Online supplement to the paper 'Quasi score-driven models'*

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

December 24, 2021

Abstract

Online supplement to the article titled "Quasi score-driven models". It offers detailed proofs of the propositions, corollaries, theorems and lemmas presented in the main core of the paper as well as details on the local ρ-improvement of the ψ-filter and some additional Monte Carlo Simulation results on the PIT test.

JEL Classification: C13, C32, C53, C58.
Keywords: score-driven models, GARCH, fat-tails, asymmetry, QLE, QMLE.

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

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

*The authors gratefully acknowledge the editor Torben Andersen, the associate editor and two anonymous referees for insightful comments and suggestions, 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. Sébastien acknowledges research support by the French National Research Agency Grant ANR-17-EURE-0020, and by the Excellence Initiative of Aix-Marseille University - A*MIDEX. Christian and Sébastien also acknowledge research support by the French National Research Agency Grant ANR-21-CE26-0007-01. Christian also thanks the ECODEC labex.

[†]VU University Amsterdam, Tinbergen Institute, 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

[§]Corresponding author: 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.

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 $KL_{(y_t,\delta)}$ denotes 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 1. (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 2. (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

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

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 (B.1), 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

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

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

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

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

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

By the Cauchy rule, the right-hand side of (B.2) tends almost surely to zero as $n\to\infty$. The existence of a stationary and ergodic solution to (5) follows.

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

$$|f_{t+1}-f_{t+1}^*|\leq\Lambda_t\Lambda_{t-1}\cdots\Lambda_{t-N}|f_{t-N}-f_{t-N}^*|. \tag{B.3}$$

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 (B.3) 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 (B.2), 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 $\mathbb{E}\ln^+\left|\varsigma(y_t,X_t,f^0)-f^0\right|<\infty$ 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

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

We thus have

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

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

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

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

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

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

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

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

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

Proof of Theorem 1

By compactness of 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

$$\liminf_{T\to\infty} \inf_{\theta^*\in V(\theta)\cap\Theta} \left\|\widehat{G}_T(\theta^*)\right\| > 0, \quad \text{a.s.} \tag{B.4}$$

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

$$\limsup_{T\to\infty} \inf_{\theta^*\in V(\theta_0)\cap\Theta} \left\|\widehat{G}_T(\theta^*)\right\| = 0, \quad \text{a.s.} \tag{B.5}$$

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 (16), (17) and (18), we have

$$\sup_{\theta\in\Theta} \|g_t(\theta) - \widehat{g}_t(\theta)\| \leq K\varrho^t u_t, \quad u_t = \sup_{\theta\in\Theta} \left(|y_t|^k + |f_t(\theta)| + 1\right)\left(1 + \left\|f_t'(\theta)\right\|\right).$$

Since $\mathbb{E}\log^+ u_t < \infty$ under the log-moment conditions and $\varrho < 1$, the Cauchy root test (see Lemma 2.1 of Straumann and Mikosch, 2006) 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,

$$\sup_{\theta\in\Theta} \left\|G_T(\theta) - \widehat{G}_T(\theta)\right\| = O(T^{-1}) \quad \text{as} \quad T \to \infty. \tag{B.6}$$

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 (21), the expectation of the right-hand side of the inequality tends to 0 as $m \to \infty$. By (22) 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 (B.4).

To show (B.5), 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.

To show the asymptotic normality, first note that

$$\frac{\partial G_T(\theta)}{\partial \theta^\top} = \frac{1}{T}\sum_{t=t_0+1}^{T} \frac{h_t(\theta)}{\sigma_t^2(\theta)} f_t''(\theta) - \frac{h_t(\theta)}{\sigma_t^4(\theta)} \frac{\partial \sigma_t^2(\theta)}{\partial \theta} f_t'^{\top}(\theta) - \frac{1}{\sigma_t^2(\theta)} f_t'(\theta) f_t'^{\top}(\theta). \tag{B.7}$$

By already given arguments, Lemma 4 and (24) show that

$$\sup_{\theta\in\Theta} \left\| \frac{\partial G_T(\theta)}{\partial \theta^\top} - \frac{\partial \widehat{G}_T(\theta)}{\partial \theta^\top} \right\| = O(T^{-1}) \quad \text{a.s.} \tag{B.8}$$

Now note that (B.7), the ergodic theorem and $\mathbb{E}_{t-1} h_t(\theta_0) = 0$ imply that

$$\dot{G}_T := \partial G_T(\theta_0)/\partial \theta^\top \to -\mathcal{J} \tag{B.9}$$

almost surely as $T \to \infty$. In view (25), 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 (B.8), 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 neighborhood, for which $\widehat{G}_T(\theta_T) = 0_p$. 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 (20), we have $\theta_T = \widehat{\theta}_T$, and thus

$$\widehat{G}_T(\widehat{\theta}_T) = 0_p. \tag{B.10}$$

Doing Taylor expansions of $G_{1T}(\cdot), \ldots, G_{pT}(\cdot)$, where $G_{iT}(\theta)$ denotes the i-th element of $G_T(\theta_T)$, and using (B.6) we have almost surely

$$0_p = \sqrt{T}\widehat{G}_T(\widehat{\theta}_T) = \sqrt{T} G_T(\theta_0) - \mathcal{J}_T \sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) + o(1), \tag{B.11}$$

where the i-th row of $\mathcal{J}_T$ is of the form $-\frac{\partial G_{iT}(\theta_{iT})}{\partial \theta^\top}$ with θ_{iT} a point between $\widehat{\theta}_T$ and θ_0. The consistency of $\widehat{\theta}_T$ entails that, for T large enough

$$\left\| \mathcal{J}_T + \dot{G}_T \right\| \leq \frac{1}{T} \sum_{t=t_0+1}^{T} \sup_{\theta \in V(\theta_0)} \left\| \frac{\partial g_t(\theta)}{\partial \theta^\top} - \frac{\partial g_t(\theta_0)}{\partial \theta^\top} \right\| \tag{B.12}$$

for any neighborhood $V(\theta_0)$ of θ_0. In view of (21) and (23) with $r = 2$, the Hölder inequality entails

$$E \sup_{\theta \in V(\theta_0)} \left\| \frac{h_t(\theta)}{\sigma_t^2(\theta)} f_t''(\theta) \right\| \leq \sqrt{E \sup_{\theta \in V(\theta_0)} \|\eta_t(\theta)\|^2 \, E \sup_{\theta \in V(\theta_0)} \left\| \frac{1}{\sigma_t(\theta)} f_t''(\theta) \right\|^2} < \infty.$$

Similarly

$$E \sup_{\theta \in V(\theta_0)} \left\| \frac{h_t(\theta)}{\sigma_t^4(\theta)} \frac{\partial \sigma_t^2(\theta)}{\partial \theta} f_t'^\top(\theta) \right\| < \infty, \quad E \sup_{\theta \in V(\theta_0)} \left\| \frac{1}{\sigma_t^2(\theta)} f_t'(\theta) f_t'^\top(\theta) \right\| < \infty.$$

In view of (B.7), it follows that, by the ergodic theorem, the right-hand side of (B.12) tends almost surely to

$$E \sup_{\theta \in V(\theta_0)} \left\| \frac{\partial g_t(\theta)}{\partial \theta^\top} - \frac{\partial g_t(\theta_0)}{\partial \theta^\top} \right\|.$$

By the dominated convergence theorem this expectation is arbitrarily small when $V(\theta_0)$ is small. In view of (B.9), it follows that $\mathcal{J}_T \to \mathcal{J}$. The asymptotic distribution of $\widehat{\theta}_T$ is then obtained from the CLT for martingale differences. The consistency of $\widehat{\mathcal{J}}_T$ is shown exactly as that of $\mathcal{J}_T$.

Proof of Corollary 1

The convergence of the Wald statistic is obvious since Theorem 1 entails that,

$$\text{under } H_0, \quad \sqrt{T} K \widehat{\theta}_T = \sqrt{T} K (\widehat{\theta}_T - \theta_0) \xrightarrow{d} N(0, K \mathcal{J}^{-1} K^\top).$$

By the arguments used to show (B.11), we have

$$\sqrt{T} \widehat{G}_T(\widehat{\theta}_{T|2}) = \sqrt{T} G_T(\theta_0) - \mathcal{J}_T \sqrt{T} \left(\widehat{\theta}_{T|2} - \theta_0 \right) + o(1) \quad \text{a.s.} \tag{B.13}$$

Note that $\widehat{G}_T(\widehat{\theta}_{T|2})$ is of the form

$$\widehat{G}_T(\widehat{\theta}_{T|2}) = \begin{pmatrix} 0_{p_1} \\ \widehat{G}_T^{(2)}(\widehat{\theta}_T^{(1)}) \end{pmatrix} = K^\top \widehat{G}_T^{(2)}(\widehat{\theta}_T^{(1)}). \tag{B.14}$$

Under H_0, we also have

$$\left(\widehat{\theta}_{T|2} - \theta_0 \right) = \widetilde{K}^\top \left(\widehat{\theta}_T^{(1)} - \theta_0^{(1)} \right).$$

The first p_1 rows of (B.13) thus yield

$$0_{p_1} = \sqrt{T}\widetilde{K}G_T(\theta_0) - \widetilde{K}\mathcal{J}_T\widetilde{K}^\top\sqrt{T}\left(\widehat{\theta}_T^{(1)} - \theta_0^{(1)}\right) + o(1) \quad \text{a.s.}$$

and the last p_2 rows yield

$$\sqrt{T}\widehat{G}_T^{(2)}(\widehat{\theta}_T^{(1)}) = \sqrt{T}KG_T(\theta_0) - K\mathcal{J}_T\widetilde{K}^\top\sqrt{T}\left(\widehat{\theta}_T^{(1)} - \theta_0^{(1)}\right) + o(1) \quad \text{a.s.}$$

Letting

$$\mathcal{J} = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}, \quad J_{11} = \widetilde{K}\mathcal{J}\widetilde{K}^\top, \quad J_{22} = K\mathcal{J}K^\top, \quad J_{12} = \widetilde{K}\mathcal{J}K^\top$$

and

$$G = \begin{pmatrix} G^{(1)} \\ G^{(2)} \end{pmatrix} \sim N\left(0_p, \mathcal{J}\right), \quad G^{(1)} = \widetilde{K}G, \quad G^{(2)} = KG,$$

we obtain

$$\sqrt{T}\widehat{G}_T^{(2)}(\widehat{\theta}_T^{(1)}) \xrightarrow{d} G^{(2)} - J_{21}J_{11}^{-1}G^{(1)} \sim N\left(0_{p_2}, J_{22} - J_{21}J_{11}^{-1}J_{12}\right).$$

By (B.14), the previous result and the well-known block matrix inversion formula $K\mathcal{J}^{-1}K^\top = (J_{22} - J_{21}J_{11}^{-1}J_{12})^{-1}$, we obtain

$$R_T \;=\; T\widehat{G}_T^{(2)\top}(\widehat{\theta}_T^{(1)})K\widehat{\mathcal{J}}_{T|2}^{-1}K^\top\widehat{G}_T^{(2)}(\widehat{\theta}_T^{(1)}) \xrightarrow{d} \chi^2_{p_2}.$$

Proof of Theorem 2

The desired result follows from the classical consistency argument found e.g. in White (1994, Theorem 3.4) or Pötscher 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

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

by the uniform invertibility obtained in Lemma 3, and

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

by application of Rao's (1962) uniform law of large numbers. The identifiable uniqueness of $\theta_0^*\in\Theta$ is implied by the uniqueness assumption $\mathbb{E}\ell_t(\theta)<\mathbb{E}\ell_t(\theta_0^*)$ for every $\theta\neq\theta_0^*$, $\theta\in\Theta$, the continuity of the limit criterion and the compactness of Θ (Pötscher 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 2

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

Without loss of generality, assume $s\in(0,1]$. Using the mean value theorem and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\left(\sum_{t=1}^{\infty}\sup_{\theta\in\Theta}\left\|\widehat{\ell}_t'(\theta)-\ell_t'(\theta)\right\|\right)^s \leq \sum_{t=1}^{\infty}\sup_{\theta\in\Theta}\left\|\widehat{\ell}_t'(\theta)-\ell_t'(\theta)\right\|^s\\
\leq&\sum_{t=1}^{\infty}\sup_{\theta\in\Theta}\sup_{f,f'}\left\|\frac{\partial\ell'(y_t,f,f',\theta)}{\partial f}\right\|^s\left|\widehat{f}_t(\theta)-f_t(\theta)\right|^s\\
&\quad+\sup_{f,f'}\left\|\frac{\partial\ell'(y_t,f,f',\theta)}{\partial f'}\right\|^s\left\|\widehat{f}_t'(\theta)-f_t'(\theta)\right\|^s\\
\leq&K\sum_{t=1}^{\infty}\sup_{\theta\in\Theta}\left(\sup_{f,f'}\left\|\frac{\partial\ell'(y_t,f,f',\theta)}{\partial f}\right\|^s+\sup_{f,f'}\left\|\frac{\partial\ell'(y_t,f,f',\theta)}{\partial f'}\right\|^s\right)\varrho^{st} \quad \text{a.s.}
\end{aligned}$$

The expectation of the previous sum is finite, and thus it is finite almost surely. The conclusion follows.

Proof of Lemma 7

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 Pötscher 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 Pötscher 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 Pötscher and Prucha (1997).

Proof of Theorem 3

For convenience, we adopt the following notation

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

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

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

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

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

We use the usual mean-value theorem expansion

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

to obtain

$$\sqrt{T}(\widetilde{\theta}_T - \theta_0^*) = -\Big(\boldsymbol{l}''_T(\theta_T^*)\Big)^{-1}\sqrt{T}\boldsymbol{l}'_T(\theta_0^*). \tag{B.15}$$

By Lemma 7, 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 $\mathbb{E}|\ell'(y_t, f_t, \theta_0)|^r < \infty$, we can thus appeal to the central limit theorem for near epoch dependent sequences in Pötscher and Prucha (1997, Theorem 10.2) to show that

$$= \lim \sqrt{T}\boldsymbol{l}'_T(\theta_0^*) \xrightarrow{d} N(0, V(\theta_0^*)) \quad \text{as } T\to\infty. \tag{B.16}$$

Note that unlike in Pötscher and Prucha (1997, Theorem 10.2), the score here is also strictly stationary and ergodic. This implies that the limit asymptotic variance $V(\theta_0^*)$ converges as the sums of the covariance terms are ensured to converge (see Theorem 13.12 in Davidson, 1994).

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., $\sup_{\theta\in\Theta}\|\boldsymbol{l}''_T(\theta) - \mathbb{E}\ell''_t(\theta)\| \xrightarrow{as} 0$, which implies

$$\boldsymbol{l}''_T(\theta_T^*) = \frac{1}{T}\sum_{t=2}^{T}\ell''_t(\theta_T^*) \xrightarrow{as} \mathbb{E}\ell''_t(\theta_0^*) \quad \text{as } T\to\infty, \tag{B.17}$$

since $\theta_T^* \xrightarrow{as} \theta_0^*$. The asymptotic distribution of $\widetilde{\theta}_T$ is obtained by combining (B.15), (B.16) and (B.17), 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

$$\Sigma(\theta_0^*) = \Big(\mathbb{E}\ell_t''(\theta_0^*)\Big)^{-1} V(\theta_0^*) \Big(\mathbb{E}\ell_t''(\theta_0^*)\Big)^{-1},$$

We now expand the score using a mean value theorem

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

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

$$\sqrt{T}\Big(\widehat{\boldsymbol{l}}_T'(\widehat{\theta}_T) - \boldsymbol{l}_T'(\widehat{\theta}_T)\Big) = \boldsymbol{l}_T''(\theta_T^{**})\sqrt{T}(\widetilde{\theta}_T - \widehat{\theta}_T). \tag{B.18}$$

We use again the uniform convergence of the Hessian and the fact that $\tilde{\theta}_T \xrightarrow{a.s.} \theta_0^*$ to conclude that

$$\boldsymbol{l}_T''(\theta_T^{**}) \xrightarrow{as} \mathbb{E}\ell_t''(\theta_0^*). \tag{B.19}$$

Since the assumption of Lemma 6 hold, we have

$$\sqrt{T} \sup_{\theta\in\Theta} \Big\| \widehat{\boldsymbol{l}}_T'(\theta) - \boldsymbol{l}_T'(\theta) \Big\| \xrightarrow{as} 0 \quad \text{as } T \to \infty \tag{B.20}$$

which in turn implies that

$$\sqrt{T} \Big\| \widehat{\boldsymbol{l}}_T'(\widehat{\theta}_T) - \boldsymbol{l}_T'(\widehat{\theta}_T) \Big\| \xrightarrow{as} 0 \quad \text{as } T \to \infty. \tag{B.21}$$

Combining (B.18), (B.19) and (B.21), 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 3

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

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

The first order conditions yield

$$R\widehat{\theta}_T^{p_0} - \boldsymbol{r} = 0, \qquad R^\top \widehat{\lambda}_T = \widehat{\boldsymbol{l}}'_T(\widehat{\theta}_T^{p_0}). \tag{B.22}$$

First recall that from Corollary 3

$$R\sqrt{T}(\widehat{\theta}_T - \theta_0) \xrightarrow{d} N\big(0, R\mathcal{I}^{-1}R^\top\big), \tag{B.23}$$

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

We know that, almost surely, $\widehat{\theta}_T \to \theta_0 \in \mathrm{int}(\Theta)$. Therefore $\widehat{\boldsymbol{l}}'_T(\widehat{\theta}_T) = 0$, at least for T large enough. By (B.20), we thus have

$$\sqrt{T}\widehat{\boldsymbol{l}}'_T(\widehat{\theta}_T^{p_0}) = \sqrt{T}\boldsymbol{l}'_T(\widehat{\theta}_T^{p_0}) + o(1), \quad \sqrt{T}\boldsymbol{l}'_T(\widehat{\theta}_T) = o(1) \quad \text{a.s.}$$

A Taylor expansion of $\sqrt{T}\boldsymbol{l}'_T(\cdot)$ around $\widehat{\theta}_T^{p_0}$ and $\widehat{\theta}_T$ then entails

$$\sqrt{T}\widehat{\boldsymbol{l}}'_T(\widehat{\theta}_T^{p_0}) = -\widehat{\mathcal{I}}\sqrt{T}(\widehat{\theta}_T^{p_0} - \widehat{\theta}_T) + o(1) \quad \text{a.s.} \tag{B.24}$$

where $\widehat{\mathcal{I}} = \frac{1}{T}\sum_{t=2}^{T}\ell''_t(\theta_T)$ for some θ_T between $\widehat{\theta}_T^{p_0}$ and $\widehat{\theta}_T$. It can be shown that $\widehat{\theta}_T^{p_0} \to \theta_0$ a.s. under H_0. Therefore (B.17) entails that $\widehat{\mathcal{I}}$ tends almost surely to $\mathcal{I}$.

Using (B.22) and (B.24), it follows that under H_0

$$R\sqrt{T}(\widehat{\theta}_T - \theta_0) \;\; = \;\; R\sqrt{T}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0}) = R\mathcal{I}^{-1}R^\top\sqrt{T}\widehat{\lambda}_T + o_P(1). \tag{B.25}$$

Using (B.23) we then obtain

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

and thus, using again (B.22),

$$T\widehat{\lambda}_T^\top R\mathcal{I}^{-1}R^\top\widehat{\lambda}_T = T\widehat{\boldsymbol{l}}_T^{'\top}(\widehat{\theta}_T^{p_0})\mathcal{I}^{-1}\widehat{\boldsymbol{l}}'_T(\widehat{\theta}_T^{p_0}) \xrightarrow{d} \chi^2_{p_0}. \tag{B.26}$$

The first convergence follows.

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

$$\begin{aligned}
\mathrm{LR}_T &:= 2T\left\{\widehat{\boldsymbol{l}}_T(\widehat{\theta}_T) - \widehat{\boldsymbol{l}}_T(\widehat{\theta}_T^{p_0})\right\} \\
&= 2T\left\{\widehat{\boldsymbol{l}}_T^{'\top}(\widehat{\theta}_T^{p_0})(\widehat{\theta}_T - \widehat{\theta}_T^{p_0}) - \frac{1}{2}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0})^\top\mathcal{I}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0})\right\} + o_P(1) \\
&= \sqrt{T}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0})^\top\sqrt{T}\widehat{\boldsymbol{l}}'_T(\widehat{\theta}_T^{p_0}) + o_P(1) \\
&= \sqrt{T}(\widehat{\theta}_T - \widehat{\theta}_T^{p_0})^\top\sqrt{T}R^\top\widehat{\lambda}_T + o_P(1) \\
&= T\widehat{\lambda}_T^\top R\mathcal{I}^{-1}R^\top\widehat{\lambda}_T + o_P(1) \xrightarrow{d} \chi^2_{p_0}
\end{aligned}$$

using (B.24), (B.22), (B.25) and (B.26).

C Monte Carlo Simulation results on the PIT test

In this appendix, we study the finite sample properties of the goodness-of-fit test proposed by Diebold, Gunther and Tay (1998) and the correction we propose to account for the estimation error when the model is estimated by ML.

Recall that under the null hypothesis of correct specification, the probability integral transform (PIT) series $\{u_t\}_{t=1}^T \overset{i.i.d.}{\sim} U(0,1)$ so that $\sqrt{T}\sum_{t=1}^T(u_t - T/2) \sim N(0,1/12)$ as $T \to \infty$. Furthermore, according to Corollary 2 in Section 3, the MLEs of the models considered in this paper are asymptotically Gaussian under the assumption of correct specification, so that $\sqrt{T}(\widehat{\theta}_T - \theta_0) \overset{d}{\to} N(0, \mathcal{I}(\theta_0)^{-1})$, where $\mathcal{I}(\theta_0)$ denotes the Fisher information matrix.

We have therefore that both statistics are jointly are asymptotically Gaussian, i.e.,

$$\left(\begin{array}{c} \sqrt{T}\left(\hat{\theta} - \theta_0\right) \\ \sqrt{T}\sum_{t=1}^T(u_t - T/2) \end{array} \right) \sim \mathcal{N}\left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \mathcal{I}(\theta_0)^{-1} & \Sigma_{\theta_0,u} \\ \Sigma'_{\theta_0,u} & 1/12 \end{array} \right) \right),$$

where $\Sigma_{\theta_0,u}$ is the asymptotic covariance which is cumbersome to derive.

However, in empirical applications, the true PIT series is never observed and has to be replaced by $\hat{u}_t(\hat{\theta})$ or $\hat{u}_t$ in short, i.e., the PIT series computed from the cumulative distribution function of the estimated residuals which depends obviously on $\hat{\theta}$. To derive the asymptotic distribution of $\sum_{t=1}^T(\hat{u}_t - T/2)$ we rely on Pierce's (1982) theorem.

Following Pierce (1982), we assume that $\sqrt{T}(\sum_{t=1}^T(\hat{u}_t - T/2))$ can be approximated as $\sqrt{T}(\sum_{t=1}^T(\hat{u}_t - T/2)) = \sqrt{T}(\sum_{t=1}^T(u_t - T/2)) + B\sqrt{T}\left(\hat{\theta} - \theta_0\right) + o_p(1)$, where $\sum_{t=1}^T u_t$ is differentiable in θ and $B = \lim_{T\to\infty} E\left(\frac{\partial \sum_{t=1}^T(u_t - T/2)}{\partial \theta'}\right)$. A simple application of Pierce (1982) allows us to show that under the above conditions, $\sqrt{T}(\sum_{t=1}^T(\hat{u}_t - T/2)) \sim N\left(0, 1/12 - B\mathcal{I}(\theta_0)^{-1}B'\right)$ when $T \to \infty$. In the Monte Carlo simulations and the empirical application, we rely on numerical procedures to compute B although explicit formulas could also be obtained.

This result suggests rejecting the null of correct specification at the 5% nominal level when $|(\sum_{t=1}^T \hat{u}_t - T/2)|/\sqrt{T(1/12 - B\mathcal{I}(\hat{\theta})^{-1}B')} > 1.96$. To illustrate the finite sample properties of this test, we generate $T = 4,000$ observations of the QSD_{ST} $GARCH(1,1) - ST$ presented in Section 5.2 and estimate the true model as well as a β_{ST} $GARCH(1,1)$ and a $GARCH(1,1) - ST$ (also presented in Section 5.2).

For simplicity, we chose the same parameter values as in Table 1 for $\mu, \omega, \varpi, \alpha$ and β. For the shape parameters of the skewed Student's t distributions, we set $1/\xi = 5.2383$ and $\kappa = 0.5$ for the density of the innovations and $1/\zeta = 19.8381$ and $\kappa = 0.3327$ for the updating equation of f_t so that the conditional density has fat-tails and is symmetric while the NIC is bounded and asymmetric. The rejection frequencies (over 1,000 replications) of the PIT test described above at the 5% nominal level applied to the QSD_{ST} $GARCH(1,1) - ST$, β_{ST} $GARCH(1,1)$ and $GARCH(1,1) - ST$ models are respectively 6.11% (size), 87.96% (power) and 45.54% (power). The true model being a QSD_{ST} $GARCH(1,1) - ST$, the first value corresponds to an empirical size while the other two figures correspond to empirical powers. These results suggest that the PIT

test has a decent size and high power to reject the misspecified models for the considered DGP. Importantly, not accounting for the estimation error leads indeed to strong size distortions and in our case to a strong lack of power when the test is applied on misspecified models as the rejection frequencies are 0% (size), 1.4% (power) and 0% (power), respectively for the same three models. This is in line with the findings of Tse (2002) and Lambert, Laurent and Veredas (2012) who have shown that not accounting 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 leads strong size distortions, with empirical sizes close to 0% for a nominal size of 1, 5 or even 10% and that accounting for the estimation error using Pierce's (1982) theorem gives empirical sizes close to the nominal sizes.

References

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

[2] 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.

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

[4] 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.

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

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

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

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

[9] 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.

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

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

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

[13] 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.

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

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