 1 and $(ii)$ of Lemma 3 evaluated at the MLE estimates of the parameters. Interestingly, the null hypothesis $\zeta = 0$ is rejected in 94.4% of the cases using an LRT (at the 5% nominal size), suggesting that downweighting of large shocks that estimate the conditional variance of these US stocks is empirically relevant. These results naturally call for the use of a GAS-type model rather than a GARCH dynamic. However, the null hypothesis $\xi = \zeta$ is rejected in 36.8% of the cases (again at the 5% nominal level) suggesting that the additional flexibility of the $\psi_T GAS - T$ over the $\beta_T GAS$ is needed in more than one-third of the cases. Furthermore all the estimated $\zeta$s are positive except for two stocks; however, in these two cases, the null hypothesis $\zeta = 0$ is not rejected, so that a GARCH(1,1) dynamic is preferable for these two series.

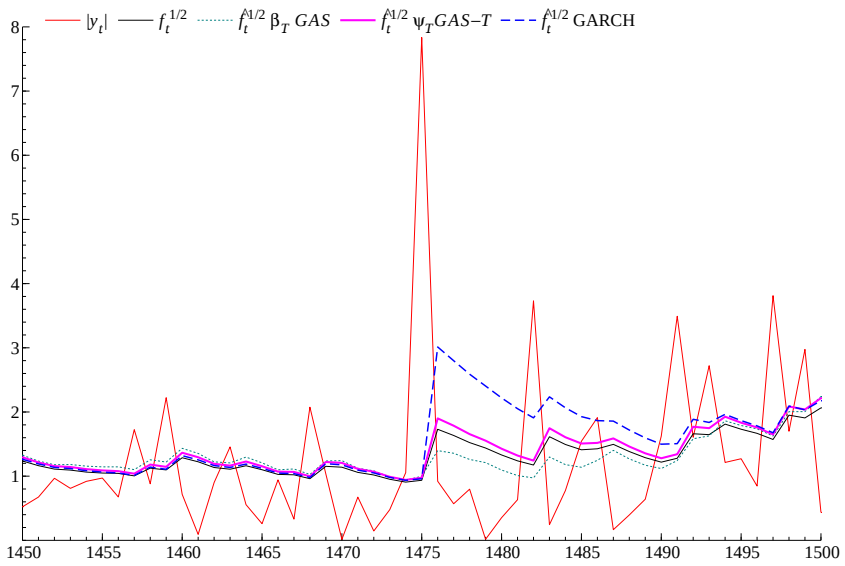


Figure 1: Fifty observations around a large shock for a DGP corresponding to a $\psi_T GAS(1,1) - T$ with $\xi = 0.2$ and $\zeta = 0.1$ (as in Table 2). The $\psi_T GAS - T$, $\beta_T GAS$ and GARCH-$T$ models are estimated by ML on 4,000 observations.

The difference $\widehat{\xi} - \widehat{\zeta}$ is plotted in Figure 2 for the 408 US stocks (sorted in alphabetical order of the ticker's name). A full (resp. empty) circle corresponds to a significant (resp. insignificant) difference (according to an LRT at the 5% nominal level). For all stocks for which $\widehat{\xi} \neq \widehat{\zeta}$, $\widehat{\xi} > \widehat{\zeta}$, suggesting that the downweighting of the $\beta_T GAS$ is too strong.

To visualize the added value of the $\psi_T GAS - T$ model over the GARCH and $\beta_T GAS$ models, we randomly selected a stock for which both the null hypotheses $H_0 : \xi = \zeta$ and $H_0 : \zeta = 0$ are rejected. We choose CenterPoint Energy (whose ticker is CNP), a domestic energy delivery company that includes electric transmission and distribution, natural gas distribution and energy services operations.

The MLEs of the GARCH-$T$, $\beta_T$ $GAS$ and $\psi_T GAS - T$ obtained on daily log-returns of CNP during the period spanning from January 1995 to the end of February 2019 (i.e., 6081 observations) are reported in Table 4 together with the log-likelihood value and the outcome of the two LRT tests presented above (with the corresponding p-values in squared brackets). Interestingly, the estimated $\xi$ parameters of the GARCH-$T$ and $\beta_T$ $GAS$ do not differ much and are at approximately 0.17, which corresponds to a degree of freedom of the Student's t-distribution just below 6. However, the log-likelihood of the $\beta_T$ $GAS$ is 4.5 points below the one of the GARCH-$T$ model. Given that the two models have the same number of parameters and are not nested, LRT cannot be employed to discriminate

Table 1: Bias, RMSE and 95% coverage probabilities and LRT. The DGP is a $\beta_T GAS(1,1)$. Sample size $T = 3,000$ or $4,000$.

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | $T = 3,000$ | | | |
| | $\mu$ | $\omega$ | $\alpha$ | $\beta$ | $\xi$ | $\zeta$ |
| | 0 | 0.03 | 0.13 | 0.84 | 0.2 | 0.2 |
| $\psi_T GAS - T$ ML | 0.000 | 0.002 | -0.001 | -0.004 | -0.001 | 0.002 |
| | (0.014) | (0.009) | (0.018) | (0.027) | (0.018) | (0.065) |
| | [0.963] | [0.940] | [0.942] | [0.942] | [0.951] | [0.899] |
| $\psi_T GAS - T$ QML | 0.000 | 0.002 | 0.001 | -0.008 | | 0.012 |
| | (0.017) | (0.013) | (0.027) | (0.040) | | (0.082) |
| | [0.951] | [0.917] | [0.928] | [0.915] | | [0.846] |
| $\beta_T GAS$ ML | 0.000 | 0.002 | 0.001 | -0.003 | -0.000 | |
| | (0.014) | (0.010) | (0.018) | (0.023) | (0.017) | |
| | [0.965] | [0.948] | [0.948] | [0.949] | [0.956] | |
| $GARCH - T$ ML | 0.000 | 0.009 | -0.038 | 0.032 | 0.005 | |
| | (0.014) | (0.010) | (0.018) | (0.023) | (0.017) | |
| | [0.962] | [0.871] | [0.313] | [0.519] | [0.928] | |
| | 1% | 5% | 10% | | | |
| $H_0 : \xi = \zeta$ | 1.403 | 7.014 | 13.627 | | | |
| $H_0 : \zeta = 0$ | 96.894 | 98.998 | 99.499 | | | |
| | | | $T = 4,000$ | | | |
| | $\mu$ | $\omega$ | $\alpha$ | $\beta$ | $\xi$ | $\zeta$ |
| | 0 | 0.03 | 0.13 | 0.84 | 0.2 | 0.2 |
| $\psi_T GAS - T$ ML | 0.000 | 0.002 | -0.001 | -0.003 | -0.000 | 0.005 |
| | (0.012) | (0.008) | (0.016) | (0.023) | (0.016) | (0.057) |
| | [0.965] | [0.948] | [0.936] | [0.946] | [0.952] | [0.907] |
| $\psi_T GAS - T$ GARCH | 0.000 | 0.002 | -0.000 | -0.007 | | 0.013 |
| | (0.015) | (0.012) | (0.023) | (0.034) | | (0.074) |
| | [0.959] | [0.930] | [0.928] | [0.933] | | [0.861] |
| $\beta_T GAS$ ML | 0.000 | 0.002 | 0.000 | -0.002 | 0.000 | |
| | (0.012) | (0.008) | (0.016) | (0.020) | (0.015) | |
| | [0.964] | [0.953] | [0.942] | [0.943] | [0.953] | |
| GARCH$-T$ ML | 0.000 | 0.007 | -0.039 | 0.035 | 0.005 | |
| | (0.012) | (0.008) | (0.016) | (0.020) | (0.015) | |
| | [0.963] | [0.881] | [0.195] | [0.414] | [0.936] | |
| | 1% | 5% | 10% | | | |
| $H_0 : \xi = \zeta$ | 1.300 | 6.100 | 12.400 | | | |
| $H_0 : \zeta = 0$ | 99.100 | 99.700 | 99.900 | | | |

Note: Monte Carlo simulation results for $T = 3,000$ (top panel) and $T = 4,000$ (bottom panel). Each panel is divided into two parts. The first part is for the estimated parameters of 4 models. Figures at the right of the name of the models are the empirical biases over 1,000 replications. Figures in parenthesis correspond to RMSEs, while those in squared brackets are the 95% coverage probabilities. The second part contains rejection frequencies of two LR tests computed from the ML estimates. Some of the figures reported in this part correspond to empirical sizes or powers depending on the DGP.

between these two models, but this result suggests that the $\beta_T$ $GAS$ underperforms with respect to the GARCH-$T$ model. Importantly, while $\widehat{\xi}$ is also close to 0.17 for the $\psi_T GAS - T$ model, $\widehat{\zeta}$ is approximately 0.04 (and therefore $1/\widehat{\zeta}$ is close to 25), suggesting that the $\beta_T$ $GAS$ downweights the large shocks far too much. To help visualize the difference between the three models, the news impact curve (NIC) of each estimated model is plotted in Figure 3. The NIC measures how new information is incorporated into the conditional variance. Since the $\psi_T GAS - T$ nests the other two models, we can write the NIC of the three models as the function

Table 2: Bias, RMSE and 95% coverage probabilities and LRT. The DGP is a $\psi_T GAS(1,1)$. Sample size $T = 3,000$ or $4,000$.

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | $T = 3,000$ | | | |
| | $\mu$ | $\omega$ | $\alpha$ | $\beta$ | $\xi$ | $\zeta$ |
| | 0 | 0.03 | 0.13 | 0.84 | 0.2 | 0.1 |
| | $\mu$ | $\omega$ | $\alpha$ | $\beta$ | $\xi$ | $\zeta$ |
| $\psi_T GAS - T$ ML | 0.000 | 0.002 | -0.001 | -0.005 | -0.001 | 0.009 |
| | (0.012) | (0.009) | (0.020) | (0.027) | (0.018) | (0.058) |
| | [0.966] | [0.951] | [0.937] | [0.944] | [0.951] | [0.867] |
| $\psi_T GAS - T$ QML | 0.000 | 0.002 | 0.001 | -0.010 | | 0.025 |
| | (0.015) | (0.012) | (0.030) | (0.041) | | (0.081) |
| | [0.954] | [0.908] | [0.922] | [0.918] | | [0.837] |
| $\beta_T GAS$ ML | 0.000 | 0.002 | 0.009 | -0.021 | -0.006 | |
| | (0.012) | (0.009) | (0.021) | (0.033) | (0.018) | |
| | [0.964] | [0.942] | [0.942] | [0.903] | [0.927] | |
| GARCH-$T$ ML | 0.000 | 0.006 | -0.029 | 0.015 | 0.002 | |
| | (0.012) | (0.009) | (0.021) | (0.033) | (0.018) | |
| | [0.964] | [0.903] | [0.527] | [0.786] | [0.949] | |
| | 1% | 5% | 10% | | | |
| $H_0 : \xi = \zeta$ | 18.838 | 38.577 | 49.900 | | | |
| $H_0 : \zeta = 0$ | 74.649 | 87.074 | 91.182 | | | |
| | | | $T = 4,000$ | | | |
| | $\mu$ | $\omega$ | $\alpha$ | $\beta$ | $\xi$ | $\zeta$ |
| | 0 | 0.03 | 0.13 | 0.84 | 0.2 | 0.1 |
| $\psi_T GAS - T$ ML | 0.000 | 0.001 | -0.001 | -0.004 | -0.001 | 0.008 |
| | (0.010) | (0.007) | (0.018) | (0.023) | (0.016) | (0.050) |
| | [0.966] | [0.955] | [0.932] | [0.942] | [0.952] | [0.895] |
| $\psi_T GAS - T$ QML | 0.000 | 0.002 | 0.001 | -0.007 | | 0.021 |
| | (0.013) | (0.011) | (0.025) | (0.034) | | (0.070) |
| | [0.953] | [0.939] | [0.927] | [0.950] | | [0.871] |
| $\beta_T GAS$ ML | 0.000 | 0.001 | 0.008 | -0.019 | -0.006 | |
| | (0.010) | (0.007) | (0.019) | (0.029) | (0.017) | |
| | [0.965] | [0.942] | [0.931] | [0.884] | [0.922] | |
| GARCH-$T$ ML | 0.000 | 0.005 | -0.030 | 0.017 | 0.002 | |
| | (0.010) | (0.007) | (0.019) | (0.029) | (0.017) | |
| | [0.966] | [0.908] | [0.412] | [0.738] | [0.949] | |
| | 1% | 5% | 10% | | | |
| $H_0 : \xi = \zeta$ | 24.400 | 44.200 | 57.200 | | | |
| $H_0 : \zeta = 0$ | 88.200 | 95.300 | 96.900 | | | |

Note: see Table 1

mapping the shocks $\epsilon_t$ to $\frac{1+\zeta}{1-2\zeta+\zeta\epsilon_t^2}\epsilon_t^2$, where $\zeta = 0$ for the GARCH model and $\zeta = \xi$ for the $\beta_T$ $GAS$. We see from Figure 3 that the NIC of the $\psi_T GAS - T$ for the CNP stock lies between the NIC of the other two models.

Finally, to see the impact of different NICs on the estimated conditional volatilities, the absolute value of the daily log-returns of CNP (thin solid red line) as well as the estimated conditional volatilities of the $\psi_T GAS - T$ (thick solid pink line), $\beta_T GAS$ (thin green dashed line) and GARCH-$T$ (thin blue line with long dashes) estimated by ML (on the full period) are plotted for the sub-period spanning from January 2002 to December 2002 in Figure 4. The three boxes highlight periods between 3 and 6 weeks around very large shocks (i.e., high absolute returns in %).

Table 3: Bias, RMSE and 95% coverage probabilities and LRT. The DGP is a $GARCH(1,1)-T$. Sample size $T = 3,000$ or $4,000$.

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | $T = 3,000$ | | | |
| | $\mu$ | $\omega$ | $\alpha$ | $\beta$ | $\xi$ | $\zeta$ |
| | 0 | 0.03 | 0.13 | 0.84 | 0.2 | 0 |
| | $\mu$ | $\omega$ | $\alpha$ | $\beta$ | $\xi$ | $\zeta$ |
| $\psi_T GAS-T$ ML | 0.000 | 0.001 | -0.000 | -0.002 | -0.001 | 0.003 |
| | (0.012) | (0.007) | (0.019) | (0.021) | (0.018) | (0.015) |
| | [0.962] | [0.938] | [0.939] | [0.941] | [0.948] | [0.852] |
| $\psi_T GAS-T$ QML | 0.000 | 0.002 | 0.003 | -0.006 | | 0.009 |
| | (0.015) | (0.010) | (0.029) | (0.032) | | (0.028) |
| | [0.953] | [0.920] | [0.926] | [0.918] | | [0.845] |
| $\beta_T GAS$ | 0.000 | 0.001 | 0.045 | -0.051 | -0.015 | |
| | (0.012) | (0.007) | (0.050) | (0.057) | (0.023) | |
| | [0.959] | [0.927] | [0.387] | [0.395] | [0.825] | |
| GARCH-$T$ ML | 0.000 | 0.001 | 0.000 | -0.002 | -0.001 | |
| | (0.012) | (0.007) | (0.050) | (0.057) | (0.023) | |
| | [0.962] | [0.941] | [0.943] | [0.950] | [0.947] | |
| | 1% | 5% | 10% | | | |
| $H_0 : \xi = \zeta$ | 97.998 | 99.800 | 99.900 | | | |
| $H_0 : \zeta = 0$ | 2.102 | 9.109 | 14.114 | | | |
| | | | $T = 4,000$ | | | |
| | $\mu$ | $\omega$ | $\alpha$ | $\beta$ | $\xi$ | $\zeta$ |
| | 0 | 0.03 | 0.13 | 0.84 | 0.2 | 0 |
| $\psi_T GAS-T$ ML | 0.000 | 0.001 | -0.001 | -0.001 | -0.001 | 0.002 |
| | (0.010) | (0.006) | (0.017) | (0.017) | (0.016) | (0.012) |
| | [0.965] | [0.946] | [0.932] | [0.948] | [0.947] | [0.883] |
| $\psi_T GAS-T$ QML | 0.000 | 0.002 | 0.002 | -0.004 | | 0.005 |
| | (0.012) | (0.009) | (0.025) | (0.026) | | (0.018) |
| | [0.957] | [0.931] | [0.933] | [0.946] | | [0.880] |
| $\beta_T GAS$ | 0.000 | 0.000 | 0.045 | -0.050 | -0.014 | |
| | (0.010) | (0.006) | (0.049) | (0.055) | (0.021) | |
| | [0.962] | [0.931] | [0.274] | [0.302] | [0.805] | |
| GARCH-$T$ ML | 0.000 | 0.001 | -0.001 | -0.001 | -0.001 | |
| | (0.010) | (0.006) | (0.049) | (0.055) | (0.021) | |
| | [0.965] | [0.948] | [0.941] | [0.951] | [0.944] | |
| | 1% | 5% | 10% | | | |
| $H_0 : \xi = \zeta$ | 99.700 | 99.900 | 99.900 | | | |
| $H_0 : \zeta = 0$ | 1.802 | 5.606 | 11.311 | | | |

Note: see Table 1

Regarding the NICs, the conditional volatility of the $\psi_T GAS-T$ lies between the one of the other two models and, importantly, is closer to the absolute returns following the large peaks. It is also clear from this graph that, for this series, the GARCH model over-estimates the volatility following the large shocks.

# 6 Conclusion

*GAS* models have received considerable attention in the time series literature. In this paper, we point out a major limitation of this general class of models that imposes a strong link between the conditional distribution of $y_t$ and the updating

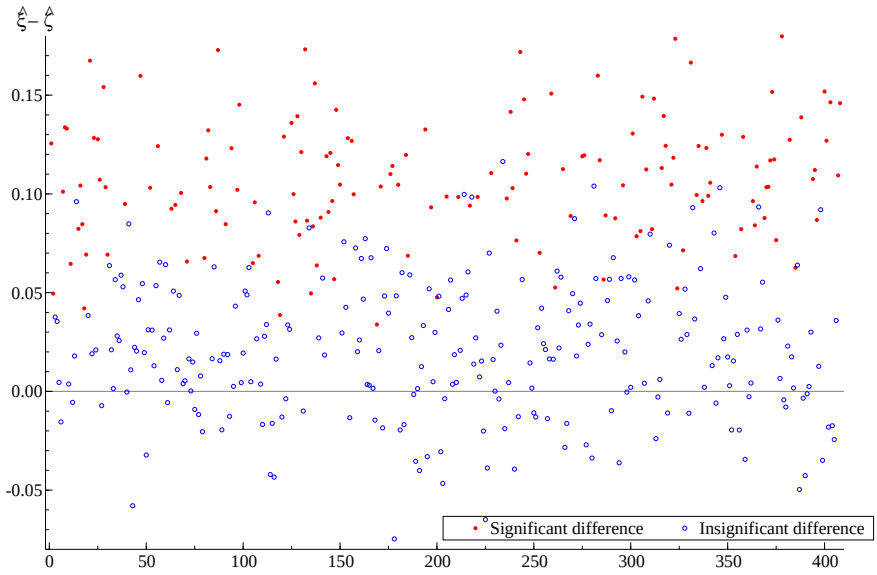


Figure 2: $\widehat{\xi}-\widehat{\zeta}$ for the $\psi_T GAS(1,1)-T$ estimated on the 408 US stocks. A full (resp. empty) circle corresponds to a significant (resp. insignificant) difference (according to an LRT at the 5% nominal level).

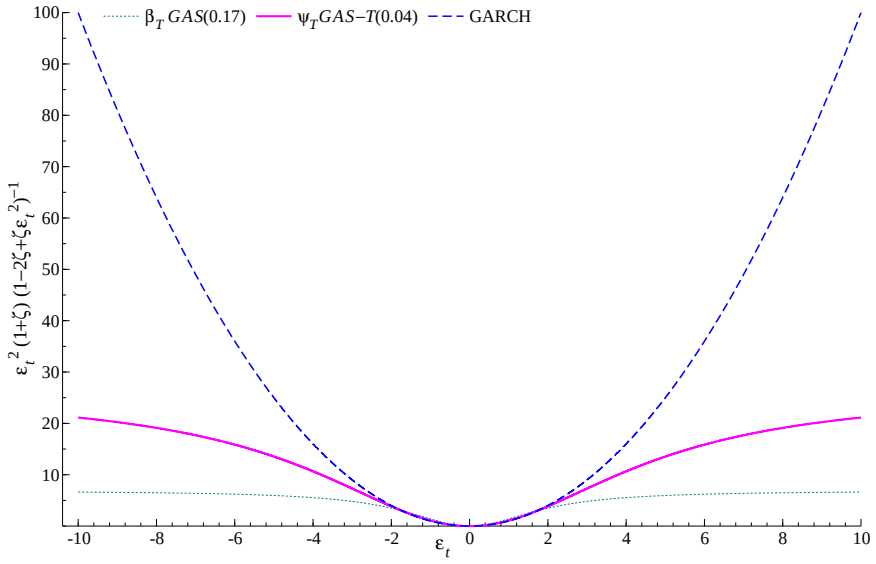


Figure 3: New impact curve for the MLEs of the $\psi_T GAS{-}T$, $\beta_T GAS$ and GARCH-$T$ for the ticker CNP (CenterPoint Energy).

Table 4: MLEs of the GARCH-$T$, $\beta_T$ $GAS$ and $\psi_T GAS-T$ for CNP during the period spanning from January 1995 to the end of February 2019 (i.e., 6,081 observations).

| | GARCH-$T$ | $\beta_T$ $GAS$ | $\psi_T GAS-T$ |
|---|---|---|---|
| $\mu$ | 0.0888 | 0.0914 | 0.0894 |
| | (0.0152) | (0.0150) | (0.0151) |
| $\omega$ | 0.0606 | 0.0428 | 0.0497 |
| | (0.0094) | (0.0084) | (0.0088) |
| $\alpha$ | 0.0951 | 0.1272 | 0.1069 |
| | (0.0089) | (0.0102) | (0.0113) |
| $\beta$ | 0.8817 | 0.8573 | 0.8791 |
| | (0.0097) | (0.0115) | (0.0113) |
| $\xi$ | 0.1703 | 0.1677 | 0.1722 |
| | (0.0095) | (0.0088) | (0.0094) |
| $\zeta$ | | | 0.0400 |
| | | | (0.0124) |
| Log-Likelihood | -10732.7 | -10737.2 | -10726.6 |
| $H_0:\xi=\zeta$ | 21.2 | | |
| | [0.00000] | | |
| $H_0:\zeta=0$ | 12.2 | | |
| | [0.00047] | | |

Note: The figures at the right of $H_0:\xi=\zeta$ and $H_0:\zeta=0$ are the values of the LRT corresponding to the specified null hypothesis (with the p-value below in squared brackets).

equation of $f_t$. We therefore propose a more general family of models called $\psi GAS$ that overcomes this problem.

We study the statistical properties of the $\psi GAS$ filter as well as the QLE, QMLE and MLE of the parameters of this model. We show how to test the relevance of some of the constraints in the $GAS$ models, linking $f_t$ to $p_t(y_t|f_t,\theta)$.

We study in detail the $\psi_T GAS-T$ model, a volatility model extending the $\beta_T GAS$ model of Harvey and Chakravarty (2008). This model relies on a standardized Student's t-density for the innovations and the score of a standardized Student's t-density in the updating equation of the conditional variance but does not restrict the degrees of freedom to be the same. The additional flexibility of this model (over the $\beta_T GAS$) is found to be significant at the 5% significance level using a standard LRT in more than one-third of the cases (out of more than 400 stocks).

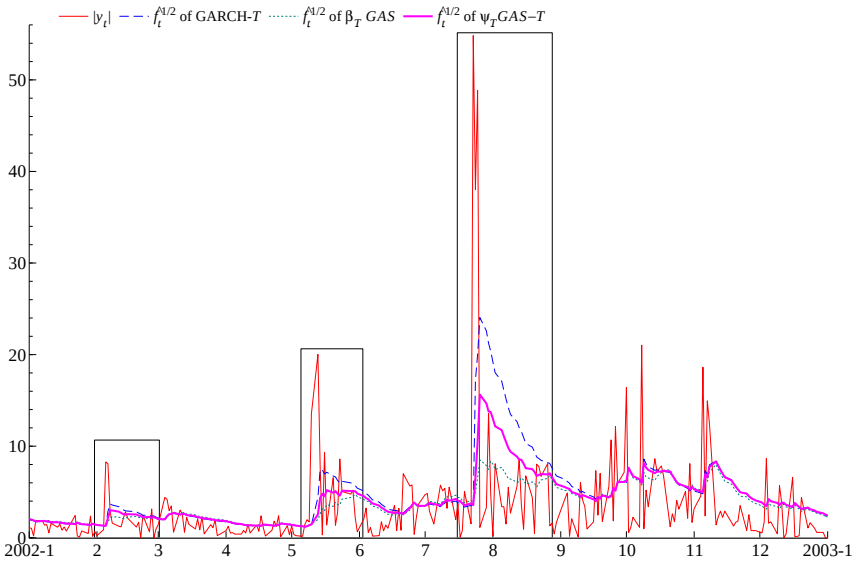


Figure 4: Absolute value of log-returns and estimated conditional volatility of the $\psi_T GAS - T$, $\beta_T GAS$ and GARCH-$T$ estimated by ML for the ticker CNP (CenterPoint Energy). The graph only shows the sub-period spanning from January 2002 to December 2002.

# References


[1] Bera, A.K. and Bilias, Y. (2002) The MM, ME, ML, EL, EF and GMM approaches to estimation: A synthesis. *Journal of Econometrics* 107, 51–86.

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

[3] Blasques F., Koopman S.J., and Lucas, A. (2014) Maximum Likelihood Estimation for Correctly Specified Generalized Autoregressive Score Models: Feedback Effects, Contraction Conditions and Asymptotic Properties. *Tinbergen institute discussion paper.*

[4] Blasques F., Koopman S.J., and Lucas, A. (2015) Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika* 102, 325–343.

[5] Bougerol, P. and Picard, N. (1992) Strict stationarity of generalized autoregressive processes. *Annals of Probability* 20, 1714–1729.

[6] Brandt, A.(1986) The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advance in Applied Probability* 18, 221–254.

[7] Chandra, A.S. and Taniguchi, M. (2001) Estimating functions for non-linear time series models. *Annals of the Institute of Statistical Mathematics* 53, 125–141.

[8] Charbonnier, P., Blanc-Feraud, L. Aubert, G. and Barlaud, M. (1997) Deterministic edge-preserving regularization in computed imaging. *IEEE Trans. Image Processing* 6, 298–311.

[9] Creal, D.D., Koopman, S.J. and Lucas A. (2012) A General Framework for Observation Driven Time-Varying Parameter Models. *Journal of Applied Econometrics* 28, 5, 777–795.

[10] Davidson, J. (1994) Stochastic limit theory: An introduction for econometricians. Oxford University Press.

[11] Davis, R.A., Dunsmuir, W.T.M. and Streett S.B. (2003) Observation-driven models for Poisson counts. *Biometrika* 90, 777–790.

[12] Domowitz, I. and White, H. (1982) Misspecified models with dependent observations. *Journal of Econometrics* 20, 35–58.

[13] Durbin, J. (1960) Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society Series B* 22, 139–153.

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

[15] Francq, C. and J-M. Zakoian (2019) GARCH models: structure, statistical inference and financial applications. Chichester: John Wiley, second edition.

[16] Freedman, D.A. and Diaconis, P.(1982) On Inconsistent $M$-Estimators. *Ann. Statist.* 10, 454–461.

[17] Godambe, V.P. (1960) An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics* 31, 1208–1212.

[18] Godambe, V. P. (1985) The foundations of finite sample estimation in stochastic processes. *Biometrika* 72, 419–428.

[19] Godambe, V.P. and Heyde, C.C. (1987) Quasi-likelihood and optimal estimation. *International Statistical Review* 55, 231–244.

[20] Gouriéroux, C., Monfort, A., and Trognon, A. (1984) Pseudo maximum likelihood methods: Theory. *Econometrica* 52, 681–700.

[21] Granger, C.W.J. (1999) Outline of Forecast Theory Using Generalized Cost Functions. *Spanish Economic Review* 1, 161-173.

[22] Hartley, R. and Zisserman, A. (2003) Multiple View Geometry in Computer Vision (2nd ed.). Cambridge University Press.

[23] Harvey, A. C. and Chakravarty, T. (2008) Beta-t-(E)GARCH. *Discussion Paper* University of Cambridge CWPE 08340.

[24] Heyde, C.C. (2008) *Quasi-likelihood and its application: a general approach to optimal parameter estimation.* Springer Science & Business Media.

[25] Jacod, J. and Sørensen, M. (2018) A review of asymptotic theory of estimating functions. *Statistical Inference for Stochastic Processes* 21, 415–434.

[26] Kabaila P. (1983) Parameter values of ARMA models minimising the one-step-ahead prediction error when the true system is not in the model set. *Journal of Applied Probability* 20, 405–408.

[27] Lecourt, C., Laurent, S. and Palm, F. (2016) Testing for Jumps in ARMA-GARCH Models, a Robust Approach. *Computational Statistics and Data Analysis* 100, 383–400.

[28] Pedersen, R.S., and Rahbek, A. (2019) Testing GARCH-X type models. *Econometric Theory* 35, 1012–1047.

[29] Potscher, B.M. and Prucha, I. R. (1997) Dynamic Nonlinear Statistical Models: Asymptotic Theory. Spinger-Verlag, Berlin.

[30] Rao, R.R. (1962) Relations between Weak and Uniform Convergence of Measures with Applications. *Ann. Math. Statist.* 33, 659–680.

[31] Varian, H.R. (1975) A Bayesian Approach to Real Estate Assessment, in Studies in Bayesian Econometrics and Statistics in Honor of L.J. Savage, eds. S.E. Fienberg and A. Zellner, Amsterdam: North Holland, 195–208.

[32] Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61, 439–447.

[33] White, H. (1982) Maximum Likelihood Estimation of Misspecified Models. *Econometrica* 50, 1–25.

[34] White, H. (1994) Estimation, Inference and Specification Analysis. Cambridge Books. Cambridge University Press.

[35] Zellner, A. (1986) Bayesian Estimation and Prediction Using Asymmetric Loss Functions. *Journal of the American Statistical Association* 81, 446–451.

# A Proofs

## Proof of Proposition 1

The first claim follows trivially by noting that

$$\begin{aligned}\rho(y_t, f_{t+1}, \theta) - \rho(y_t, f_t, \theta) &= \psi(y_t, f_t^*, \theta)(f_{t+1} - f_t) \\ &= \alpha\psi(y_t, f_t^*, \theta)\psi(y_t, f_t, \theta) + o(1) \\ &= \alpha\psi(y_t, f_t, \theta)^2 + o(1) > 0,\end{aligned}$$

where the first equality is an application of the mean value theorem, the second equality is obtained since $f_{t+1} - f_t = \omega + \alpha\psi(y_t, f_t, \theta) + (\beta - 1)f_t$ with $\omega + (\beta - 1)f_t = o(1)$, the third equality follows by continuity of $\psi$ and hence writing $\psi(y_t, f_t^*, \theta)^2 = \psi(y_t, f_t, \theta)^2 + o(1)$ as $f_t \to f_t^*$. Finally, the inequality is obtained by setting $\omega$, $\beta - 1$ and $f_{t+1} - f_t$ small enough such that the inequality holds.

The second claim is easily achieved since

$$\begin{aligned}\rho(y_{t+1}, f_{t+1}, \theta) - \rho(y_t, f_t, \theta) &= \rho(y_{t+1}, f_{t+1}, \theta) - \rho(y_t, f_{t+1}, \theta) \\ &\quad + \rho(y_t, f_{t+1}, \theta) - \rho(y_t, f_t, \theta) \\ &= \rho'_\eta(y_{t+1}, f_{t+1}, \theta)(\eta(y_{t+1}) - \eta(y_t)) \\ &\quad + \psi(y_{t+1}, f_t^*, \theta)(f_{t+1} - f_t) \\ &= \rho'_\eta(y_{t+1}, f_{t+1}, \theta) \cdot o(1) + \alpha\psi(y_t, f_t^*, \theta)\psi(y_t, f_t, \theta) + o(1) \\ &= \alpha\psi(y_t, f_t, \theta)^2 + o(1) > 0,\end{aligned}$$

where in the first equality we add and subtract $\rho(y_t, f_{t+1}, \theta)$, the second equality uses the mean-value theorem twice, and the final inequality is obtained by setting $\eta(y_{t+1}) - \eta(y_t)$, $\omega$, $\beta - 1$ and $f_{t+1} - f_t$ small enough.

## Proof of Lemma 1

For all $t \in \mathbb{Z}$ and $n \in \mathbb{N}$, let

$$f_{t+1}^{(n)} = \varphi(z_t, f_t^{(n-1)}) \tag{41}$$

with $f_t^{(0)} = f^0$. Note that

$$f_{t+1}^{(n)} = \varphi_n(z_t, z_{t-1}, \ldots, z_{t-n+1}),$$

for some measurable function $\varphi_n : E^n \to F$. For all fixed $n$, the sequence $(f_t^{(n)})_{t\in\mathbb{Z}}$ is stationary and ergodic. If for all $t$, the limit $f_t = \lim_{n\to\infty} f_t^{(n)}$ exists a.s., then by taking the limit of both sides of (41), it can be seen that the process $(f_t)$ is solution of (5). When it exists, the limit is a measurable function of the form $f_t = \psi_\infty(z_{t-1}, z_{t-2}, \dots)$, and is therefore stationary and ergodic. To show the existence of $\lim_{n\to\infty} f_t^{(n)}$, it sufficies to prove that, a.s., $(f_t^{(n)})_{n\in\mathbb{N}}$ is a Cauchy sequence in the complete space $F$.

By the mean value theorem we have

$$\begin{aligned}\sup_{f,\widetilde{f}\in F,\, f\neq\widetilde{f}} \left|\frac{\varphi(z_t,f)-\varphi(z_t,\widetilde{f})}{f-\widetilde{f}}\right| &\le \Lambda_t := \sup_{f\in F}\left|\frac{\partial\varphi(z_t,f)}{\partial f}\right| \\ &= \sup_{f\in F}\left|\alpha\frac{\partial\psi(g(f,\epsilon_t),X_t,f,\theta)}{\partial f}+\beta\right|.\end{aligned}$$

It follows that

$$\left|\frac{f_{t+1}^{(n)}-f_{t+1}^{(n-1)}}{f_t^{(n-1)}-f_t^{(n-2)}}\right| = \left|\frac{\varphi(z_t,f_t^{(n-1)})-\varphi(z_t,f_t^{(n-2)})}{f_t^{(n-1)}-f_t^{(n-2)}}\right| \le \Lambda_t,$$

and thus

$$\left|f_{t+1}^{(n)}-f_{t+1}^{(n-1)}\right| \le \Lambda_t\left|f_t^{(n-1)}-f_t^{(n-2)}\right| \le \Lambda_t\Lambda_{t-1}\cdots\Lambda_{t-n+2}\left|\varphi(z_{t-n+1},f^0)-f^0\right|.$$

For $n<m$, we then have

$$\begin{aligned}\left|f_{t+1}^{(m)}-f_{t+1}^{(n)}\right| &\le \sum_{k=0}^{m-n-1}\left|f_{t+1}^{(m-k)}-f_{t+1}^{(m-k-1)}\right| \\ &\le \sum_{k=0}^{m-n-1}\Lambda_t\Lambda_{t-1}\cdots\Lambda_{t-m+k+2}\left|\varphi(z_{t-m+k+1},f^0)-f^0\right| \\ &\le \sum_{j=n}^{\infty}\Lambda_t\Lambda_{t-1}\cdots\Lambda_{t-j+1}\left|\varphi(z_{t-j},f^0)-f^0\right|. \qquad (42)\end{aligned}$$

Note that (i) implies that $E\ln^+|\varphi(z_t,f^0)-f^0|<\infty$. Therefore

$$\limsup_{t\to\infty}\frac{\ln|\varphi(z_t,f^0)-f^0|}{t}\le 0 \quad \text{a.s.}$$

The process $(\Lambda_t)$ being stationary and ergodic, (ii) then entails

$$\begin{aligned}&\limsup_{j\to\infty}\ln\left(\Lambda_t\Lambda_{t-1}\cdots\Lambda_{t-j+1}\left|\varphi(z_{t-j},f^0)-f^0\right|\right)^{1/j} \\ =\ &\limsup_{j\to\infty}\frac{1}{j}\sum_{k=1}^{j}\ln\Lambda_{t-k+1}+\frac{\ln|\varphi(z_{t-j},f^0)-f^0|}{j}\le E\ln\Lambda_1<0.\end{aligned}$$

By the Cauchy rule, the right-hand side of (42) tends almost surely to zero as $n \to \infty$. The existence of a stationary and ergodic solution to (5) follows.

Assume that there exists another stationary process $(f_t^*)$ such that $f_{t+1}^* = \varphi(z_t, f_t^*)$. For all $N \geq 0$ we have

$$|f_{t+1} - f_{t+1}^*| \leq \Lambda_t \Lambda_{t-1} \cdots \Lambda_{t-N} |f_{t-N} - f_{t-N}^*|. \tag{43}$$

Since $\Lambda_t \Lambda_{t-1} \cdots \Lambda_{t-N} \to 0$ a.s. as $N \to \infty$, and $|f_{t-N} - f_{t-N}^*| = O_P(1)$ by stationarity, the right-hand side of (43) tends to zero in probability. Since the left-hand side does not depend on $N$, we have $P(|f_{t+1} - f_{t+1}^*| > \varepsilon) = 0$ for all $\varepsilon > 0$, and thus $P(f_{t+1} = f_{t+1}^*) = 1$, which establishes the uniqueness.

## Proof of Lemma 2

By (42) we have

$$\left|f_{t+1} - f^0\right| \leq \left|\varphi(z_t, f^0) - f^0\right| + \sum_{j=1}^{\infty} \Lambda_t \Lambda_{t-1} \cdots \Lambda_{t-j+1} \left|\varphi(z_{t-j}, f^0) - f^0\right|.$$

Note that the variables $\Lambda_t$ are independent, $E \log \Lambda_t < 0$, $E\left|\varphi(z_t, f^0) - f^0\right|^r < \infty$ and $E\Lambda_t^r < \infty$. The arguments of the proof of Lemma 2.3 in Berkes, Horváth and Kokoszka (2003) (see also Corollary 2.3 in Francq and Zakoian, 2019) then entail that there exists $s \in (0, r \wedge 1)$, such that $E\Lambda_t^s < 1$, and thus $E\left|f_{t+1} - f^0\right|^s < \infty$ and the conclusion follows.

## Proof of Lemma 3

The filter satisfies the SRE

$$f_{t+1}(\theta) = \varsigma_\theta(y_t, X_t, f_t(\theta))$$

for some function $\varsigma = \varsigma_\theta$ such that $E \ln^+ |\varsigma(y_t, X_t, f^0) - f^0| < \infty$ and $E \log \Lambda_t(\theta) < 0$ with

$$\Lambda_t(\theta) = \sup_{f \in F} \left| \frac{\partial \varsigma(y_t, X_t, f)}{\partial f} \right| = \sup_{f \in F} \left| \alpha \frac{\partial \psi(y_t, X_t, f, \theta)}{\partial f} + \beta \right|.$$

As in the proof of Lemma 1, the solution of the SRE is obtained by taking the almost sure limit, as $n \to \infty$, of

$$f_{t+1}^{(n)}(\theta) = \varsigma(y_t, X_t, f_t^{(n-1)}(\theta))$$

with $f_t^{(0)}(\theta) = f^0$. Now, note that

$$\sup_{\theta \in \Theta} |f_{t+1}(\theta) - \widehat{f}_{t+1}(\theta)| \leq \Lambda_t \Lambda_{t-1} \cdots \Lambda_1 \sup_{\theta \in \Theta} |f_1(\theta) - \widehat{f}_1(\theta)|,$$

where $\Lambda_t = \sup_{\theta\in\Theta} \Lambda_t(\theta)$. By (ii) one can choose $\varrho$ such that

$$1 > \varrho > e^{E \ln \sup_\theta \Lambda_1} > 0,$$

so that

$$\lim_{t\to\infty} \frac{1}{t} \ln \varrho^{-t} \Lambda_t \Lambda_{t-1} \cdots \Lambda_1 = -\ln \varrho + E \ln \Lambda_1 < 0$$

and the conclusion follows.

## Proof of Lemma 4

Let $\theta$ be a fixed value of the parameter. Under the conditions of Lemma 3, the process $Z_t = (\epsilon_t, X_t^\top, f_t(\theta))^\top$ is stationary and ergodic. The processes $(A_t)$ and $(b_t)$ are thus also stationary and ergodic. The sequence $\{f_t'(\theta)\}_t$ satisfies the linear stochastic difference equation

$$f_{t+1}'(\theta) = A_t + b_t f_t'(\theta),$$

where $(A_t, b_t)$ is strictly stationary and ergodic, and under (i) $E \log^+ \|A_1\| < \infty$ and $E \log^+ |b_1| < \infty$. By Brandt (1986) and Bougerol and Picard (1992), or simply by applying the Cauchy rule, it is known that there exists a stationary, ergodic and non anticipative solution $\left\{f_{t+1}'(\theta)\right\}_t$ to the stochastic difference equation if

$$\gamma := E \log |b_t| < 0,$$

which is implied by (ii) of Lemma 3.

In the sequel, $\varrho$ denotes a generic constant of the interval $(0, 1)$, and $K$ denotes a positive constant or a random variable measurable with respect to $\{z_t, t \le 0\}$. Let

$$\frac{\partial \widehat{\psi}_t}{\partial \theta} = \left.\frac{\partial \psi(y, X, f, \theta)}{\partial \theta}\right|_{(y,X,f,\theta)=(y_t,X_t,\widehat{f}_t(\theta),\theta)}$$

and similar notations for the other derivatives. For $i = 1, \ldots, p$, Taylor expansions show that

$$\frac{\partial \psi_t}{\partial \theta_i} = \frac{\partial \widehat{\psi}_t}{\partial \theta_i} + \left.\frac{\partial^2 \psi(y, X, f, \theta)}{\partial \theta_i \partial f}\right|_{(y,X,f,\theta)=(y_t,X_t,f^*,\theta)} \left\{f_t(\theta) - \widehat{f}_t(\theta)\right\},$$

where $f^*$ is between $f_t(\theta)$ and $\widehat{f}_t(\theta)$. By Lemma 3, we have $|f_t(\theta) - \widehat{f}_t(\theta)| \le K\varrho^t$. Dropping "$(\theta)$" in the notations, other similar Taylor expansions thus show that

$$\left\|A_t - \widehat{A}_t + (b_t - \widehat{b}_t) f_t'\right\| \le K \varrho_t,$$

where $\varrho_t = u_t \varrho^t$ with $E \log^+ u_t < \infty$, using (ii). We thus have

$$\left\| f'_{t+1} - \widehat{f}'_{t+1} \right\| = \left\| A_t - \widehat{A}_t + (b_t - \widehat{b}_t) f'_t + \widehat{b}_t \left( f'_t - \widehat{f}'_t \right) \right\| \leq K \varrho_t + c_t \left\| f'_t - \widehat{f}'_t \right\|,$$

where

$$c_t = |b_t| + K \varrho_t \geq |b_t| + |\widehat{b}_t - b_t| \geq |\widehat{b}_t|.$$

We obtain

$$\left\| f'_{t+1} - \widehat{f}'_{t+1} \right\| \leq K \left\{ \varrho_t + c_t \varrho_{t-1} + \cdots + c_t \cdots c_2 \varrho_1 + c_t \cdots c_1 \| f'_1 - \widehat{f}'_1 \| \right\}.$$

Now note that, by the dominated convergence theorem, $\lim_{\tau \to 0} E \log(|b_1| + \tau) = \gamma < 0$. Therefore, there exists $\tau > 0$ such that

$$\varrho < e^{E \log(|b_1| + \tau)} < 1,$$

and then

$$\frac{\varrho_i}{\prod_{j=1}^i c_j + \tau} \leq \frac{\varrho_i}{\prod_{j=1}^i |b_j| + \tau} \leq K \left( \frac{\varrho}{e^{E \log(|b_1| + \tau)}} \right)^i \leq K \text{ a.s.}$$

We thus have

$$\begin{aligned} \left\| f'_{t+1} - \widehat{f}'_{t+1} \right\| &\leq K \sum_{i=1}^t \varrho_i \frac{\prod_{j=1}^t (c_j + \tau)}{\prod_{j=1}^i (c_j + \tau)} + K \prod_{j=1}^t (c_j + \tau) \\ &\leq K \prod_{j=1}^t (c_j + \tau) \left\{ 1 + \sum_{i=1}^t \varrho_i \right\}. \end{aligned}$$

Note also that $E \log(|b_1| + \widetilde{\tau}) < 0$ implies

$$(|b_1| + \widetilde{\tau}) \cdots (|b_t| + \widetilde{\tau}) \leq K \widetilde{\varrho}^t \quad \text{ a.s., when } e^{E \log(|b_1| + \widetilde{\tau})} < \widetilde{\varrho} < 1.$$

Since $\limsup_{t \to \infty} (\log \varrho_t)/t \leq \log \rho + \limsup_{t \to \infty} (\log u_t)/t < 0$, using $E \log^+ u_t < \infty$, it follows that $\varrho_t$ converges almost surely to 0 as $t \to \infty$. When $\tau < \widetilde{\tau}$ we then have $0 \leq c_t + \tau < |b_t| + \widetilde{\tau}$ for $t$ large enough, and thus

$$(c_1 + \tau) \cdots (c_t + \tau) \leq K \widetilde{\varrho}^t \quad \text{ a.s.}$$

For any $\varrho_* \in (\widetilde{\varrho}, 1)$ we then have

$$\frac{1}{\varrho_*^t} \left\| f'_{t+1} - \widehat{f}'_{t+1} \right\| \leq K \left( \frac{\widetilde{\varrho}}{\varrho_*} \right)^t \left( 1 + \sum_{i=1}^\infty \varrho_i \right) \to 0$$

a.s. as $t \to \infty$.

The second-order derivatives are treated in the same way, and the conclusion follows.

## Proof of Theorem 1

By compactness of $\Theta$, the strong consistency is obtained by showing that for any $\theta \neq \theta_0$, there exists a neighbourhood $V(\theta)$ of $\theta$ such that

$$\liminf_{T\to\infty} \inf_{\theta^*\in V(\theta)\cap\Theta} \left\|\widehat{G}_T(\theta^*)\right\| > 0, \quad \text{a.s.} \tag{44}$$

and that for any neighbourhood $V(\theta_0)$ of $\theta_0$

$$\limsup_{T\to\infty} \inf_{\theta^*\in V(\theta_0)\cap\Theta} \left\|\widehat{G}_T(\theta^*)\right\| = 0, \quad \text{a.s.} \tag{45}$$

Let

$$G_T(\theta) = \frac{1}{T}\sum_{t=t_0+1}^{T} g_t(\theta).$$

For any neighbourhood $V(\theta)$ of $\theta$, we have

$$\inf_{\theta^*\in V(\theta)\cap\Theta} \left\|\widehat{G}_T(\theta^*)\right\| \geq \inf_{\theta^*\in V(\theta)\cap\Theta} \|G_T(\theta^*)\| - \sup_{\theta\in\Theta} \left\|G_T(\theta) - \widehat{G}_T(\theta)\right\|.$$

By (15), (16) and (17), we have

$$\sup_{\theta\in\Theta} |g_t(\theta) - \widehat{g}_t(\theta)| \leq K\varrho^t u_t, \quad u_t = \sup_{\theta\in\Theta} \left(|y_t|^k + |f_t(\theta)| + 1\right)\left(1 + \left\|\frac{\partial f_t(\theta)}{\partial\theta}\right\|\right).$$

Since $E\log^+ u_t < \infty$ under the log-moment conditions and $\varrho < 1$, the Cauchy root test shows that

$$\sum_{t=1}^{\infty} \sup_{\theta\in\Theta} |g_t(\theta) - \widehat{g}_t(\theta)| < \infty \ \text{ a.s.},$$

which entails that, almost surely, $\sup_{\theta\in\Theta} \left\|G_T(\theta) - \widehat{G}_T(\theta)\right\| \to 0$ as $T \to \infty$. Now note that

$$\inf_{\theta^*\in V(\theta)\cap\Theta} \|G_T(\theta^*)\| \geq \|G_T(\theta)\| - \sup_{\theta^*\in V(\theta)\cap\Theta} \|G_T(\theta^*) - G_T(\theta)\|,$$

with

$$\sup_{\theta^*\in V(\theta)\cap\Theta} \|G_T(\theta^*) - G_T(\theta)\| \leq \frac{1}{T}\sum_{t=t_0+1}^{T} \sup_{\theta^*\in V(\theta)\cap\Theta} \|g_t(\theta^*) - g_t(\theta)\|.$$

Let $V_m(\theta)$ be the ball of center $\theta$ and radius $1/m$. By the ergodic theorem applied to $\left\{\sup_{\theta^*\in V_m(\theta)\cap\Theta} \|g_t(\theta^*) - g_t(\theta)\|\right\}_t$, we have

$$\limsup_{T\to\infty} \sup_{\theta^*\in V_m(\theta)\cap\Theta} \|G_T(\theta^*) - G_T(\theta)\| \leq E \sup_{\theta^*\in V_m(\theta)\cap\Theta} \|g_t(\theta^*) - g_t(\theta)\|.$$

By Fatou's lemma, the continuity of $g_t(\cdot)$ and (20), the expectation of the right-hand side of the inequality tends to 0 as $m \to \infty$. By (21) and the ergodic theorem, we have

$$\lim_{T\to\infty} \|G_T(\theta)\| = \|G(\theta)\| > 0$$

when $\theta \neq \theta_0$. We thus have shown (44).

To show (45), it suffices to use the same arguments, noting that

$$\limsup_{T\to\infty} \inf_{\theta^* \in V(\theta_0)\cap\Theta} \left\|\widehat{G}_T(\theta^*)\right\| \leq \lim_{T\to\infty} \left\|\widehat{G}_T(\theta_0)\right\| = \|G(\theta_0)\| = 0.$$

The proof of the consistency is complete.

By already given arguments, Lemma 4 and (23) show that, almost surely,

$$\sup_{\theta\in\Theta} \left\| \frac{\partial G_T(\theta)}{\partial \theta} - \frac{\partial \widehat{G}_T(\theta)}{\partial \theta} \right\| = O(T^{-1}) \quad \text{a.s.} \tag{46}$$

Now note that the ergodic theorem and $E_{t-1}h_t(\theta_0) = 0$ imply that

$$\dot{G}_T := \partial G_T(\theta_0)/\partial \theta^\top \to -\mathcal{J}$$

almost surely as $T \to \infty$. In view (24), we can thus assume that $\dot{G}_T$ is invertible. The mapping $f_T : \Theta \to \Theta$ then defined by

$$f_T(\theta) = \theta - \dot{G}_T^{-1}\widehat{G}_T(\theta)$$

satisfies

$$\left\| \frac{\partial f_T(\theta)}{\partial \theta} \right\| \leq \left\|\dot{G}_T^{-1}\right\| \left\| \dot{G}_T - \frac{\partial \widehat{G}_T(\theta)}{\partial \theta^\top} \right\| < 1$$

for $T$ large enough on some neighborhood of $\theta_0$, using (46), the ergodic theorem and the continuity of $\partial G(\theta)/\partial\theta^\top$. The contraction $f_T$ thus admits a unique fixed-point $\theta_T$ on this neightborhood, for which $\widehat{G}_T(\theta_T) = 0$. See Jacod and Sørensen (2017) and the references therein for examples of applications of the fixed-point theorem to show the asymptotic existence of an estimator. In view of (19), we have $\theta_T = \widehat{\theta}_T$, and thus

$$\widehat{G}_T(\widehat{\theta}_T) = 0.$$

The rest of the proof follows by Taylor expansions, using standard arguments.

## Proof of Theorem 2

The desired result follows from the classical consistency argument found e.g. in White (1994, Theorem 3.4) or Potscher and Prucha (1997, Lemma 3.1). First we

show that the sample log-likelihood converges uniformly to a deterministic limit criterion. Next we show that $\theta_0^*$ is the identifiably unique maximizer of the limit criterion.

The uniform convergence of the criterion follows from

$$\begin{aligned}
&\sup_{\theta\in\Theta}\Big|\frac{1}{T}\sum_{t=2}^{T}\ell(y_t,\widehat{f}_t(\theta),\theta)-\mathbb{E}\ell(y_t,f_t(\theta),\theta)\Big| \\
&\quad\le \frac{1}{T}\sum_{t=2}^{T}\sup_{\theta\in\Theta}\Big|\ell(y_t,\widehat{f}_t(\theta),\theta)-\ell(y_t,f_t(\theta),\theta)\Big| \\
&\qquad + \sup_{\theta\in\Theta}\Big|\frac{1}{T}\sum_{t=2}^{T}\ell(y_t,f_t(\theta),\theta)-\mathbb{E}\ell(y_t,f_t(\theta),\theta)\Big| \\
&\quad\le \frac{1}{T}\sum_{t=2}^{T}\sup_{\theta\in\Theta}\sup_{f}\Big|\frac{\partial\ell(y_t,f,\theta)}{\partial f}\Big|\sup_{\theta\in\Theta}|\widehat{f}_t(\theta)-f_t(\theta)| \\
&\qquad + \sup_{\theta\in\Theta}\Big|\frac{1}{T}\sum_{t=2}^{T}\ell(y_t,f_t(\theta),\theta)-\mathbb{E}\ell(y_t,f_t(\theta),\theta)\Big|,
\end{aligned}$$

$$\text{where}\quad \frac{1}{T}\sum_{t=2}^{T}\sup_{\theta\in\Theta}\sup_{f}\Big|\frac{\partial\ell(y_t,f,\theta)}{\partial f}\Big|\sup_{\theta\in\Theta}|\widehat{f}_t(\theta)-f_t(\theta)| \xrightarrow{as} 0 \quad\text{as}\quad T\to\infty$$

by the uniform invertibility obtained in Lemma 3,

$$\text{and}\quad \sup_{\theta\in\Theta}\Big|\frac{1}{T}\sum_{t=2}^{T}\ell(y_t,f_t(\theta),\theta)-\mathbb{E}\ell(y_t,f_t(\theta),\theta)\Big| \xrightarrow{as} 0 \quad\text{as}\quad T\to\infty$$

by application of Rao's (1962) uniform law of large numbers. The identifiable uniqueness of $\theta_0^*\in\Theta$ is implied by the uniqueness assumption $\mathbb{E}\ell(y_t,f_t,\theta)<\mathbb{E}\ell(y_t,f_t,\theta_0^*)$ for every $\theta\neq\theta_0^*$, $\theta\in\Theta$, the continuity of the limit criterion and the compactness of $\Theta$ (Potscher and Prucha, 1997). The interpretation of $\theta_0^*$ as the minimizer of the expected KL is well known and available e.g. in White (1994).

## Proof of Lemma 5

Immediate under the assumptions of Theorem 2 as long as the level sets of the limit log-likelihood function are regular. In our case, the regularity of the level sets is easily implied by continuity (see Lemma 4.2 in Postcher and Prucha, 1997).

## Proof of Corollary 1

The proof is the same as for Theorem 2 after showing that the data $\{y_t\}_{t\in\mathbb{Z}}$ is strictly stationary and ergodic. This follows by application of Lemma 1 at $\theta_0 \in \Theta$ and by continuity of $y_t$ in $f_t$ and $\epsilon_t$.

## Proof of Lemma 6

The first claim of Lemma 6 is obtained by noting that Conditions *(i)* and *(ii)* imply

$$|f_{t+1} - f^*_{t+1}| \leq a|y_t - y^*_t| + b|f_t - f^*_t|$$

$$\text{with} \quad a := |\alpha| \sup_{f,y} \Big|\frac{\partial\psi(y,f,\theta)}{\partial y}\Big| < \infty \quad \text{and} \quad b := \sup_{f,y} \Big|\alpha\frac{\partial\psi(y,f,\theta)}{\partial f} + \beta\Big| < 1.$$

Since $\{y_t\}$ is NED of size $-q$ on some process $\{e_t\}_{t\in\mathbb{Z}}$ and has two bounded moments $\sup_t \mathbb{E}|y_t|^2 < \infty$, we conclude by Theorem 6.10 of Potscher and Prucha (1997) that $\{\widehat{f}_t\}$ is also NED of size $-q$ on $\{e_t\}_{t\in\mathbb{Z}}$.

The second claim is obtained by noting that the filter $\widehat{f}_t(\theta_0)$ and its derivative are both uniformly bounded $\sup_t |f'_t(\theta_0)| < M < \infty$,

$$\begin{aligned}\sup_t \|\widehat{f}'_{t+1}(\theta_0)\| &\leq \sup_t \sum_{j=0}^{t-1} |B(y_t,\widehat{f}_t,\theta_0)|^j \|A(y_t,\widehat{f}_t,\theta_0)\| + \sup_t |B_1(y_t,\widehat{f}_t,\theta_0)|^t |\widehat{f}'_1(\theta_0)| \\ &\leq (1 - \sup_{y,f}|B(y,f,\theta_0)|)^{-1} \sup_{y,f}\|A(y,f,\theta_0)\| \\ &\quad + \sup_{y,f}|B_1(y,f,\theta_0)||\widehat{f}'_1(\theta_0)| \;\leq\; M \;<\; \infty\end{aligned}$$

$$\text{because} \quad A(y_t,\widehat{f}_t,\theta_0) \;=\; \frac{\partial\omega_0}{\partial\theta} + \frac{\partial\alpha_0}{\partial\theta}\psi(y_t,\widehat{f}_t,\theta_0) + \alpha\frac{\partial\psi(y_t,\widehat{f}_t,\theta_0)}{\partial\theta} + \frac{\partial\beta_0}{\partial\theta}\widehat{f}_t \;,$$

$$\text{and} \quad B(y_t,\widehat{f}_t,\theta_0) \;=\; \alpha_0\frac{\partial\psi(y_t,\widehat{f}_t,\theta_0)}{\partial f} + \beta_0$$

with $\sup_{y,f}\|A(y_t,\widehat{f}_t,\theta_0)\| < \infty$ and $\sup_{y,f}|B(y_t,\widehat{f}_t,\theta_0)| < 1$. Next, we verify that the derivative filter satisfies

$$\|\widehat{f}'_{t+1} - \widehat{f}'^*_{t+1}\| \leq a_y|y_t - y_t| + a_f|\widehat{f}_t - \widehat{f}^*_t| + b\|\widehat{f}'_t - \widehat{f}'^*_t\|,$$

$$\begin{aligned}\text{where} \quad a_y :=& \Big\|\frac{\partial\alpha}{\partial\theta}\Big\| \sup_{y,f}\Big|\frac{\partial\psi(y,f,\theta)}{\partial y}\Big| + |\alpha| \sup_{y,f}\Big\|\frac{\partial^2\psi(y,f,\theta)}{\partial\theta\partial y}\Big\| \\ &+ |\alpha| \sup_{y,f}\Big|\frac{\partial^2\psi(y,f,\theta)}{\partial f\partial y}\Big| M < \infty,\end{aligned}$$

$$a_f := \Big\|\frac{\partial\alpha}{\partial\theta}\Big\| \sup_{y,f}\Big|\frac{\partial\psi(y,f,\theta)}{\partial f}\Big| + |\alpha| \sup_{y,f}\Big\|\frac{\partial^2\psi(y,f,\theta)}{\partial\theta\partial f}\Big\| + \Big\|\frac{\partial\beta}{\partial\theta}\Big\|$$
$$+ |\alpha| \sup_{y,f}\Big|\frac{\partial^2\psi(y,f,\theta)}{\partial f^2}\Big| M < \infty,$$
$$\text{and} \quad |b| := \sup_{y,f}\Big|\alpha\frac{\partial\psi(y,f,\theta)}{\partial f} + \beta\Big| < 1.$$

Since $\{(y_t, \widehat{f}_t)\}$ id NED of size $-q$ on some process $\{e_t\}_{t\in\mathbb{Z}}$ with $\sup_t \mathbb{E}|y_t|^2 < \infty$ and $\sup_t |\widehat{f}_t| < \infty$, we conclude again by Theorem 6.10 of Potscher and Prucha (1997) that $\{\widehat{f}'_t\}$ is also NED of size $-q$ on $\{e_t\}_{t\in\mathbb{Z}}$.

Finally, we conclude that the score $\{\widehat{\ell}'_t(\theta_0)\}_{t\in\mathbb{N}}$ is also NED of size $-q$ on $\{e_t\}_{t\in\mathbb{Z}}$ by the Lipschitz assumption and Theorem 6.7 and Corollary 6.8 of Potscher and Prucha (1997).

## Proof of Theorem 3

For convenience, we adopt the following notation

$$\boldsymbol{\ell}_T(\theta) := \frac{1}{T}\sum_{t=2}^{T} \ell(y_t, f_t(\theta), \theta)$$

and furthermore, we let $\widehat{\boldsymbol{\ell}}'_T(\theta) := \partial\widehat{\boldsymbol{\ell}}_T(\theta)/\partial\theta$, $\boldsymbol{\ell}'_T(\theta) := \partial\boldsymbol{\ell}_T(\theta)/\partial\theta$ and $\boldsymbol{\ell}''_T(\theta) := \partial\boldsymbol{\ell}_T(\theta)/(\partial\theta\partial\theta')$.

Below, we first obtain the asymptotic normality of the estimator $\widetilde{\theta}_T$ which maximizes the criterion $\ell_T$, i.e.,

$$\widetilde{\theta}_T \in \arg\max_{\theta\in\Theta} \boldsymbol{\ell}_T(\theta),$$

and also show that $\widehat{\theta}_T$ has the same asymptotic distribution as $\widetilde{\theta}_T$.

We use the usual mean-value theorem expansion

$$\boldsymbol{\ell}'_T(\widetilde{\theta}) - \boldsymbol{\ell}'_T(\theta_0^*) = \boldsymbol{\ell}''_T(\theta_T^*)(\widetilde{\theta}_T - \theta_0^*),$$

to obtain

$$\sqrt{T}(\widetilde{\theta}_T - \theta_0^*) = -\Big(\boldsymbol{\ell}''_T(\theta_T^*)\Big)^{-1}\sqrt{T}\boldsymbol{\ell}'_T(\theta_0^*). \tag{47}$$

By Lemma 6, we have that the score sequence $\{\ell'_t(\theta_0^*)\}_{t\in\mathbb{Z}}$ is near epoch dependent of size $-1$ on a $\phi$-mixing sequence of size $-r/(r-1)$ for some $r > 2$. Given the moment bounds $\mathbb{E}|\ell'(y_t, f_t, \theta_0)|^2 < \infty$, we can thus appeal to the central limit theorem for near epoch dependent sequences in Potscher and Prucha (1997, Theorem 10.2) to show that

$$\sqrt{T}\boldsymbol{\ell}'_T(\theta_0^*) \xrightarrow{d} N(0, V(\theta_0^*)) \quad \text{as } T \to \infty. \tag{48}$$

Additionally, by the stationary and ergodic behavior of the limit filter and its derivatives obtained in Lemma 4 and the uniform moment bound on the Hessian,

$$\mathbb{E}\sup_{\theta\in\Theta}|\ell''(y_t, f_t, \theta)| < \infty.$$

The uniform convergence of the Hessian over $\Theta$ is obtained by Rao's (1962) uniform law of large numbers (i.e., $\sup_{\theta\in\Theta}\|\boldsymbol{\ell}''_T(\theta) - \mathbb{E}\ell''_t(\theta)\| \xrightarrow{as} 0$, which implies

$$\boldsymbol{\ell}''_T(\theta^*_T) = \frac{1}{T}\sum_{t=2}^{T}\ell''_t(\theta^*_T) \xrightarrow{as} \mathbb{E}\ell''_t(\theta^*_0) \quad \text{as} \quad T\to\infty, \tag{49}$$

since $\theta^*_T \xrightarrow{as} \theta^*_0$. The asymptotic distribution of $\widetilde{\theta}_T$ is obtained by combining (47), (48) and (49), i.e.,

$$\sqrt{T}(\widetilde{\theta}_T - \theta^*_0) \xrightarrow{d} N\big(0, \Sigma(\theta^*_0)\big),$$

where the asymptotic variance is given by

$$\Sigma(\theta^*_0) = \Big(\mathbb{E}\widehat{\ell}''_t(\theta^*_0)\Big)^{-1}\Big(\mathbb{E}\widehat{\ell}'_t(\theta_0)\mathbb{E}\widehat{\ell}'_t(\theta_0)^\top\Big)\Big(\mathbb{E}\widehat{\ell}''_t(\theta^*_0)\Big)^{-1}.$$

We now expand the score using a mean value theorem

$$\boldsymbol{\ell}'_T(\widetilde{\theta}_T) - \boldsymbol{\ell}'_T(\widehat{\theta}_T) = \boldsymbol{\ell}''_T(\theta^*_T)(\widetilde{\theta}_T - \widehat{\theta}_T)$$

and notice that $\boldsymbol{\ell}'_T(\widetilde{\theta}_T) = \widetilde{\boldsymbol{\ell}}'_T(\widehat{\theta}_T) = 0$ to obtain

$$\sqrt{T}\Big(\widetilde{\boldsymbol{\ell}}'_T(\widehat{\theta}_T) - \boldsymbol{\ell}'_T(\widehat{\theta}_T)\Big) = \boldsymbol{\ell}''_T(\theta^*_T)\sqrt{T}(\widetilde{\theta}_T - \widehat{\theta}_T). \tag{50}$$

We use again the uniform convergence of the Hessian to conclude that

$$\boldsymbol{\ell}''_T(\theta^*_T) \xrightarrow{as} \mathbb{E}\ell''_t(\theta^*_0). \tag{51}$$

Finally, we use the uniform bounded moment on the score $\mathbb{E}\sup_{\theta\in\Theta}|\ell'(y_t, f_t, \theta)| < \infty$ and Lemma 4 to obtain,

$$\sqrt{T}\sup_{\theta\in\Theta}\big|\widetilde{\boldsymbol{\ell}}'_T(\theta) - \boldsymbol{\ell}'_T(\theta)\big| \xrightarrow{as} 0 \quad \text{as } T\to\infty$$

which in turn implies that

$$\sqrt{T}\big|\widetilde{\boldsymbol{\ell}}'_T(\widehat{\theta}_T) - \boldsymbol{\ell}'_T(\widehat{\theta}_T)\big| \xrightarrow{as} 0 \quad \text{as } T\to\infty. \tag{52}$$

Combining (50), (51) and (52), we conclude that $\sqrt{T}|\widetilde{\theta}_T - \widehat{\theta}_T| \xrightarrow{as} 0$ as $T\to\infty$. This delivers the desired result

$$\sqrt{T}(\widehat{\theta}_T - \theta^*_0) \xrightarrow{d} N\big(0, \Sigma(\theta^*_0)\big).$$

## Proof of Corollary 2

The proof is the same as for Theorem 3 with the exception that the score satisfies a central limit theorem for martingale difference sequences at $\theta_0$ and hence does not need the NED property. Additionally, the stationarity the data $\{y_t\}_{t\in\mathbb{Z}}$ follows by application of Lemma 1 at $\theta_0 \in \Theta$ and by continuity of $y_t$ in $f_t$ and $\epsilon_t$.

## Proof of Theorem 4

Recall that the constrained estimator $(\widehat{\theta}_T^r)$ is such that $(\widehat{\theta}_T^r, \widehat{\lambda}_T)$ is a critical point of the Lagrangian function

$$\mathcal{L}(\theta, \lambda) = \widehat{\boldsymbol{\ell}}_T(\theta) - \lambda^\top (R\theta - \boldsymbol{r}).$$

The first order conditions yield

$$R\widehat{\theta}_T^r - \boldsymbol{r} = 0, \qquad R^\top \widehat{\lambda}_T = \frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta}. \tag{53}$$

First recall that from Corollary 2

$$R\sqrt{T}(\widehat{\theta}_T - \theta_0) \xrightarrow{d} N\big(0, R\mathcal{I}^{-1}R^\top\big), \tag{54}$$

where $\mathcal{I} = -\mathbb{E}\ell_t''(\theta_0)$.

We know that $\widehat{\theta}_T \to \theta_0$ a.s., and it can be shown that $\widehat{\theta}_T^r \to \theta_0$ a.s. under $H_0$. A Taylor expansion then entails

$$\sqrt{T}\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta} + o_P(1) = \sqrt{T}\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T)}{\partial \theta} - \mathcal{I}\sqrt{T}(\widehat{\theta}_T^r - \widehat{\theta}_T) = -\mathcal{I}\sqrt{T}(\widehat{\theta}_T^r - \widehat{\theta}_T). \tag{55}$$

Using (53), it follows that under $H_0$

$$R\sqrt{T}(\widehat{\theta}_T - \theta_0) \;\; = \;\; R\sqrt{T}(\widehat{\theta}_T - \widehat{\theta}_T^r) = R\mathcal{I}^{-1}R^\top\sqrt{T}\widehat{\lambda}_T + o_P(1). \tag{56}$$

Using (54) we then obtain

$$\sqrt{T}\widehat{\lambda}_T = \left(R\mathcal{I}^{-1}R^\top\right)^{-1} R\sqrt{T}(\widehat{\theta}_T - \theta_0) + o_P(1) \xrightarrow{d} N\left\{0, \left(R\mathcal{I}^{-1}R^\top\right)^{-1}\right\}$$

and thus, using again (53),

$$T\widehat{\lambda}_T^\top R\mathcal{I}^{-1}R^\top\widehat{\lambda}_T = T\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta^\top}\mathcal{I}^{-1}\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta} \xrightarrow{d} \chi_r^2. \tag{57}$$

The first convergence follows.

To derive the asymptotic distribution of $\mathrm{LR}_T$ we use the usual argument which involves expanding $\widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T)$ around $\widehat{\theta}_T^r$ to obtain

$$\begin{aligned}
\mathrm{LR}_T &:= 2T\Big(\widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T) - \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)\Big) \\
&= 2T\Big(\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta^\top}(\widehat{\theta}_T - \widehat{\theta}_T^r) - \frac{1}{2}(\widehat{\theta}_T - \widehat{\theta}_T^r)^\top \mathcal{I}(\widehat{\theta}_T - \widehat{\theta}_T^r)\Big) + o_P(1) \\
&= \sqrt{T}(\widehat{\theta}_T - \widehat{\theta}_T^r)^\top \sqrt{T}\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)}{\partial \theta} + o_P(1) \\
&= \sqrt{T}(\widehat{\theta}_T - \widehat{\theta}_T^r)^\top \sqrt{T} R^\top \widehat{\lambda}_T + o_P(1) \\
&= T\widehat{\lambda}_T^\top R \mathcal{I}^{-1} R^\top \widehat{\lambda}_T + o_P(1) \xrightarrow{d} \chi^2_q
\end{aligned}$$

noting $\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^r)/\partial\theta = 0$ and using (53), (55), (56) and (57).