Quasi score-driven models*

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

April 10, 2021

Abstract

This paper introduces the class of quasi score-driven (QSD) models. This new class inherits and extends the basic ideas behind the development of score-driven (SD) models and addresses a number of unsolved issues in the score literature. In particular, the new class of models (i) generalizes many existing models, including SD models, (ii) disconnects the updating equation from the log-likelihood implied by the conditional density of the observations, (iii) allows testing of the assumptions behind SD models that link the updating equation of the conditional moment to the conditional density, (iv) allows QML estimation of SD models, (v) and allows explanatory variables to enter the updating equation.

We establish the asymptotic properties of the QLE, QMLE and MLE of the proposed QSD model as well as the 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 QSD models to estimate the conditional variance of 400 US stocks.

*The authors gratefully acknowledge Kris Boudt, Paul Embrechts, Patrick Gagliardini, Peter Reinhard Hansen, Andrew Harvey, Elvezio Ronchetti, Bilel Sanhaji and Olivier Scaillet for helpful discussions as well as the participants of the GFRI seminar in Geneva, the Econometrics seminar at KULeuven, the Quantact seminar in Montreal, the Econometrics and Business Statistics seminar at CREATES in Aarhus and the 22nd Dynamic Econometrics Conference in Oxford.

†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

Score-driven (SD) models, also known as generalized autoregressive score (GAS) models or dynamic conditional score (DCS) models, have been proposed independently by Harvey and Chakravarty (2008) and Creal, Koopman and Lucas (2012). They 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 $\{f_t\}_{t\in\mathbb{Z}}$,

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

The conditional distribution $p(y_t|f_t,\theta)$ is often defined by means of a time series model of the type

$$y_t = g(f_t,\epsilon_t), \tag{1}$$

where ϵ_t is an $i.i.d.$ random variable that can be interpreted as an error term and g is a differentiable measurable function. $g(f_t,\epsilon_t) := f_t + \epsilon_t$ describes a location model while $g(f_t,\epsilon_t) := \sqrt{f_t}\epsilon_t$ corresponds to a volatility or duration model.

An SD model for f_t is a model of the form

$$f_{t+1} = \omega + \alpha S(f_t)\frac{\partial \log p(y_t|f_t,\theta)}{\partial f_t} + \beta f_t, \tag{2}$$

where $S(f_t)$ is a scaling function for the score, e.g., the inverse of an information matrix.

There are more than 220 papers referenced on the `gasmodel.com` website that build upon this modelling strategy and have applications in various areas in financial econometrics (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) but also in macroeconomics or public health.

The success of SD models is attributable to the following features: (i) these models nest and extend existing observation-driven models such as $GARCH$, (ii) their estimation does not require sophisticated techniques (the conditional likelihood is readily available), (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' predictive performance is outstanding.

They have 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), $\frac{\partial \log p_t(y_t,\theta)}{\partial f_t}$ downweights and even bounds the effect of large shocks on the conditional variance; by 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, which

suggests that $GARCH$ models tend to 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 (2) that SD models impose a strong link between the conditional distribution of y_t, i.e., $p(y_t|f_t, \theta)$, 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 (2) of SD models but allow the updating equation of f_t to be disconnected from the density of the innovations if needed. Our family of models, called quasi score-driven (or QSD) models, therefore encompasses not only SD models but also many other existing models. For instance, Banulescu-Radu, Hansen, Huang and Matei (2018) use a volatility model whose dynamic is derived from a QSD model obtained from a Student's t log-likelihood to bound the effect of large shocks, although the model is estimated by Gaussian ML.

We study the statistical properties of QSD models in the case where the conditional moment of interest depends on some covariates. We also study three estimation methods for this model. Since QSD models disconnect the dynamics in f_t and the density of the innovations, unlike SD models, they permit the consideration of the estimation of the parameters by QML or by ML with a score function in (2) taken with respect to a conditional density other than the one of y_t. We also study the estimation of QSD models using the quasi-likelihood estimator (QLE), which encompasses the QMLE. QLE can actually be seen as an extension of QMLE where the mean-variance relationship of the exponential family is relaxed.

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 χ^2 distribution, which offers a strategy to test some restrictions implied by standard SD models.

We study in detail a QSD volatility model extending the $Beta_T\ GARCH$ model of Harvey and Chakravarty (2008). This model, called $QSD_T\ GARCH-T$, relies on a Student's t density for the innovations and the score of a Student's t log-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 $Beta_T\ GARCH$) is found to be significant at the 5% significance level in more than 50% of the cases out of 400 US stocks. We also show that when an asymmetric density is used instead of a symmetric Student's t density, the SD model is rejected in favour of its QSD extension in all cases.

The rest of the paper is structured as follows. Section 2 presents the quasi score 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 QSD_T $GARCH-T$ model, i.e., the stationarity, invertibility and positivity conditions, its estimation, some hypothesis tests as well as the finite sample properties of the QMLE and MLE via a Monte Carlo simulation. The empirical application is presented in Section 5. Finally, Section 6 concludes. All proofs are given in the appendix.

2 The QSD filtering equation

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

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

where ω, α, β are real parameters, θ is an element of a parameter space $\Theta \subset \mathbb{R}^p$, ψ is a differentiable measurable function and X_t is a vector of exogenous random variables.

2.1 Some examples

The class of models defined by (1) and (3) is very general. For instance, when $p(y_t|f_t, \theta)$ is a Student's t density with conditional variance f_t and $\psi(y_t, X_t, f_t, \theta) := y_t^2$, we obtain the $GARCH-T$ model of Bollerslev (1987) and a $GARCH-T$ model with additive explanatory variables X_t in the conditional variance when $\psi(y_t, X_t, f_t, \theta) := y_t^2 + \varpi^\top X_t$, where ϖ is a parameter vector.

While the QSD filtering equation (3) is very general, the parallel with SD models is clearer when specifying ψ as follows

$$\psi(y_t, X_t, f_t, \theta) := \frac{\partial \rho(y_t, X_t, f_t, \theta)}{\partial f_t} S(f_t). \tag{4}$$

Interestingly, when $\rho(y_t, X_t, f_t, \theta)$ is the log-likelihood function of y_t, i.e., $\rho(y_t, X_t, f_t, \theta) = \log p(y_t|f_t, \theta)$, we recover the class of score models described in (2).[1] Importantly, when $\rho(y_t, f_t, \theta)$ is a log-likelihood function but $\rho(y_t, f_t, \theta) \neq \log p(y_t|f_t, \theta)$, ψ is no longer proportional to $\frac{\partial \log p_t(y_t,\theta)}{\partial f_t}$ but is proportional to $\frac{\partial \rho(y_t,f_t,\theta)}{\partial f_t}$. Consequently, the model is called the quasi score-driven (QSD) model.

The $GARCH-T$ model presented above is an example of a QSD model, where $p(y_t|f_t, \theta)$ is a Student's t density with conditional variance f_t but $\rho(y_t, f_t, \theta)$ is a Gaussian log-likelihood with conditional variance f_t while $S(f_t) = -\mathbb{E}\left[\frac{\partial^2 \rho(y_t,f_t,\theta)}{\partial f_t \partial f_t^\top}\right]^{-1} = 2f_t^2$, i.e., the inverse of the information matrix of ρ.

[1] Notice that when exogenous variables X_t are not present in ψ and ρ, we simply write $\psi(y_t, f_t, \theta)$ and $\rho(y_t, f_t, \theta)$ instead of $\psi(y_t, X_t, f_t, \theta)$ and $\rho(y_t, X_t, f_t, \theta)$.

In this formulation, QMLE also becomes a naturally viable alternative to MLE. Indeed, in a volatility model, large shocks can be downweighted by an appropriate choice of ρ, e.g., a Student's t log-likelihood, without imposing y_t to follow the same conditional distribution; i.e, y_t can be assumed to be conditionally Gaussian as in Banulescu-Radu, Hansen, Huang and Matei (2018). This is in contrast to SD models where there is a strong link between the innovation density and the updating equation, which makes it unnatural to use QMLE.

We also note that, in our new model formulation, we can formulate updating equations that Winsorize or censor outliers, regardless of the conditional distribution $p(y_t|f_t, \theta)$. More generally, QSD models yield filtering equations that employ many popular loss functions used in 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" SD models that are unable to deliver an updating equation with a leverage effect when the innovation density is left-skewed (as shown in Example 2 below). 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 and 2 cover cases of location and volatility filtering involving non-linear asymmetric criteria as well as fat-tailed and skewed innovations.

EXAMPLE 1. (Asymmetric forecast with symmetric innovations) *Consider a location model where $y_t = f_t + \epsilon_t$ and ϵ_t is i.i.d.. When ϵ_t is also assumed to follow a $N(0, \sigma^2)$ distribution, the score is symmetric in ϵ_t, and the resulting SD model is given by*

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

In many applications it is desirable to weight differently positive and negative shocks, irrespectively of the shape of the distribution of ϵ_t.

As an example, consider the negative linex loss function introduced by Varian (1975), i.e.,

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

The QSD filtering equation obtained with this loss function is

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

and can therefore provide asymmetric forecasts even when the density of the innovations is symmetric.

EXAMPLE 2. (Volatility model with leverage effect and left-skewed innovations) *Stock returns are typically heavy tailed and left-skewed (see Giot and Laurent, 2003 among others). As such, SD models of the conditional volatility* $y_t = \sqrt{f_t}\epsilon_t$ *employing skewed distributions such as the skewed Gaussian or skewed 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, which predicts higher volatility after negative returns. Depending on the skewed Student's t distribution that is adopted, pure score models may thus be unable to capture the leverage effect. This issue does not affect the larger class of* QSD *models*

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

since the ψ *function can adopt nonlinear functional forms independently of the density of the innovations* ϵ_t*. This model will be discussed in more detail in Section 5.2.*

2.2 Stationarity and invertibility of QSD models

We now give general conditions for stationarity and invertibility of the QSD model defined by (1) and (3). These general conditions will be illustrated on the QSD_T $GARCH-T$ model presented in Section 4. In an earlier version of this paper (Blasques et al., 2020), the assumptions are also made more explicit on various specific examples.

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. We assume that E is a convex subspace of $\mathbb{R}^d$ and F is an interval; see Bougerol (1993) and Straumann and Mikosch (2006) for major references on SRE theory.

The results given in this section can be considered a direct application of the general theory developed in Bougerol (1993), but we will provide explicit and self-contained proofs. Our results are also closely in line with those of Straumann and

Mikosch (2006), which focus on volatility models, or those reported for SD models in Blasques et al. (2020), which do not include exogenous variables.

Lemma 1 details conditions for the QSD 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) $\mathbb{E}\log\Lambda_t<0$ *with* $\Lambda_t=\sup_f\left|\alpha\frac{\partial\psi(g(f,\epsilon_t),X_t,f,\theta)}{\partial f}+\beta\right|<0$.

Then, there exist unique strictly stationary and ergodic solutions $\{f_t\}_{t\in\mathbb{Z}}$ and $\{y_t\}_{t\in\mathbb{Z}}$ to Equations (1)-(3).

Lemma 2 gives conditions for the existence of bounded unconditional moments.

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

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

for some $r>0$, then the stationary solution to Equations (1)-(3) satisfies $\mathbb{E}|f_t|^s<\infty$ for some $s>0$.

Note that if the exogenous variables appear linearly, i.e., if $\psi(g(f,\epsilon_t),X_t,f,\theta)=\psi(g(f,\epsilon_t),f,\theta)+\varpi^\top X_t$, then Λ_t only depends on ϵ_t, and therefore the assumption that (Λ_t) is *i.i.d.* is satisfied without having to assume X_t to be *i.i.d.* itself.

Assume that, for some $\theta=\theta_0$ satisfying the assumptions of Lemma 1, (y_t) is the stationary solution to (1)-(3), and recall that the time-varying parameter f_t depends on the true but unknown parameter θ_0. For all θ, 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 when $\beta\neq0$, $f_t(\theta)$ is not computable 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

$$\widehat{f}_{t+1}(\theta)=\omega+\alpha\psi(y_t,X_t,\widehat{f}_t(\theta),\theta)+\beta\widehat{f}_t(\theta),\quad t\geq1,\tag{7}$$

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 Θ to F.

Lemma 3 gives sufficient conditions for the invertibility of the QSD 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*

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

When (8) holds, the model is said to be uniformly invertible. This property will be essential to find a consistent estimator of θ_0 and to approximate the time-varying parameter f_t.

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 Θ is a compact subspace of $\mathbb{R}^p$ with $p \geq 3$. Without loss of generality, assume that $\theta = (\theta_1, \ldots, \theta_p)^\top$

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 $\widehat{f}_1'(\theta) \in \mathbb{C}(\Theta, \mathbb{R}^p)$. $\widehat{A}_t$ 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 ψ 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

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

3 Estimating the QSD models

In contrast to SD models, QSD models disentangle the parameters involved in f_t of those involved in the conditional distribution. We first consider the case where the time-varying parameter of interest is $f_t = f_t(\theta_0)$, where θ_0 is a p-dimensional vector. It makes sense to estimate θ_0, trying to be as agnostic as possible on the distribution of the innovations. In Section 3.1, we consider the class of the so-called quasi-likelihood estimators (QLEs), which encompasses the usual QMLEs. Section 3.2 will consider the MLE, which is the natural estimator for SD and QSD models, for which the conditional distribution is entirely specified by θ_0. In that case, θ_0 contains the parameters governing the dynamic of f_t and possibly the shape parameters of the conditional distribution $p(\cdot \mid f, \theta_0)$.

3.1 The QLE approach

Assume that (y_t) is a stationary process satisfying the QSD models (1)-(3) with $f_t = f_t(\theta_0)$.

To estimate θ_0 using very weak assumptions, the estimating functions theory can be used. This is a general estimation method 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 an optimal estimating function that applies in finite $i.i.d.$ samples, as well as for stochastic processes.

In time series models, there generally exists an "unbiased estimating function" $h_t = h_t(\theta_0) \in \mathbb{R}$, depending on y_t and $f_t = f_t(\theta_0)$, such that

$$\mathbb{E}_{t-1}(h_t) = 0,$$

where $\mathbb{E}_{t-1}$ 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 $\mathbb{E}(\epsilon_t) = 0$, 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 θ_0 can be obtained by solving an "estimating equation" of the form

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

where $a_t = a_t(\theta)$ is a p-dimensional vector of $\mathcal{F}_t$-measurable weights. Godambe (1985) shows that, within the class of the estimating functions of this form and

under mild assumptions, the optimal choice of a_t is

$$a_{t-1} = \frac{1}{\sigma_t^2(\theta)} \mathbb{E}_{t-1} \left(\frac{\partial f_t(\theta)}{\partial \theta} \right), \tag{14}$$

where $\sigma_t^2(\theta) = \mathbb{E}_{t-1} h_t^2(\theta)$ (possibly multiplied by an unimportant non-zero constant).

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

3.1.1 Conditional moment estimation

We now consider the case where $f_t = \mathbb{E}_{t-1}(y_t^k)$ for some $k > 0$. Location models correspond to $k = 1$ and volatility models to $k = 2$. We set $h_t(\theta) = y_t^k - f_t(\theta)$.

Of course, 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

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

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 an approximation 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 (ϵ_t) is *i.i.d.* with a mean of 0 and variance σ_ϵ^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 Θ 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

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

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

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

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

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

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

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

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:

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

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 (1)-(3) with θ replaced by θ_0 and $\mathbb{E}_{t-1}(y_t^k)=f_t(\theta_0)$ 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 $\mathbb{E}(\log^+|y_t|^k)<\infty$, $\mathbb{E}(\log^+\sup_{\theta\in\Theta}|f_t(\theta)|)<\infty$ and $\mathbb{E}(\log^+\sup_{\theta\in\Theta}\|\partial f_t(\theta)/\partial\theta\|)<\infty$. Suppose further (16), (17), (20) with $r=2$, (21), Θ is a compact subset of $\mathbb{R}^p$ and $\theta_0\in\Theta$. Then, for any sequence $\widehat{\theta}_T$ satisfying (19) and for T large enough, we have $\widehat{\theta}_T\overset{as}{\to}\theta_0$ as $T\to\infty$.*

Moreover, if θ_0 belongs to the interior of Θ, (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 Gourieroux, Monfort and Trognon (1984), it is known that in some location models of the form $y_t=f_t(\theta)+\epsilon_t$, the parameter θ can be estimated consistently by a*

quasi-maximum likelihood estimator (QMLE), which does not assume a particular distribution for ϵ_t but coincides with the MLE when the distribution of ϵ_t belongs to the linear exponential family, i.e., when, with respect to some σ-finite measure, ϵ_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 ϵ_t does not belong to that family, the QLE and QMLE are generally consistent, but the QLE may be more efficient.

The following examples compare the QLE and QMLE in volatility models.

EXAMPLE 3. (Standard volatility models and link with the Gaussian QMLE) *Consider the case where (1) is of the form $y_t = \sqrt{f_t}\epsilon_t$, where ϵ_t is i.i.d. with a mean of 0 and variance of 1. The usual QMLE of the volatility parameter θ_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 $\sigma_t^2 = f_t^2\,[\mathbb{E}(\epsilon_t^4) - 1]$, it is easy to see that in this case, the QMLE coincides with the optimal QLE.

EXAMPLE 4. (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 $\mathbb{E}_{t-1}(Z_t) = f_t$ 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 $\mathbb{E}_{t-1}(y_t^2) = f_t$ and $var_{t-1}(y_t^2) = \sigma_t^2$. When σ_t^2 is not proportional to f_t^2, the sequence (y_t) does not follow the standard volatility model of Example 3, and the QMLE is not the optimal QLE of θ_0 involved in $f_t = f_t(\theta_0)$.*

For the sake of illustration of Example 4, we run a Monte Carlo simulation in which we simulate $T = 4,000$ observations using the above model where 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 σ_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., ω, α and β, 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 $p(y_t|f_t, \theta)$ 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 $p(\cdot|f, \theta)$. To estimate the p parameters of the model, the MLE is often much more efficient than the QMLE and QLE when the conditional distribution $p(y_t|f_t, \theta)$ 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.

3.2.1 Strong consistency

Theorem 2 establishes the consistency of the MLE $\widehat{\theta}_T$ for QSD models satisfying the stationarity and invertibility conditions stated in Section 2.2. This theorem allows model misspecification and ensures only the convergence of the MLE $\widehat{\theta}_T$ to the pseudo-true parameter θ_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

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

Let $\ell_t(\theta) = \ell(y_t, f_t(\theta), \theta)$ and $\widehat{\ell}_t(\theta) = \ell(y_t, \widehat{f}_t(\theta), \theta)$.

THEOREM 2. (Consistency of MLE under misspecification) *Let the conditions of Lemma 3 hold. Suppose further that ℓ is continuous, $\ell(y_t, \cdot, \theta)$ is differentiable with $E\sup_{\theta\in\Theta}\sup_f \left|\frac{\partial \ell(y_t,f,\theta)}{\partial f}\right|^s < \infty$ for some $s > 0$, the parameter space Θ is compact,*

$\mathbb{E}\sup_{\theta\in\Theta}|\ell_t(\theta)| < \infty$ *and there exists* $\theta_0^* \in \Theta$ *such that* $\mathbb{E}\ell_t(\theta) < \mathbb{E}\ell_t(\theta_0^*)$ *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 QSD 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 Pöstcher and Prucha (1997, Lemma 4.2) and highlights that when the uniqueness fails, the estimator can still be consistent with 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 Pöstcher and Prucha, 1997).

Lemma 5. (Set consistency of MLE under possible misspecification) *Let the conditions of Lemma 3 hold. Suppose further that* ℓ *is continuous,* Θ *is compact, and* $\mathbb{E}\sup_{\theta\in\Theta}|\ell_t(\theta)| < \infty$. *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*

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

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 QSD 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 θ_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 (1) and (3) 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* Θ. *Suppose further that* Θ *is compact and* $\mathbb{E}\sup_{\theta\in\Theta}|\ell_t(\theta)| < \infty$. *Finally, assume* $\mathbb{E}\log^+|f_t| < \infty$ *and* $\mathbb{E}\ell_t(\theta_0) > \mathbb{E}\ell_t(\theta)\ \forall\ \theta \neq \theta_0$. *Then,* $\widehat{\theta}_T \xrightarrow{as} \theta_0 \in \Theta$.

3.2.2 Asymptotic normality

We now turn to the asymptotic normality of the ML and QML estimators for the static parameters of QSD models.

When the QSD model is correctly specified, we can use the martingale difference sequence property of the score at θ_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 Pötscher and Prucha (1997, Definition 6.3). This lemma is written for robust QSD models with bounded updates delivered by a uniformly bounded ψ function 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 (Pötscher and Prucha, Chapter 10).[2]

Let $\widehat{\ell}'_t(\theta_0)$ denote the score evaluated at θ_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 $(\widehat{f}_t, \widehat{f}'_t)$. Define similarly the second-order derivatives $\widehat{\ell}''_t(\theta)$. Let $\ell'_t(\theta)$ and $\ell''_t(\theta)$ be obtained by replacing $\widehat{f}_t(\theta)$ with $f_t(\theta)$ in $\widehat{\ell}'_t(\theta)$ and $\widehat{\ell}''_t(\theta)$.

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

Then, $\{\widehat{f}_t(\theta_0)\}_{t\in\mathbb{N}}$ is NED of size $-q$ on $\{e_t\}_{t\in\mathbb{Z}}$. Additionally, if $|\beta_0| < 1$ and

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

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

(vii) $\sup_{y,X,f} \left|\frac{\partial^2 \psi(y,X,f,\theta_0)}{\partial f \partial y}\right| < \infty$; *(viii)* $\sup_{y,X,f} \left|\frac{\partial^2 \psi(y,X,f,\theta_0)}{\partial^2 f}\right| < \infty$

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, assume the score $\widehat{\ell}'_t(\theta_0)$ is Lipschitz on $(y_t, \widehat{f}_t, \widehat{f}'_t)$,

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

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

[2]In general, the concept of mixing sequences is mostly suitable for linear processes; see e.g. pages 65-68 in Pötscher and Prucha (1997) for a detailed discussion of this point. As such, NED offers a way forward to describe fading memory in nonlinear autoregressive processes, which mixing does not.

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

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

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 Pötscher and Prucha (1997). In this theorem, we assume that the data are NED on an underlying ϕ-mixing sequence of size $-r/(r-1)$. The same result can, however, be obtained for α-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 NED of size* -1 *on a* ϕ*-mixing sequence of size* $-r/(r-1)$ *for some* $r > 2$*, and that*

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

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*

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

where

$$\mathcal{J}(\theta_0^*) = \lim_{T\to\infty} T^{-1}\mathbb{E}\Big(\sum_{t=1}^{T}\widehat{\ell}_t'(\theta_0)\Big)\Big(\sum_{t=1}^{T}\widehat{\ell}_t'(\theta_0)^\top\Big).$$

COROLLARY 2. (Asymptotic normality of MLE under correct specification) *Let* $\{y_t\}_{t\in Z}$ *be generated by (1) and (3) under some* $\theta_0 \in \Theta$*, and let the conditions of Corollary 1 and Lemma 4 hold. Suppose further that*

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

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.*

As a final result, we obtain the asymptotic distribution of the score (also called the Lagrange Multiplier) and log-likelihood ratio tests for testing p_0 linear restrictions on the $p > p_0$ dimensional parameter vector θ_0. The null hypothesis of interest is $\mathrm{H}_0 : R\theta_0 = \boldsymbol{r}$, where R is a given full rank $p_0 \times (p - p_0)$ matrix and $\boldsymbol{r}$ is a given p_0-dimensional vector. Below, $\widehat{\theta}_T^{p_0}$ denotes the MLE of θ_0 in the model constrained by the null.

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

$$\begin{aligned} \mathrm{LM}_T &= T\frac{\partial\widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial\theta^\top}\widehat{\mathcal{I}}^{-1}\frac{\partial\widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial\theta} \xrightarrow{d} \chi^2_{p_0}, \\ \mathrm{LR}_T &= 2T\Big(\widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T) - \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})\Big) \xrightarrow{d} \chi^2_{p_0} \quad \textit{as } T \to \infty, \end{aligned}$$

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

$$\widehat{\mathcal{I}} = -\frac{\partial^2\widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial\theta\partial\theta^\top} \qquad \textit{or} \qquad \widehat{\mathcal{I}} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\ell}_t'(\widehat{\theta}_T^{p_0})\widehat{\ell}_t'^{\top}(\widehat{\theta}_T^{p_0}). \tag{25}$$

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* $\widehat{\ell}_t'(\widehat{\theta}_T^{p_0})$ *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* $\widehat{\ell}_t'(\widehat{\theta}_T^{p_0})$*.*

4 The QSD_T $GARCH(1,1) - T$ model

In this section, we illustrate our general results on an extension of one of the most popular SD volatility models, namely, the $Beta_T$ $GARCH(1,1)$ of Harvey and Chakravarty (2008).

4.1 An extension of the $Beta_T$ $GARCH(1,1)$

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

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

with $\omega > 0$, $\alpha > 0$ and $\beta \geq 0$ to ensure positivity and avoid triviality. Harvey and Chakravarty (2008) show that (26) is the updating equation of an SD model when the $i.i.d.$ sequence (ϵ_t) follows a standardized Student's t distribution T_ν with $\nu > 2$ degrees of freedom (and variance 1). Note that ν plays two roles in this model: it determines the shape of the density of the innovations ϵ_t and bounds the effects of large shocks ϵ_t on future values of the conditional variance (i.e., f_{t+1}) when $\nu < \infty$. It is convenient to reparametrize (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 $Beta_T$ $GARCH(1,1)$ model but disconnect the updating equation of the conditional variance and the density of the innovations. To do so, we also assume $\epsilon_t \sim T_{1/\xi}$, but we introduce an additional parameter ζ in the updating equation that is not related to ξ. We also consider the possibility of introducing a vector X_t of positive exogenous variables. The model, called the QSD_T $GARCH(1,1) - T$ model, is parameterized as follows:[3]

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

where $\epsilon_t \overset{i.i.d.}{\sim} T_{1/\xi}$ with $0 \leq \xi < 1/2$ and $\varpi \geq 0$ is a parameter vector of the same dimension as X_t. In the absence of exogenous variables, this model is identical to a $Beta_T$ $GARCH(1,1)$ model when $\xi = \zeta$ and a standard $GARCH(1,1)$ model with standardized Student's t innovations when $\zeta = 0$.

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

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

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 choose for Ψ 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

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

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 $c = 1,000$. Note that in the empirical application we do not find a single series (out of 400 stocks) for which ζ is significantly negative.

[3]The name QSD_T $GARCH - T$ refers to the fact that the model involves two Student's t distributions. Indeed, the conditional log-density $\log p(y_t|f_t,\theta)$ is a Student's t log-density with $1/\xi$ degrees of freedom while $\rho(y_t,f_t,\theta)$ is another Student's t log-density with $1/\zeta$ degrees of freedom. Consequently, the updating equation of f_t depends on ζ and not on ξ.

4.1.1 Stationarity and positivity conditions

Let us consider the stationarity of the general QSD_T $GARCH(1,1) - T$ model (29) without assuming a particular distribution for (ϵ_t). For the moment, we just assume that (ϵ_t, X_t) is stationary and ergodic with $\mathbb{E}(\epsilon_t^2) = 1$ and $E\|X_t\|^s < \infty$ for some $s > 0$. By the Cauchy root test, it is easy to show that there exists a stationary (ergodic) solution to this model, explicitly given by

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

$\omega_t = \omega + \varpi^\top X_t$, when

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

Note that, since $\alpha \neq 0$, (29) corresponds to (3) with

$$\psi \left(g(f, \epsilon_t), X, f, \theta \right) = \frac{\varpi^\top X}{\alpha} + \Psi_t \epsilon_t^2 f, \qquad \frac{\partial \psi(g(f, \epsilon_t), X, f, \theta)}{\partial f} = \Psi_t \epsilon_t^2.$$

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

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

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

4.1.2 Invertibility of the filter

We now assume that (ϵ_t) is an *i.i.d.* sequence and that conditions (31)-(32) hold true. When $E\|X_t\|^r < \infty$ for some $r > 0$, all conditions of Lemma 2 are satisfied, which shows that the stationary solution of the QSD_T $GARCH(1,1) - T$ model is such that $\mathbb{E}|y_t|^s < \infty$ for some $s > 0$. Assume also that

$$\theta = (\omega, \alpha, \beta, \zeta, \varpi)^\top \in \Theta \subset [\underline{\omega}, \overline{\omega}] \times [\underline{\alpha}, \overline{\alpha}] \times [\underline{\beta}, \overline{\beta}] \times [\underline{\zeta}, \overline{\zeta}] \times [\underline{\varpi}, \overline{\varpi}] \tag{34}$$

with $0 < \underline{\omega} < \overline{\omega}$, $0 < \underline{\alpha} < \overline{\alpha}$, $0 \leq \underline{\beta} \leq \overline{\beta} < 1$, $0 \leq \underline{\zeta} \leq \overline{\zeta} < 1/2$ and $0 \leq \underline{\varpi} \leq \overline{\varpi}$ (componentwise). Since we have

$$\psi(y_t, X_t, f_t, \theta) \;=\; \frac{\varpi^\top X_t}{\alpha} + \Psi \left(\frac{1+\zeta}{1-2\zeta+\zeta\frac{y_t^2}{f_t}} \right) y_t^2,$$

condition (i) of Lemma 3 is satisfied when $\|X_t\|$ and y_t admit a small-order moment. Since

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

the uniform invertibility condition (ii) is satisfied when

$$\mathbb{E}\log\left(c_4\overline{\alpha}\frac{1+\overline{\zeta}}{\left(1-2\overline{\zeta}+\overline{\zeta}\frac{y_t^2}{\underline{f}}\right)^2}\overline{\zeta}\frac{y_t^4}{\underline{f}}+\overline{\beta}\right)<0, \tag{35}$$

where $\underline{f}=\underline{\omega}/(1-\underline{\beta})$ is a lower bound for the time-varying volatility. Note that the expectation of the left-hand side of (35) 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 Θ), the supremum involved in condition (ii) of Lemma 3 can be computed numerically.

4.1.3 Derivatives of the filter

With the notation (34), (9) holds with

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

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 $f_t \geq \omega > 0$, it can be seen that $\mathbb{E}\,\|A_t\|^s < \infty$ for some $s > 0$ when $\|X_t\|$ and y_t admit a small-order moment. Therefore, $\mathbb{E}\log^+\|A_t\| < \infty$, and (i) of Lemma 4 is satisfied. Similarly, it can be seen that the other conditions of that lemma hold true.

4.2 Statistical inference

We now consider the estimation of the $QSD_T\ GARCH-T$ model. We thus assume the standardized Student's t conditional distribution

$$p(y \mid f, \theta) = \frac{1}{\sqrt{f}} p_\nu \left(\frac{y}{\sqrt{f}} \right), \tag{36}$$

$$p_\nu(\epsilon) = \frac{1}{\sqrt{\pi(\nu-2)}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\epsilon^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}, \tag{37}$$

where $p_\nu(\epsilon)$ is the standardized Student's t distribution (with mean 0, unit variance and $\nu > 2$) and denoted as T_ν in short. Setting $\xi = 1/\nu$ and imposing $0 \leq \xi < 1/2$, the Gaussian conditional distribution case corresponds to $\xi = 0$. Let $\theta = (\omega, \alpha, \beta, \zeta, \xi, \varpi^\top)^\top$, Θ a compact subset of $(0,\infty)^2 \times [0,1] \times (-1, 1/2) \times [0, 1/2) \times [0,\infty)^{d-1}$ 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

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

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

4.2.1 Testing the $Beta_T\ GARCH(1,1)$ restriction

The standard $Beta_T\ GARCH(1,1)$ is obtained when $0 < \zeta = \xi < 1/2$ and $\Psi(x) = x$. It is thus of interest to test the null hypothesis $H_0 : \xi_0 = \zeta_0$. This hypothesis can be written as $H_0 : K\theta_0 = 0$ with $K = (0_3^\top, 1, -1, 0_{d-1}^\top)$. 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 χ_1^2 under H_0. The test of critical region $\{W_T^{\zeta,\xi} > \chi_1^2(1-\alpha^*)\}$ thus has the asymptotic level α^*.

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

$$LM_T^{\zeta,\xi} = T \frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial \theta^\top} \widehat{\mathcal{I}}^{-1} \frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial \theta}, \quad LR_T^{\zeta,\xi} = 2T \Big(\widehat{\ell}_T(\widehat{\theta}_T) - \widehat{\ell}_T(\widehat{\theta}_T^{p_0}) \Big).$$

For the Wald statistics, it is natural to take $\widehat{\Sigma} = \widehat{\mathcal{I}}^{-1}$, where $\widehat{\mathcal{I}}$ is defined by (25), replacing $\widehat{\theta}_T^{p_0}$ by $\widehat{\theta}_T$.

4.2.2 Testing the $GARCH - T$ restriction

The $GARCH(1,1)-T$ volatility model 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 QSD_T $GARCH-T$ model defined by (36) and (29), with $-1 < \zeta < 1/2$ and Ψ satisfying (31). Another possibility would be to test $\zeta_0 = 0$ in the model defined by (36) 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 Θ 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

$$W_T^\zeta = T\widehat{\theta}_T^\top \boldsymbol{e}_4 \left(\boldsymbol{e}_4^\top \widehat{\Sigma} be_4\right)^{-1} \boldsymbol{e}_4^\top \widehat{\theta}_T, \quad \boldsymbol{e}_4^\top = (0_3^\top, 1, 0_d^\top),$$

has asymptotic level α^*.

4.2.3 Testing the standard $GARCH$

The parameter of main interest is often the volatility $f_t = f(\theta_0)$ with $\theta_0 = (\omega_0, \alpha_0, \beta_0, \zeta_0, \varpi^\top)^\top$ and Θ changed accordingly. It is then desirable to estimate θ_0 without assuming (36) or any other particular conditional distribution.

The benchmark estimator in this framework is the QMLE

$$\widehat{\theta}_{QMLE} = \arg\min_{\theta\in\Theta} \frac{1}{T} \sum_{t=t_0+1}^{T} \frac{y_t^2}{\widehat{f}_t(\theta)} + \log \widehat{f}_t(\theta). \tag{38}$$

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

$$\widehat{\theta}_T = \arg\min_{\theta\in\Theta} \left\| \frac{1}{T} \sum_{t=t_0+1}^{T} \frac{y_t^2 - \widehat{f}_t(\theta)}{\widehat{\sigma}_t^2(\theta)} \frac{\partial \widehat{f}_t(\theta)}{\partial \theta} \right\| \tag{39}$$

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 (38) and (39) are equivalent. However, they are not equivalent if, for instance, one takes $\widehat{\sigma}_t^2(\theta) = \widehat{f}_t(\theta)$.

4.3 Monte Carlo simulation

In this section, we present a Monte Carlo experiment that studies the finite-sample properties of the $QSD_T\ GARCH(1,1)-T$ model as well as some models nested in this specification.

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

$$y_t = \mu + \sqrt{f_t}\epsilon_t \tag{40}$$

$$\epsilon_t \sim T_{1/\xi} \tag{41}$$

$$f_{t+1} = \omega + \varpi X_t + \alpha\frac{1+\zeta}{1-2\zeta+\zeta\epsilon_t^2}\epsilon_t^2 f_t + \beta f_t \tag{42}$$

with $0 < \xi < 1/2$ and $-1 < \zeta < 1/2$ and where $T_{1/\xi}$ is defined in (37). Recall that this model is a pure $Beta_T\ GARCH(1,1)$ in the absence of an explanatory variable and when $\xi = \zeta$, and it is a $GARCH(1,1)-T$ model when $\zeta = 0$. The explanatory variable X_t is taken from the empirical application and corresponds to historical data of the square of the VIX index converted from an annual to a daily horizon.

In all simulations, we set the parameters to a rounded value of the average (over all stocks) of the estimated parameters obtained in Section 5.1.[4] More specifically, we set $\mu = 0.06, \omega = 0.08, \varpi = 0.13, \alpha = 0.10$ and $\beta = 0.83$. In the first simulation (i.e., Table 1), we set $\xi = \zeta = 0.2$ so that the model is a $Beta_T\ GARCH(1,1)$ 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 $QSD_T\ GARCH(1,1)-T$ 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., $GARCH(1,1)-T$, $Beta_T\ GARCH(1,1)$ and $QSD_T\ GARCH(1,1)-T$) are estimated by ML. The fourth model is the $QSD_T\ GARCH(1,1)$ estimated by Gaussian QML (ξ is therefore not estimated). Note that during the optimization, the positivity of the conditional variance of the $QSD_T\ GARCH(1,1)-T$ models is imposed by replacing (42) by (29), as discussed in Section 4.

In all cases considered in this section, σ_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 into 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 into two parts.

[4]Unlike in the empirical application, we do not consider an $AR(1)$ specification in the simulations.

The first one contains summary statistics on the estimated parameters, while the second reports rejection frequencies of two LRTs. Figures at the right of the name of the models are the empirical biases over 1,000 replications. Figures in parentheses 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 LRTs computed from the ML estimates. The first one is for the null hypothesis that the true model is a $Beta_T\ GARCH(1,1)$, 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. For this reason, for ease of reading of the results, an asterisk is added after the name of the models that do not nest the DGP.

Some comments are in order.

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

- When the true model is a $Beta_T\ GARCH(1,1)$ (see Table 1), the inverse of the degree of freedom of innovations ξ is slightly more precisely estimated with the $Beta_T\ GARCH(1,1)$ model than with the $QSD_T\ GARCH(1,1)-T$, but the difference is marginal. Indeed, the RMSE is only 0.001 higher for the latter when $T = 3,000$ and almost identical when $T = 4,000$. Furthermore, while the biases of ξ and ζ are small, the RMSE of ζ is between three and four times higher than for ξ. This is a consequence of the fact that the identification of ζ is only possible from the observations for which the shocks are truncated, whereas all observations can be used to identify ξ. 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 α) are biased when wrongly imposing the assumption that $\zeta = 0$, as shown in Tables 1 and 2.

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

- As expected again, the QML of the $QSD_T\ GARCH(1,1)$ is less precise than its ML version. The biases are higher than for the ML, while the RMSEs are approximately 20-25% higher.

- The coverage probabilities of the parameters of the QSD_T $GARCH(1,1)-T$ are satisfactory, except for ζ. For a sample size of 3,000 observations, the true value of ζ belongs to the 95% confidence interval drawn from the asymptotic distribution in approximately 83 to 84% (resp. 79 to 83%) of the cases for the MLE (resp. QMLE). 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 ζ 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 ζ are on average too small compared to the RMSE of the estimated ζ parameter.

- While statistical inference on ζ relying on its standard error (e.g., t-tests and Wald tests) requires a very large sample, the LRT on ζ 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 to 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 $GARCH-T$ 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 $GARCH(1,1)-T$ (see Table 3) and decent power when the true model is a QSD_T $GARCH(1,1)-T$ with $\xi = 0.2$ and $\zeta = 0.1$ (approximately 30% for a nominal size of 5% and a sample size of 4,000 observations). The power of course increases with the distance between ξ and ζ.

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 QSD_T $GARCH(1,1)-T$ 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 QSD_T $GARCH(1,1)-T$ (thick solid pink line), $Beta_T$ $GARCH(1,1)$ (thin green dashed line), $GARCH(1,1)-T$ (thin blue line with long dashes) and the true conditional volatility (black solid line). It is clear from this graph that unlike the QSD_T $GARCH(1,1)-T$, the $GARCH-T$ model overestimates the volatility during approximately two weeks (i.e., approximately 15 observations) following the large shock (occurring at observation 1475), while the $Beta_T$ $GARCH(1,1)$ underestimates the volatility during the same period.

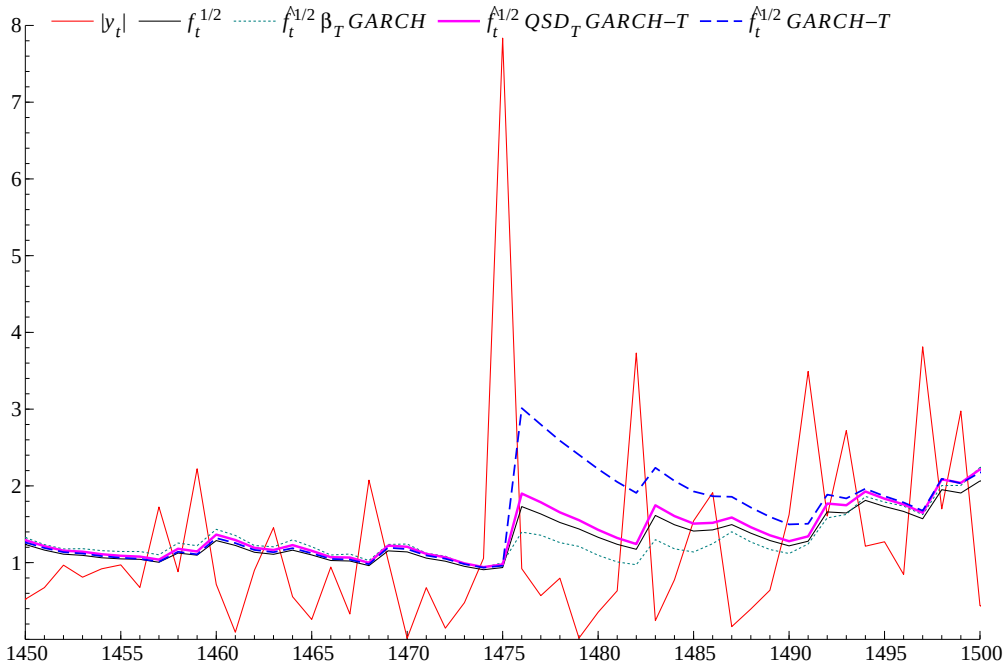


Figure 1: Fifty observations around a large shock for a DGP corresponding to a QSD_T $GARCH(1,1) - T$ with $\xi = 0.2$ and $\zeta = 0.1$ (as in Table 2). The $GARCH(1,1) - T$, $Beta_T$ $GARCH(1,1)$ and QSD_T $GARCH(1,1) - T$ models are estimated by ML on 4,000 observations.

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

	$T = 3,000$						
	μ	ω	ϖ	α	β	ξ	ζ
	0.06	0.08	0.13	0.10	0.83	0.2	0.2
$GARCH(1,1) - T$ ML*	0.000	0.022	0.031	-0.038	0.022	0.002	
	(0.026)	(0.032)	(0.045)	(0.019)	(0.035)	(0.017)	
	[0.961]	[0.937]	[0.918]	[0.289]	[0.737]	[0.947]	
$\beta_T\ GARCH(1,1)$	0.000	0.008	0.011	-0.000	-0.006	-0.001	
	(0.026)	(0.032)	(0.045)	(0.019)	(0.035)	(0.017)	
	[0.964]	[0.960]	[0.949]	[0.945]	[0.947]	[0.956]	
$QSD_T\ GARCH(1,1)$ QML	0.001	0.009	0.014	0.001	-0.018		0.025
	(0.031)	(0.048)	(0.087)	(0.029)	(0.066)		(0.111)
	[0.960]	[0.922]	[0.913]	[0.913]	[0.909]		[0.793]
$QSD_T\ GARCH(1,1) - T$ ML	0.000	0.007	0.011	-0.002	-0.009	-0.001	0.008
	(0.026)	(0.031)	(0.045)	(0.020)	(0.039)	(0.018)	(0.089)
	[0.963]	[0.962]	[0.944]	[0.934]	[0.950]	[0.957]	[0.852]
	1%	5%	10%				
$H_0 : \xi = \zeta$	0.901	6.607	12.713				
$H_0 : \zeta = 0$	81.481	92.793	95.195				
	$T = 4,000$						
	μ	ω	ϖ	α	β	ξ	ζ
	0.06	0.08	0.13	0.10	0.83	0.2	0.2
$GARCH(1,1) - T$ ML*	0.000	0.017	0.024	-0.039	0.027	0.003	
	(0.021)	(0.027)	(0.035)	(0.017)	(0.029)	(0.016)	
	[0.970]	[0.925]	[0.942]	[0.196]	[0.684]	[0.950]	
$\beta_T\ GARCH(1,1)$	0.000	0.005	0.007	-0.000	-0.004	-0.000	
	(0.021)	(0.027)	(0.035)	(0.017)	(0.029)	(0.016)	
	[0.966]	[0.939]	[0.955]	[0.938]	[0.959]	[0.953]	
$QSD_T\ GARCH(1,1)$ QML	0.001	0.007	0.009	-0.000	-0.013		0.021
	(0.025)	(0.039)	(0.051)	(0.025)	(0.049)		(0.100)
	[0.967]	[0.940]	[0.935]	[0.927]	[0.936]		[0.804]
$QSD_T\ GARCH(1,1) - T$ ML	0.000	0.005	0.007	-0.001	-0.006	-0.001	0.008
	(0.021)	(0.027)	(0.036)	(0.018)	(0.033)	(0.016)	(0.078)
	[0.967]	[0.941]	[0.949]	[0.933]	[0.942]	[0.955]	[0.864]
	1%	5%	10%				
$H_0 : \xi = \zeta$	1.100	6.600	12.900				
$H_0 : \zeta = 0$	93.700	97.600	98.200				

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 the four models. The figures at the right of the name of the models are the empirical biases over 1,000 replications. The figures in parentheses 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. Models highlighted with an asterisk after their name do not nest the DGP.

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

			$T = 3,000$				
	μ	ω	ϖ	α	β	ξ	ζ
	0.06	0.08	0.13	0.10	0.83	0.2	0.1
$GARCH(1,1)-T$ ML*	0.000	0.017	0.025	-0.028	0.007	0.001	
	(0.025)	(0.030)	(0.042)	(0.021)	(0.041)	(0.018)	
	[0.961]	[0.939]	[0.926]	[0.506]	[0.872]	[0.950]	
$\beta_T\ GARCH(1,1)$*	0.000	0.006	0.009	0.008	-0.021	-0.004	
	(0.025)	(0.030)	(0.042)	(0.021)	(0.041)	(0.018)	
	[0.966]	[0.964]	[0.946]	[0.954]	[0.944]	[0.944]	
$QSD_T\ GARCH(1,1)$ QML	0.001	0.008	0.013	0.002	-0.016		0.035
	(0.029)	(0.045)	(0.079)	(0.031)	(0.059)		(0.103)
	[0.958]	[0.914]	[0.909]	[0.910]	[0.901]		[0.817]
$QSD_T\ GARCH(1,1)-T$ ML	0.000	0.007	0.011	-0.001	-0.009	-0.001	0.013
	(0.025)	(0.031)	(0.044)	(0.021)	(0.039)	(0.018)	(0.075)
	[0.965]	[0.962]	[0.939]	[0.921]	[0.939]	[0.957]	[0.841]
	1%	5%	10%				
$H_0 : \xi = \zeta$	10.010	25.125	36.637				
$H_0 : \zeta = 0$	56.356	73.674	80.981				
			$T = 4,000$				
	μ	ω	ϖ	α	β	ξ	ζ
	0.06	0.08	0.13	0.10	0.83	0.2	0.1
$GARCH(1,1)-T$ ML*	0.000	0.014	0.019	-0.029	0.012	0.001	
	(0.020)	(0.025)	(0.034)	(0.019)	(0.035)	(0.016)	
	[0.969]	[0.932]	[0.944]	[0.405]	[0.843]	[0.953]	
$\beta_T\ GARCH(1,1)$*	0.000	0.004	0.005	0.007	-0.018	-0.004	
	(0.020)	(0.025)	(0.034)	(0.019)	(0.035)	(0.016)	
	[0.965]	[0.935]	[0.944]	[0.946]	[0.950]	[0.933]	
$QSD_T\ GARCH(1,1)$ QML	0.001	0.008	0.010	0.001	-0.013		0.031
	(0.023)	(0.039)	(0.055)	(0.026)	(0.047)		(0.092)
	[0.964]	[0.944]	[0.942]	[0.928]	[0.942]		[0.851]
$QSD_T\ GARCH(1,1)-T$ ML	0.000	0.005	0.006	-0.001	-0.006	-0.001	0.012
	(0.020)	(0.025)	(0.034)	(0.019)	(0.031)	(0.016)	(0.063)
	[0.966]	[0.945]	[0.943]	[0.931]	[0.958]	[0.956]	[0.879]
	1%	5%	10%				
$H_0 : \xi = \zeta$	13.300	29.800	40.100				
$H_0 : \zeta = 0$	71.600	85.500	90.500				

Note: see Table 1

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

			$T = 3,000$				
	μ	ω	ϖ	α	β	ξ	ζ
	0.06	0.08	0.13	0.10	0.83	0.2	0
$GARCH(1,1)-T$ ML	0.000	0.006	0.008	0.000	-0.005	-0.001	
	(0.025)	(0.027)	(0.037)	(0.046)	(0.056)	(0.020)	
	[0.965]	[0.957]	[0.944]	[0.949]	[0.938]	[0.956]	
β_T $GARCH(1,1)$*	0.000	0.004	0.005	0.040	-0.046	-0.010	
	(0.025)	(0.027)	(0.037)	(0.046)	(0.056)	(0.020)	
	[0.966]	[0.954]	[0.935]	[0.518]	[0.730]	[0.903]	
QSD_T $GARCH(1,1)$ QML	0.001	0.006	0.008	0.004	-0.009		0.013
	(0.029)	(0.038)	(0.052)	(0.030)	(0.043)		(0.038)
	[0.953]	[0.924]	[0.924]	[0.921]	[0.911]		[0.834]
QSD_T $GARCH(1,1)-T$ ML	0.000	0.005	0.007	0.000	-0.005	-0.001	0.004
	(0.025)	(0.026)	(0.036)	(0.019)	(0.029)	(0.018)	(0.019)
	[0.967]	[0.958]	[0.945]	[0.940]	[0.936]	[0.953]	[0.842]
	1%	5%	10%				
$H_0 : \xi = \zeta$	92.893	97.598	98.599				
$H_0 : \zeta = 0$	1.702	8.008	14.915				
			$T = 4,000$				
	μ	ω	ϖ	α	β	ξ	ζ
	0.06	0.08	0.13	0.10	0.83	0.2	0
$GARCH(1,1)-T$ ML	0.000	0.004	0.004	-0.000	-0.002	-0.001	
	(0.020)	(0.023)	(0.029)	(0.044)	(0.052)	(0.019)	
	[0.970]	[0.946]	[0.956]	[0.945]	[0.948]	[0.955]	
β_T $GARCH(1,1)$*	0.000	0.002	0.002	0.039	-0.044	-0.010	
	(0.020)	(0.023)	(0.029)	(0.044)	(0.052)	(0.019)	
	[0.970]	[0.926]	[0.942]	[0.395]	[0.646]	[0.877]	
QSD_T $GARCH(1,1)$ QML	0.001	0.006	0.007	0.003	-0.007		0.008
	(0.023)	(0.032)	(0.041)	(0.025)	(0.034)		(0.025)
	[0.968]	[0.937]	[0.945]	[0.926]	[0.945]		[0.866]
QSD_T $GARCH(1,1)-T$ ML	0.000	0.003	0.004	-0.000	-0.002	-0.001	0.003
	(0.020)	(0.022)	(0.029)	(0.017)	(0.023)	(0.016)	(0.015)
	[0.971]	[0.944]	[0.956]	[0.947]	[0.961]	[0.954]	[0.870]
	1%	5%	10%				
$H_0 : \xi = \zeta$	98.300	99.500	99.700				
$H_0 : \zeta = 0$	1.500	5.700	10.900				

Note: see Table 1

5 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, as well as a few stocks for which one of the competing models encountered convergence problems. We are left with 400 stocks.

The application is divided in two. We first consider volatility models based on symmetric Student's t densities and then extend our analysis by considering models with skewed Student's t densities.

5.1 Symmetric densities

A $GARCH(1,1)-T$, a β_T $GARCH(1,1)$ and a QSD_T $GARCH(1,1)-T$ are estimated by ML on all series of log-returns.[5] The explanatory variable X_t used in the conditional variance equation f_{t+1} of all models is the square of the VIX index converted from an annual to a daily horizon.

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.

The first two columns of Table 4 contain the rejection frequencies at the 5% nominal level (over the 400 stocks) of two LRTs. Interestingly, the null hypothesis $H_0 : \zeta = 0$ is rejected in approximately 90% of the cases, suggesting that downweighting large shocks in the conditional variance of these US stocks is empirically relevant.

These results naturally call for the use of an SD model rather than a $GARCH$ dynamic. However, the null hypothesis $H_0 : \xi = \zeta$ is rejected in more than 50% of the cases (again at the 5% nominal level) suggesting that the additional flexibility of the QSD_T $GARCH(1,1)-T$ over the β_T $GARCH(1,1)$ is empirically relevant. Furthermore, all the estimated ζ values are positive, and thus specification (28) can be used instead of (29).

The difference $\widehat{\xi} - \widehat{\zeta}$ is plotted in Figure 2 for the 400 US stocks (sorted in alphabetical order of the ticker name). A full (resp. empty) circle corresponds to a significant (resp. insignificant) difference (according to an LRT at the 5% nominal level). For all (but one) stocks for which $\widehat{\xi} \neq \widehat{\zeta}$, $\widehat{\xi} > \widehat{\zeta}$, suggesting that the downweighting of the β_T $GARCH(1,1)$ is too strong.

To visualize the added value of the QSD_T $GARCH(1,1)-T$ model over the $GARCH-T$ and β_T $GARCH(1,1)$ models, we randomly selected a stock for

[5] To account for possible serial correlation in the data, an AR(1) specification is used for all series.

Table 4: Rejection frequencies (at 5%) of two LRTs and a goodness-of-fit test for the models with symmetric Student's t densities

LRT		PIT test		
$H_0 : \xi = \zeta$	$H_0 : \zeta = 0$	$GARCH - T$	β_T $GARCH$	QSD_T $GARCH - T$
53.5	89.5	40.0	46.0	45.7

The figures reported in the table are rejection frequencies over the 400 stocks and a nominal size of 5%. The figures in the two columns under LRT are for the LRT whose null hypothesis is reported just above. These hypothesis are tested on the MLEs of the QSD model. The figures reported in the three columns under the PIT test χ^2 test are for the PIT test applied on the residuals of the corresponding model.

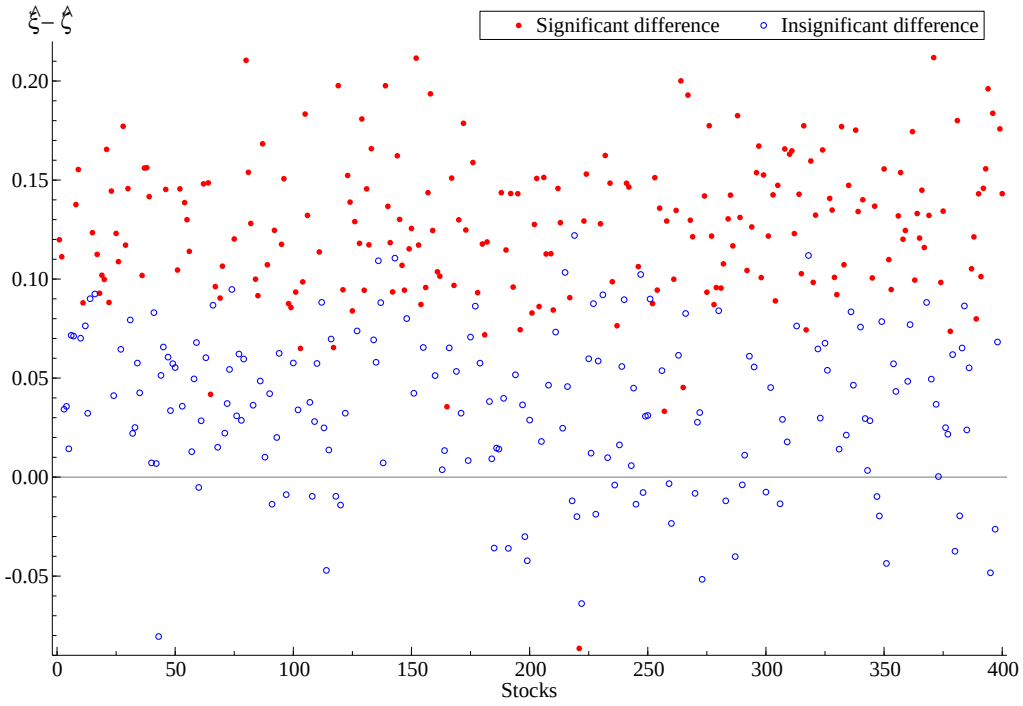


Figure 2: $\widehat{\xi}-\widehat{\zeta}$ for the QSD_T $GARCH(1,1)-T$ estimated on the 400 US stocks. A full (resp. empty) circle corresponds to a significant (resp. insignificant) difference (according to an LRT at the 5% nominal level).

which both the null hypotheses $H_0 : \xi = \zeta$ and $H_0 : \zeta = 0$ are rejected. We choose Centene (whose ticker is CNC), a multi-line healthcare enterprise.

The MLEs of parameters entering the conditional variance equation of the three models obtained for daily log-returns of CNC during the period spanning from December 2001 to the end of February 2019 (i.e., 4330 observations) are reported in Table 5 together with the log-likelihood values and the two LRT statistics of the tests presented above (with the corresponding p-values in squared brackets). Interestingly, the estimated ξ parameters of the $GARCH(1,1)-T$ and β_T $GARCH(1,1)$ do not differ much and are approximately equal to 0.25, which corresponds to a degree of freedom of the Student's t distribution close to 4. The log-likelihood of the β_T $GARCH(1,1)$ is very close to the one of the $GARCH(1,1)-T$ model but slightly lower. Given that the two models are not nested, LRTs cannot be used to discriminate between these two models. Importantly, while $\widehat{\xi}$ is also close to 0.25 for the QSD_T $GARCH(1,1)-T$ model, $\widehat{\zeta}$ is approximately 0.05 (and therefore $1/\widehat{\zeta}$ is close to 20), suggesting that the β_T $GARCH(1,1)$ downweights the large shocks far too much. To help visualize the differences among 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 QSD_T $GARCH(1,1)-T$ nests the other two models, we can write the NIC of the three models as the function mapping the shocks ϵ_t to $\frac{1+\zeta}{1-2\zeta+\zeta\epsilon_t^2}\epsilon_t^2$, where $\zeta = 0$ for the $GARCH(1,1)-T$ model and $\zeta = \xi$ for the β_T $GARCH(1,1)$. We see from Figure 3 that the NIC of the QSD_T $GARCH(1,1)-T$ for the CNC stock lies between the NIC of the other two models.

To see the impact of different NICs on the estimated conditional volatilities, the absolute value of the daily log-returns of CNC (thin solid red line) as well as the estimated conditional volatilities of the $GARCH(1,1)-T$ (thin blue line with long dashes), β_T $GARCH(1,1)$ (thin green dashed line) and QSD_T $GARCH(1,1)-T$ (thick solid pink line) estimated by ML (on the full period) are plotted for the sub-period spanning from the beginning of March 2008 to the end of April 2008 in Figure 4. The box highlights a period of several days around a large shock (i.e., an absolute return of approximately 25 %) and for which we see large differences between the estimated conditional standard deviations of the three models. In line with what we observed in Figure 1 for simulated data, the conditional volatility of the $GARCH(1,1)-T$ model is much higher than that of the β_T $GARCH(1,1)$ for several days after the shock while the conditional volatility of the QSD_T $GARCH(1,1)-T$ lies between the two.

As pointed out by a referee, the ability of the QSD_T $GARCH(1,1)-T$ to outperform the other two models in most cases does not mean this model adequately captures the main features of the data. We therefore apply the goodness-of-fit test proposed by Diebold, Gunther and Tay (1998). They show that under the null hypothesis of correct specification, the probability integral trans-

Table 5: MLEs of the parameters of the conditional variance equation of the $GARCH(1,1)-T$, β_T $GARCH(1,1)$ and QSD_T $GARCH(1,1)-T$ for CNC during the period spanning from December 2001 to the end of February 2019 (i.e., 4,330 observations).

	$GARCH(1,1)-T$	β_T $GARCH(1,1)$	QSD_T $GARCH(1,1)-T$
ω	1.1278	0.9857	0.9833
	(0.1932)	(0.1850)	(0.1842)
ϖ	0.8074	0.6570	0.6890
	(0.1573)	(0.1469)	(0.1493)
α	0.1574	0.2256	0.2015
	(0.0266)	(0.0283)	(0.0320
β	0.5394	0.5076	0.5577
	(0.0603)	(0.0621)	(0.0590)
ξ	0.2593	0.2568	0.2584
	(0.0136)	(0.0135)	(0.0137)
ζ			0.0479
			(0.0270)
Log-Likelihood	-9630.6	-9630.8	-9626.4
$H_0:\xi=\zeta$	8.812		
	[0.003]		
$H_0:\zeta=0$	8.4169		
	[0.004]		

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).

form (PIT) series $\{u_t\}_{t=1}^T \overset{i.i.d.}{\sim} U(0,1)$ so that $u_t \overset{i.i.d.}{\sim} (1/2, 1/12)$ and therefore $\sum_{t=1}^T u_t \sim N(T/2, T/12)$ as $T \to \infty$, which suggests rejecting the null of correct specification at the 5% nominal level when $|(\sum_{t=1}^T u_t - T/2)|/\sqrt{T/12} > 1.96$. This result is only true when the PIT series is observed without error, while in practice, it is computed from the cumulative distribution function of the residuals. Since the MLEs of the models considered in this section are asymptotically Gaussian under the assumption of correct specification, we can apply Pierce's (1982) theorem to derive the asymptotic distribution of $\sum_{t=1}^T u_t$ conditional on the MLEs of the model fitted on the data.[6]

[6]The same approach is used by Tse (2002) and Lambert, Laurent and Veredas (2012) to account for the estimation error when testing, respectively, the null hypothesis of no ARCH effects or no conditional skewness in the standardized residuals of a GARCH-type model. Monte Carlo simulation results, not reported here to save space, show that the PIT test accounting for the estimation error by applying Pierce's (1982) theorem has no significant size distortion, while the one ignoring the estimation error has a strong size distortion (i.e., an empirical size close

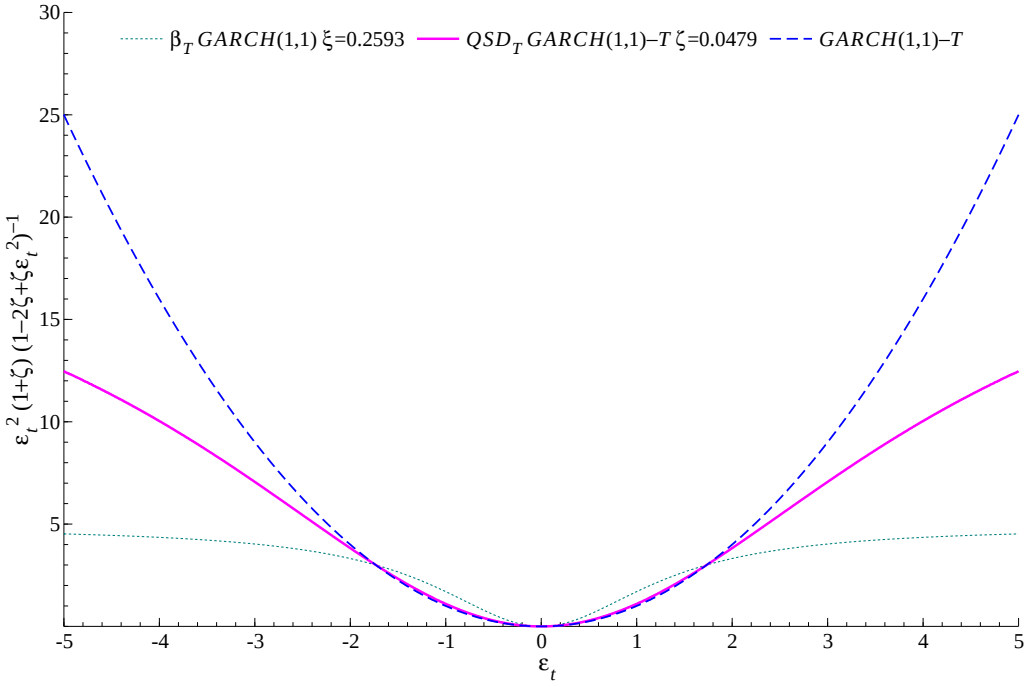


Figure 3: News impact curve for the MLEs of the QSD_T $GARCH(1,1)-T$, β_T $GARCH(1,1)$ and $GARCH(1,1)-T$ for the ticker CNC (Centene).

The rejection frequencies (over the 400 stocks and for a nominal size of 5%) of the above PIT test (accounting for estimation error) are reported in the last three columns of Table 4 (under PIT) for the three models considered in this section. The results clearly suggest that the three models are rejected in more than 40% of the cases, which calls for the use of a different model.

5.2 Skewed densities

The distribution of daily stock returns is known to have heavy tails (as illustrated in the previous section) and is also often found to be left-skewed, as shown, for instance, by Giot and Laurent (2003) and Harvey and Lange (2016). It is also known that negative shocks have a deeper impact on the volatility than positive shocks of the same magnitude (so-called leverage effect). Hansen and Lunde (2005) find that the leverage is actually empirically more relevant than tails for forecasting volatility.

Harvey and Lange (2016) are the first to consider a skewed distribution in the

to 0% for a nominal size of 1, 5 or even 10%). The results are similar to those reported in Tse (2002) and Lambert, Laurent and Veredas (2012).

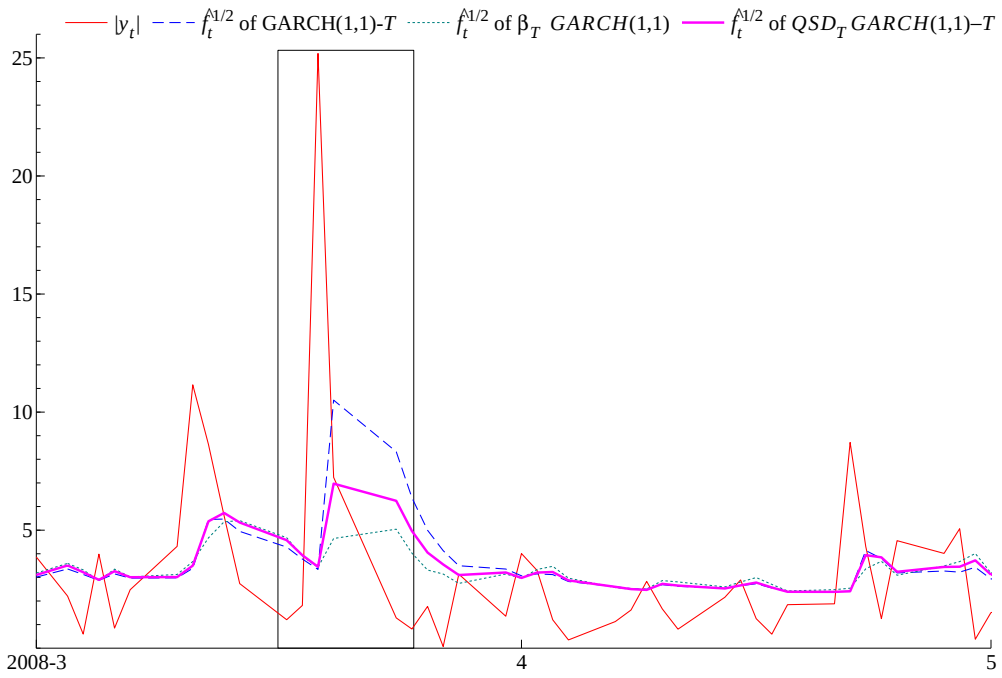


Figure 4: Absolute value of log-returns and estimated conditional volatility of the $GARCH(1,1)-T$, β_T $GARCH(1,1)$ and QSD_T $GARCH(1,1)-T$ estimated by ML for the ticker CNC (Centene). The graph only shows the sub-period spanning from the beginning of March 2008 to the end of April 2008.

context of SD volatility models. More specifically, they propose an SD model in which the innovations are assumed to follow an asymmetric generalized t distribution. In this section, we consider a constrained version of this density corresponding to the non-standardized skewed Student's t (ST) of Zhu and Galbraith (2010) with degree of freedom $1/\xi$ and asymmetry parameter κ, which we denote as $ST_{1/\xi,\kappa}$. If $y_t = \sqrt{f_t}\epsilon_t$ and ϵ_t follows a $ST_{1/\xi,\kappa}$ distribution, the log-likelihood can be written as

$$\log p\left(y_t \mid f_t, \theta\right) = \ln(\delta K_{1/\xi}) - \frac{1}{2}\ln\left(f_t\right) - \left(\frac{1/\xi+1}{2}\right)\ln\left(1+\frac{\xi\epsilon_t^2}{4\kappa_t^2}\right), \tag{43}$$

where $K_v \equiv \Gamma((v+1)/2)/[\sqrt{\pi v}\Gamma(v/2)]$ (with $\Gamma(\cdot)$ the gamma function), $\kappa_t = \kappa$ (resp. $1-\kappa$) if $\epsilon_t \leq 0$ (resp. $\epsilon_t > 0$) and

$$\begin{aligned} m &= \tfrac{4}{1-\xi}\left[(1-\kappa)^2 - \kappa^2\right]K_{1/\xi} \\ \delta^2 &= \tfrac{4}{1-2\xi}\left[\kappa^3 + (1-\kappa)^3\right] - m^2. \end{aligned}$$

Importantly, the ST density is symmetric when $\kappa = 1/2$ and left-skewed (resp.

right skewed) when $\kappa > 1/2$ (resp. $\kappa < 1/2$). Interestingly, Zhu and Galbraith (2010) show that the skewed Student's t distributions of Hansen (1994) and Fernandez and Steel (1998) can be recovered as special cases of the above ST density.

While Harvey and Lange (2016) consider the asymmetric generalized t distribution in the context of an SD model for the log of the conditional variance (i.e., in the spirit of an EGARCH model), we report below the specification of an SD model for the conditional variance when the innovations follow the above $ST_{1/\xi,\kappa}$ distribution.[7] This model, denoted $Beta_{ST}\ GARCH(1,1)$, takes the form

$$f_{t+1} = \omega + \alpha \frac{(1+\xi)}{4\kappa_t^2 + \xi\epsilon_t^2}\epsilon_t^2 f_t + \beta f_t. \tag{44}$$

When $\kappa > 1/2$, the ST density is left-skewed, and in this case, the $Beta_{ST}$ $GARCH(1,1)$ produces volatilities that are smaller after negative shocks than after positive shocks of the same magnitude. Indeed, when $\kappa = 0.6$, $\kappa_t^2 = 0.6^2$ when $\epsilon_t \leq 0$ while $\kappa_t^2 = 0.4^2$ when $\epsilon_t > 0$. Unfortunately, this contradicts the leverage effect. This issue does not affect the larger class of QSD models because ρ can be disconnected from $\log p\left(y_t \mid f_t, \theta\right)$. To preserve the flexibility and generality of the ST density, an $ST_{1/\xi,\kappa}$ distribution can be assumed for the innovations ϵ_t, while $\rho\left(y_t, X_t, f_t, \theta\right)$ can be chosen to be the log-likelihood of an ST density with different parameters, i.e., a $ST_{1/\zeta,\tau}$ distribution, which leads to the following $QSD_{ST}\ GARCH(1,1) - ST$ model:

$$f_{t+1} = \omega + \alpha \frac{(1+\zeta)}{4\tau_t^2 + \zeta\epsilon_t^2}\epsilon_t^2 f_t + \beta f_t, \tag{45}$$

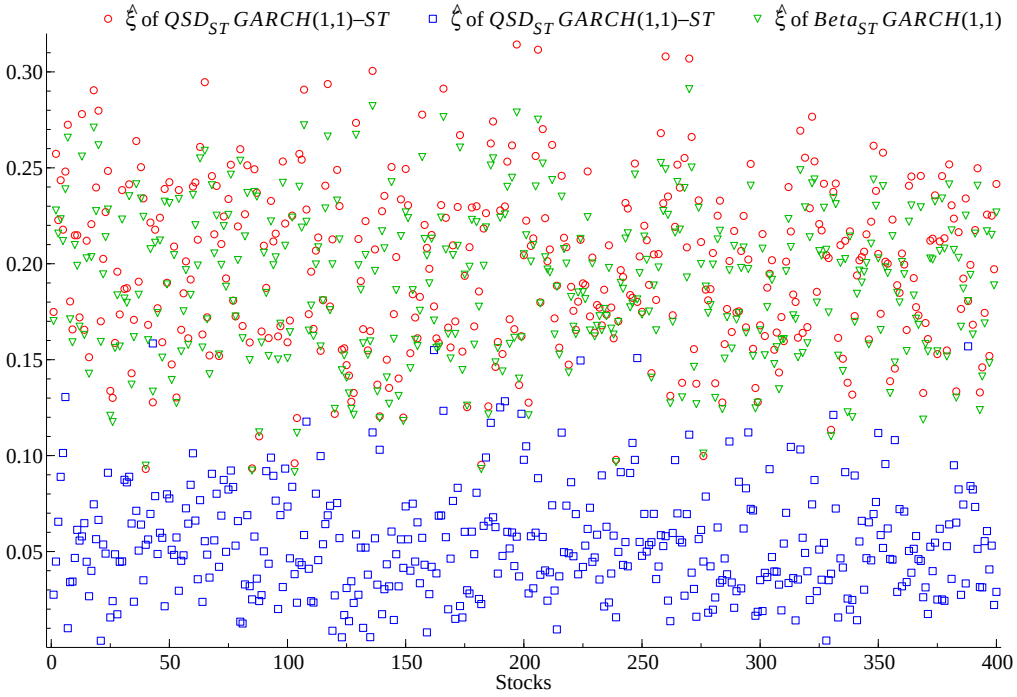
where τ_t is defined similarly to κ_t. A value of $\tau < 0.5$ can therefore produce a leverage effect, irrespective of the value taken by κ.

A $GARCH(1,1) - ST$, a $\beta_{ST}\ GARCH(1,1)$ and a $QSD_{ST}\ GARCH(1,1) - ST$ are estimated on the same 400 stocks. To help visualize the difference between the last two models, their MLEs of the tail parameters ξ and ζ are plotted in Panel (a) of Figure 5, while the asymmetry parameters κ and τ are plotted in Panel (b). Two conclusions emerge from this graph. First, the estimated values of ξ and κ of the $\beta_{ST}\ GARCH$ are systematically very close to those of the $QSD_{ST}\ GARCH - ST$. Second, in most cases, $\hat{\xi}$ and $\hat{\zeta}$ are quite different. More importantly, $\hat{\tau}$ is far below 0.5 in all cases, while $\hat{\kappa}$ is on average close to 0.5 (although very often significantly different from 0.5). This suggests that the additional flexibility of the QSD_{ST} $GARCH - ST$ is needed in all cases.

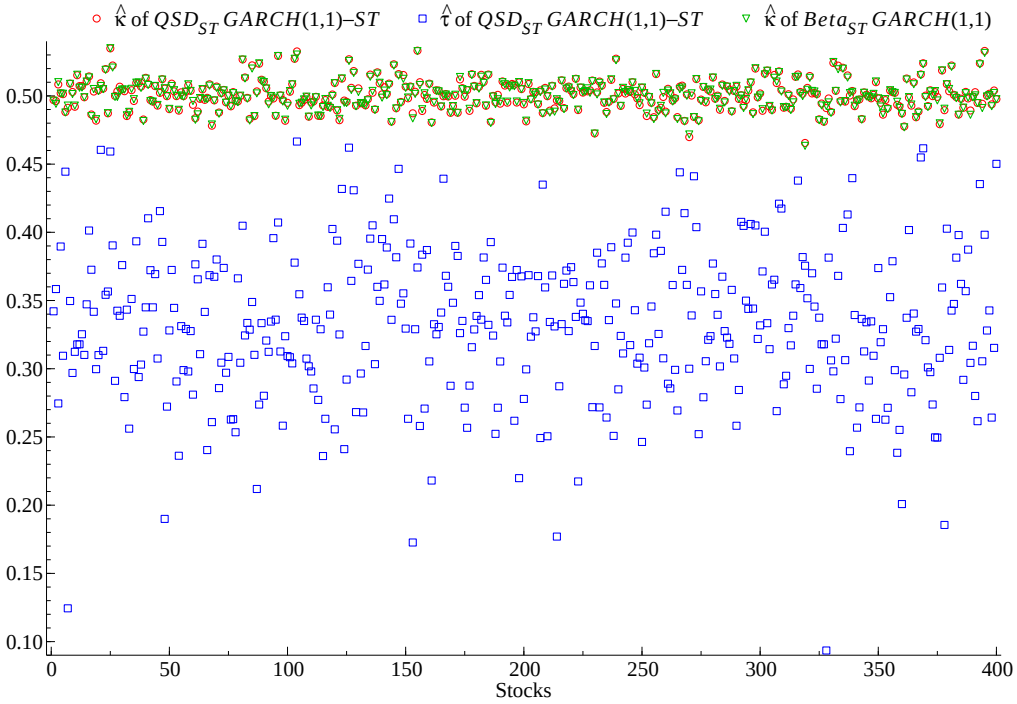
[7] Harvey and Lange (2016) rely on a non-standardized ST density because a study of the NIC of a $Beta_{ST}\ GARCH(1,1)$ derived from a standardized ST distribution shows that small shocks can lead to negative values of f_t even when the usual positivity constraints are imposed. We therefore also consider a non-standardized ST distribution for the sake of comparison.

Figure 5: Estimates of the shape parameters of the QSD_{ST} $GARCH-ST$ and β_{ST} $GARCH$ models

(a) $\hat{\xi}$ and $\hat{\zeta}$



(b) $\hat{\kappa}$ and $\hat{\tau}$



An NIC evaluated at the average of the estimated parameters of the three models is also plotted in Figure 6. For the β_{ST} $GARCH(1,1)$ model, the averages of $1/\zeta$ and κ are respectively 5.2 and 0.5, while for the QSD_{ST} $GARCH(1,1)-ST$ model, they are respectively equal to 18.9 and 0.32. This figure clearly suggests that the QSD_{ST} $GARCH(1,1)-ST$ model bounds the effect of large shocks while allowing negative shocks to have a deeper impact on future values of the volatility.

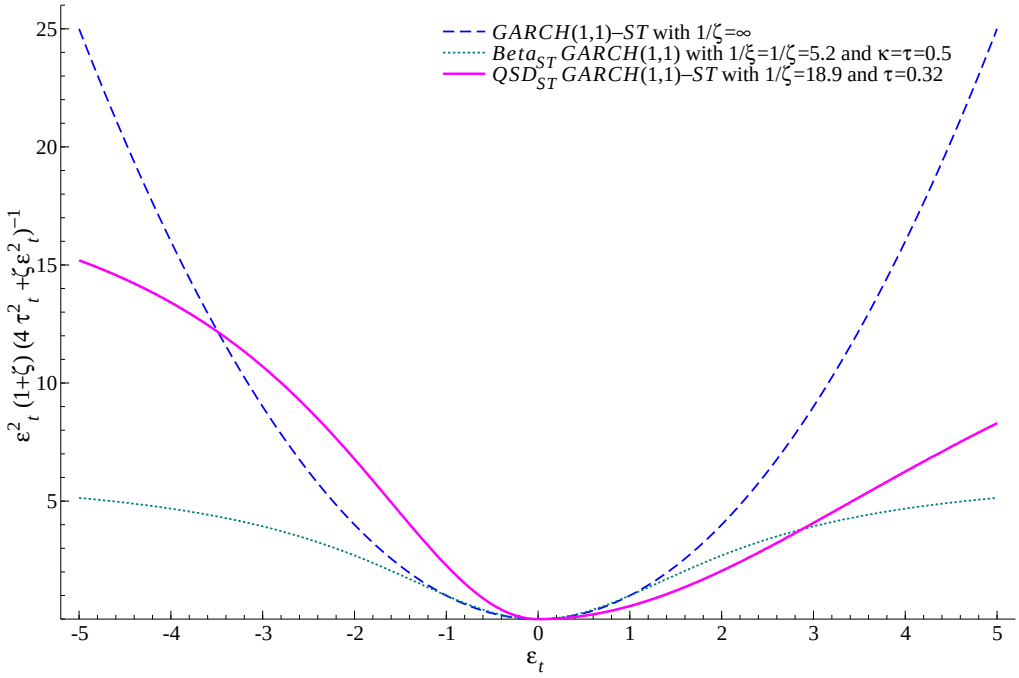


Figure 6: News impact curve for the averaged values (over the 400 stocks) of the MLEs of the $GARCH(1,1)-ST$, β_{ST} $GARCH(1,1)$ and QSD_{ST} $GARCH(1,1)-ST$ models.

To complement the above analysis, the rejection frequencies of the LRTs of the null hypotheses $H_0 : \xi = \zeta; \kappa = \tau$ (i.e., QSD_{ST} $GARCH(1,1) - ST = \beta_{ST}$ $GARCH(1,1)$) and $H_0 : \zeta = 0$ (i.e., QSD_{ST} $GARCH(1,1)-ST = GARCH(1,1)-ST$) are reported in the first two columns of Table 6, while the rejection frequencies of the PIT test presented in the previous section are reported in the last three columns (under PIT test) for the three models considered in this section.

The results suggest that the restriction in the β_{ST} $GARCH$ that imposes $\xi = \zeta$ and $\kappa = \tau$ is rejected in all cases, while the GARCH restriction that imposes $\zeta = 0$ (i.e., no bounding on the effect of large shocks) is rejected in 98.7% of the cases. Interestingly, although the skewed Student's t distribution is more flexible than the symmetric Student's t distribution, the β_{ST} $GARCH$ is rejected more often than

the symmetric Student's t density considered in Section 5.1 (i.e., in 72.2% of the cases compared to 46% for the symmetric Student's t distribution). Importantly, the QSD_{ST} $GARCH-ST$ model is rejected in only 11.3% of the cases (recall that we expect already 5% of rejections because of the multiple tests and the type I errors).

Table 6: Rejection frequencies (at 5%) of two LRTs and a goodness-of-fit test for the models with skewed Student's t densities

LRT		PIT test		
$H_0: \xi = \zeta$ $\kappa = \tau$	$H_0: \zeta = 0$	$GARCH-ST$	β_{ST} $GARCH$	QSD_{ST} $GARCH-ST$
100.0	98.8	33.1	72.2	11.3

Note: see Table 4.

6 Conclusion

SD models have received considerable attention in the time series literature. In this paper, we relax a restriction imposed by this general class of models. Specifically, we break the strict link between the shape of the conditional distribution of y_t and the loss function used to design the updating equation of f_t. We thus arrive at a more general family of models called QSD. This class of models allows researchers to design parameter updating equations that are guided by a multitude of statistical loss functions beyond the log-likelihood function.

We study the statistical properties of the QSD 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 SD models, linking f_t to $p_t(y_t|f_t, \theta)$.

We study in detail the QSD_T $GARCH(1,1)-T$ model, a volatility model extending the β_T $GARCH(1,1)$ 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 (compared with the β_T $GARCH(1,1)$) is found to be significant at the 5% significance level using a standard LRT in more than 50% of the cases (out of 400 stocks). However, we find that this model is rejected in more than 40% of the cases when using the PIT test of Diebold, Gunther and Tay (1998) extended to account for the estimation uncertainty.

Finally, we also propose a volatility model derived from the ST density of Zhu and Galbraith (2010), denoted as QSD_{ST} $GARCH(1,1)-ST$. Unlike pure SD models, this volatility model allows us to introduce the leverage effect in a QSD

model even when the series are left-skewed. We show that, according to the PIT test, this model captures the most important features of more than 88% of the stocks considered in the empirical application.

References

[1] Banulescu-Radu, D., Hansen, P.R., Huang, Z. and Matei, M. (2018) Volatility During the Financial Crisis Through the Lens of High Frequency Data: A Realized GARCH Approach. *SSRN: http://dx.doi.org/10.2139/ssrn.3178890.*

[2] 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.

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

[4] Blasques, F., Francq, C., and Laurent, S. (2020) A New Class of Robust Observation-Driven Models. *Tinbergen institute discussion paper No. 20-073/III*

[5] Blasques F., Brummelen J., Koopman S.J., and Lucas, A., (2020) Maximum Likelihood Estimation of Score-Driven Time Series Models. *Tinbergen institute discussion paper.*

[6] 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.

[7] Bollerslev, T. (1987): A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return, *Review of Economics and Statistics*, 69, 542–547.

[8] Bougerol, P. (1993) Kalman filtering with random coefficients and contractions. *SIAM Journal on Control and Optimization* 31, 942–959.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] Diebold, F.X., Gunther, T. and Tay, A. (1998) Evaluating Density Forecasts, with Applications to Financial Risk Management, *International Economic Review*, 39, 863–883.

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

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

[19] Fernández, C. and M. Steel (1998) On Bayesian Modelling of Fat Tails and Skewness, *Journal of the American Statistical Association* 93, 359–371.

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

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

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

[23] Giot, P. and Laurent, S. (2003) Value-at-Risk for Long and Short Trading Positions. *Journal of Applied Econometrics*, 18, 641-663.

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

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

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

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

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

[29] Hansen, B. (1994) Autoregressive Conditional Density Estimation. *International Economic Review*, 35, 705–730.

[30] Hansen, P.R. and Lunde, A. (2005) A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics* 20/7, 873–889.

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

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

[33] Harvey, A. C. and Lange, R.-J. (2016) Volatility Modeling with a Generalized t Distribution. *Journal of Time Series Analysis* 38, 175–190.

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

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

[36] 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.

[37] Lambert, P., Laurent S. and Veredas, D. (2012) Testing Conditional Asymmetry. A Residual-Based Approach. *Journal of Economics Dynamics and Control*, 36/8, 1129–1247.

[38] 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.

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

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

[41] Pierce, D.A. (1982) The Asymptotic Effect of Substituting Estimators for Parameters in Certain Types of Statistics. The Annals of Statistics, 10, 475–478.

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

[43] Straumann, D., and Mikosch, T. (2006) Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach. *The Annals of Statistics* 34, 2449–2495.

[44] Tse, Y.K. (2002) Residual-based diagnostics for conditional heteroscedasticity models. *The Econometrics Journal* 5, 358–373.

[45] van der Vaart, A.W. (2000) Asymptotic statistics. Cambridge university press.

[46] 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.

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

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

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

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

[51] Zhu, D. and Galbraith, J.W. (2010). A Generalized Asymmetric Student-t Distribution with Application to Financial Econometrics. *Journal of Econometrics* 157, 297–305.

A Local ρ-improvement of the ψ-filter

Compared to score models, the additional flexibility of the class of ψ-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 step $f_{t+1} - f_t|$ is small, 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 δ-neighbourhood of y_t. The QSD model allows for a generalization of this idea whereby ψ_t is a derivative of some local differentiable distance (metric) function $D_{(y_t,\delta)}$,

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

As illustrated below, the distance function $D_{(y_t,\delta)}$ is implicitly defined by the loss criterion used to build the updating equation of the QSD model.

Proposition 1 highlights the trivial but relevant notion that the ψ-update can be used as a Newton-type algorithm when the ρ-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 (ω, β) sufficiently close to the values $(0, 1)$. For completeness, a short justification for Proposition 1 is given in Appendix B. Naturally, since QSD models nest score models (in particular, when the log likelihood is used as a loss function for the update), there are a range of settings under which where these updates are equivalent. Definition 1 introduces the notion of *ψ-equivalent update* as being an update that always steps in the same direction as the ψ-update.

Definition 1. (ψ-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 ψ-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 ρ-improvement of ψ-updates) *Let ρ be continuously differentiable in f_t. Then, there exists a $\delta_f > 0$, and (ω, β) in a neighborhood of $(0, 1)$ such that*

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

if and only if f_t is ψ-equivalent. Additionally, let ρ_η and η be such that

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

with ρ_η continuously differentiable in $\eta(y)$. Then, for $\eta(y_{t+1})$ sufficiently close to $\eta(y_t)$, we have

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

if and only if f_t is ψ-equivalent.

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

EXAMPLE 5. (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 ψ-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 6. (Volatility model) *The same reasoning applies to a volatility model. Here, one might set $\eta(y_t) = y_t^2$ so that the ψ-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.

B 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 ψ 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 ω, $\beta - 1$ 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 $\eta(y_{t+1}) - \eta(y_t)$, ω, $\beta - 1$ and $f_{t+1} - f_t$ small enough.

Proof of Lemma 1

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

$$f_{t+1}^{(n)} = \varphi(z_t, f_t^{(n-1)}) \tag{46}$$

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

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

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 (46), 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

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

It follows that

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

and thus

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

For $n < m$, we then have

$$\begin{aligned}\left|f_{t+1}^{(m)} - f_{t+1}^{(n)}\right| &\leq \sum_{k=0}^{m-n-1}\left|f_{t+1}^{(m-k)} - f_{t+1}^{(m-k-1)}\right| \\ &\leq \sum_{k=0}^{m-n-1}\Lambda_t\Lambda_{t-1}\cdots\Lambda_{t-m+k+2}\left|\varphi(z_{t-m+k+1}, f^0) - f^0\right| \\ &\leq \sum_{j=n}^{\infty}\Lambda_t\Lambda_{t-1}\cdots\Lambda_{t-j+1}\left|\varphi(z_{t-j}, f^0) - f^0\right|. \qquad (47)\end{aligned}$$

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

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

The process (Λ_t) being stationary and ergodic, (ii) then entails

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

By the Cauchy rule, the right-hand side of (47) 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

$$|f_{t+1} - f_{t+1}^*| \leq \Lambda_t\Lambda_{t-1}\cdots\Lambda_{t-N}|f_{t-N} - f_{t-N}^*|. \qquad (48)$$

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 (48) 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 (47) 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 Λ_t are independent, $\mathbb{E} \log \Lambda_t < 0$, $\mathbb{E} \left|\varphi(z_t, f^0) - f^0\right|^r < \infty$ and $\mathbb{E}\Lambda_t^r < \infty$. 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 $\mathbb{E}\Lambda_t^s < 1$, and thus $\mathbb{E} \left|f_{t+1} - f^0\right|^s < \infty$ 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 $\mathbb{E} \ln^+ |\varsigma(y_t, X_t, f^0) - f^0| < \infty$ and $\mathbb{E} \log \Lambda_t(\theta) < 0$ 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 ϱ such that

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

so that

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

and the conclusion follows.

Proof of Lemma 4

Let θ 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) $\mathbb{E}\log^+\|A_1\| < \infty$ and $\mathbb{E}\log^+|b_1| < \infty$. 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

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

which is implied by (ii) of Lemma 3.

In the sequel, ϱ denotes a generic constant of the interval $(0,1)$, and K denotes a positive constant or a random variable measurable with respect to $\{z_t, t \leq 0\}$. 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 $|f_t(\theta) - \widehat{f}_t(\theta)| \leq K\varrho^t$. Dropping "$(\theta)$" in the notations, other similar Taylor expansions thus show that

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

where $\varrho_t = u_t\varrho^t$ with $\mathbb{E}\log^+ u_t < \infty$, 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, $\lim_{\tau\to 0}\mathbb{E}\log(|b_1|+\tau) = \gamma < 0$. Therefore, there exists $\tau > 0$ such that

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

and then

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

We thus have

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

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

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

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

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

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

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

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

$$\liminf_{T\to\infty} \inf_{\theta^*\in V(\theta)\cap\Theta} \left\|\widehat{G}_T(\theta^*)\right\| > 0, \quad \text{a.s.} \tag{49}$$

and that for any neighbourhood $V(\theta_0)$ of θ_0

$$\limsup_{T\to\infty} \inf_{\theta^*\in V(\theta_0)\cap\Theta} \left\|\widehat{G}_T(\theta^*)\right\| = 0, \quad \text{a.s.} \tag{50}$$

Let

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

For any neighbourhood $V(\theta)$ of θ, 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 $\mathbb{E}\log^+ u_t < \infty$ 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 θ 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

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

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 (49).

To show (50), 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,

$$\sup_{\theta\in\Theta} \left\| \frac{\partial G_T(\theta)}{\partial\theta} - \frac{\partial \widehat{G}_T(\theta)}{\partial\theta} \right\| = O(T^{-1}) \quad \text{a.s.} \tag{51}$$

Now note that the ergodic theorem and $\mathbb{E}_{t-1}h_t(\theta_0) = 0$ 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 θ_0, using (51), the ergodic theorem and the continuity of $\partial G(\theta)/\partial\theta^\top$. The contraction f_T thus admits a unique fixed-point θ_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 θ_0^* is the identifiably unique maximizer of the limit criterion.

The uniform convergence of the criterion follows from

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

where

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

by the uniform invertibility obtained in Lemma 3, and

$$\sup_{\theta\in\Theta}\left|\frac{1}{T}\sum_{t=2}^{T}\ell_t(\theta)-\mathbb{E}\ell_t(\theta)\right| \overset{as}{\to} 0 \quad \text{as} \quad T\to\infty$$

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 $\mathbb{E}\ell_t(\theta)<\mathbb{E}\ell_t(\theta_0^*)$ for every $\theta\neq\theta_0^*$, $\theta\in\Theta$, the continuity of the limit criterion and the compactness of Θ (Potscher and Prucha, 1997). The interpretation of θ_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 ϵ_t.

Proof of Lemma 6

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

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

with

$$a = |\alpha| \sup_{y,X,f} \left| \frac{\partial \psi(y, X, f, \theta_0)}{\partial y} \right| < \infty \quad \text{and} \quad b = \sup_{y,X,f} \left| \alpha_0 \frac{\partial \psi(y, X, f, \theta_0)}{\partial f} + \beta_0 \right| < 1.$$

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 $\{e_t\}_{t\in\mathbb{Z}}$.

Let $\widehat{f}_t = \widehat{f}_t(\theta_0)$ and $\widehat{f}'_t = \widehat{f}'_t(\theta_0)$. To show the second claim, first note that (7) and Condition *(iii)* entail

$$\sup_t \left| \widehat{f}_t \right| \leq \frac{1}{1 - |\beta_0|} \left\{ |\omega_0| + |\alpha_0| \sup_{y,X,f} \left| \psi(y, X, f, \theta_0) \right| \right\} + |\beta_0|^{t-1} \left| \widehat{f}_1 \right| < \infty.$$

In view of (11), we also have $\left\| \widehat{f}'_{t+1} \right\| \leq \overline{a} + b \left\| \widehat{f}'_t \right\|$ for all $t \geq 1$, where

$$\begin{aligned} \overline{a} = & \left\| \frac{\partial \omega_0}{\partial \theta} \right\| + \sup_{y,X,f} \left| \psi(y, X, f, \theta_0) \right| \left\| \frac{\partial \alpha_0}{\partial \theta} \right\| \\ & + |\alpha_0| \sup_{y,X,f} \left| \frac{\partial \psi(y, X, f, \theta_0)}{\partial \theta} \right| + \sup_t \left| \widehat{f}_t \right| \left\| \frac{\partial \beta_0}{\partial \theta} \right\| < \infty, \end{aligned}$$

using Condition *(iv)*. Therefore we have shown that $\sup_t \left\| \widehat{f}_t \right\| + \left\| \widehat{f}'_t \right\| \leq M < \infty$.

Now, noting that $\widehat{f}'_{t+1} = \Psi(y_t, X_t, \widehat{f}_t, \widehat{f}'_t)$, let $\widehat{f}'^*_{t+1} = \Psi(y^*_t, X_t, \widehat{f}^*_t, \widehat{f}'^*_t)$. The derivative filter satisfies

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

where, by Conditions *(i)-(ii)* and *(v)-(viii)*,

$$\begin{aligned} a_y = & \left\| \frac{\partial \alpha_0}{\partial \theta} \right\| \sup_{y,X,f} \left| \frac{\partial \psi(y, X, f, \theta_0)}{\partial y} \right| + |\alpha_0| \sup_{y,X,f} \left\| \frac{\partial^2 \psi(y, X, f, \theta_0)}{\partial \theta \partial y} \right\| \\ & + |\alpha_0| \sup_{y,X,f} \left| \frac{\partial^2 \psi(y, X, f, \theta_0)}{\partial f \partial y} \right| M < \infty, \\ a_f = & \left\| \frac{\partial \alpha_0}{\partial \theta} \right\| \sup_{y,X,f} \left| \frac{\partial \psi(y, X, f, \theta_0)}{\partial f} \right| + |\alpha_0| \sup_{y,X,f} \left\| \frac{\partial^2 \psi(y, X, f, \theta_0)}{\partial \theta \partial f} \right\| + \left\| \frac{\partial \beta_0}{\partial \theta} \right\| \\ & + |\alpha_0| \sup_{y,X,f} \left| \frac{\partial^2 \psi(y, X, f, \theta_0)}{\partial f^2} \right| M < \infty. \end{aligned}$$

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 ℓ_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

$$\sqrt{T}(\widetilde{\theta}_T - \theta_0^*) = -\Big(\boldsymbol{\ell}''_T(\theta_T^*)\Big)^{-1}\sqrt{T}\boldsymbol{\ell}'_T(\theta_0^*). \tag{52}$$

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 ϕ-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

$$= \lim \sqrt{T}\boldsymbol{\ell}'_T(\theta_0^*) \xrightarrow{d} N(0, V(\theta_0^*)) \quad \text{as } T\to\infty, \tag{53}$$

$$\text{where} \quad V(\theta_0^*) = \lim_{T\to\infty} \mathrm{Var}(\sqrt{T}\ell'_T(\theta_0^*)).$$

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 Θ 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

$$\boldsymbol{\ell}''_T(\theta_T^*) = \frac{1}{T}\sum_{t=2}^{T}\ell''_t(\theta_T^*) \xrightarrow{as} \mathbb{E}\ell''_t(\theta_0^*) \quad \text{as} \quad T\to\infty, \tag{54}$$

since $\theta_T^* \xrightarrow{as} \theta_0^*$. The asymptotic distribution of $\widetilde{\theta}_T$ is obtained by combining (52), (53) and (54), 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

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

where

$$\mathcal{J}(\theta_0^*) = \lim_{T\to\infty} T^{-1}\mathbb{E}\Big(\sum_{t=1}^{T} \widehat{\ell}_t'(\theta_0)\Big)\Big(\sum_{t=1}^{T} \widehat{\ell}_t'(\theta_0)^\top\Big).$$

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

$$\sqrt{T}\Big(\widehat{\boldsymbol{\ell}}_T'(\widehat{\theta}_T) - \boldsymbol{\ell}_T'(\widehat{\theta}_T)\Big) = \boldsymbol{\ell}_T''(\theta_T^*)\sqrt{T}(\widetilde{\theta}_T - \widehat{\theta}_T). \tag{55}$$

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

$$\boldsymbol{\ell}_T''(\theta_T^*) \xrightarrow{as} \mathbb{E}\ell_t''(\theta_0^*). \tag{56}$$

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,

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

which in turn implies that

$$\sqrt{T}\big|\widehat{\boldsymbol{\ell}}_T'(\widehat{\theta}_T) - \boldsymbol{\ell}_T'(\widehat{\theta}_T)\big| \xrightarrow{as} 0 \quad \text{as } T\to\infty. \tag{57}$$

Combining (55), (56) and (57), 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 θ_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 ϵ_t.

Proof of Theorem 4

Recall that the constrained estimator $(\widehat{\theta}_T^{p_0})$ is such that $(\widehat{\theta}_T^{p_0}, \widehat{\lambda}_T)$ 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

$$R\widehat{\theta}_T^{p_0} - \boldsymbol{r} = 0, \qquad R^\top \widehat{\lambda}_T = \frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial \theta}. \tag{58}$$

First recall that from Corollary 2

$$R\sqrt{T}(\widehat{\theta}_T - \theta_0) \xrightarrow{d} N\big(0, R\mathcal{I}^{-1}R^\top\big), \tag{59}$$

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 $\widehat{\theta}_T^{p_0} \to \theta_0$ a.s. under H_0. A Taylor expansion then entails

$$\sqrt{T}\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial \theta} + o_P(1) = \sqrt{T}\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T)}{\partial \theta} - \mathcal{I}\sqrt{T}(\widehat{\theta}_T^{p_0} - \widehat{\theta}_T) = -\mathcal{I}\sqrt{T}(\widehat{\theta}_T^{p_0} - \widehat{\theta}_T). \tag{60}$$

Using (58), it follows that under H_0

$$R\sqrt{T}(\widehat{\theta}_T - \theta_0) \;\; = \;\; R\sqrt{T}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0}) = R\mathcal{I}^{-1}R^\top\sqrt{T}\widehat{\lambda}_T + o_P(1). \tag{61}$$

Using (59) we then obtain

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

and thus, using again (58),

$$T\widehat{\lambda}_T^\top R\mathcal{I}^{-1}R^\top\widehat{\lambda}_T = T\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial \theta^\top}\mathcal{I}^{-1}\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial \theta} \xrightarrow{d} \chi_r^2. \tag{62}$$

The first convergence follows.

To derive the asymptotic distribution of LR_T we use the usual argument which involves expanding $\widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T)$ around $\widehat{\theta}_T^{p_0}$ to obtain

$$\begin{aligned}
\mathrm{LR}_T &:= 2T\Big(\widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T) - \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})\Big) \\
&= 2T\Big(\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial \theta^\top}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0}) - \frac{1}{2}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0})^\top\mathcal{I}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0})\Big) + o_P(1) \\
&= \sqrt{T}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0})^\top\sqrt{T}\frac{\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})}{\partial \theta} + o_P(1) \\
&= \sqrt{T}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0})^\top\sqrt{T}R^\top\widehat{\lambda}_T + o_P(1) \\
&= T\widehat{\lambda}_T^\top R\mathcal{I}^{-1}R^\top\widehat{\lambda}_T + o_P(1) \xrightarrow{d} \chi_q^2
\end{aligned}$$

noting $\partial \widehat{\boldsymbol{\ell}}_T(\widehat{\theta}_T^{p_0})/\partial\theta = 0$ and using (58), (60), (61) and (62).