Robust estimation and inference for heavy tailed GARCH

JONATHAN B. HILL[1]

[1]*Department of Economics, University of North Carolina, Chapel Hill, NC 27599-3305, USA.*
E-mail: jbhill@email.unc.edu; url: www.unc.edu/~jbhill

We develop two new estimators for a general class of stationary GARCH models with possibly heavy tailed asymmetrically distributed errors, covering processes with symmetric and asymmetric feedback like GARCH, Asymmetric GARCH, VGARCH and Quadratic GARCH. The first estimator arises from negligibly trimming QML criterion equations according to error extremes. The second imbeds negligibly transformed errors into QML score equations for a Method of Moments estimator. In this case, we exploit a sub-class of redescending transforms that includes tail-trimming and functions popular in the robust estimation literature, and we re-center the transformed errors to minimize small sample bias. The negligible transforms allow both identification of the true parameter and asymptotic normality. We present a consistent estimator of the covariance matrix that permits classic inference without knowledge of the rate of convergence. A simulation study shows both of our estimators trump existing ones for sharpness and approximate normality including QML, Log-LAD, and two types of non-Gaussian QML (Laplace and Power-Law). Finally, we apply the tail-trimmed QML estimator to financial data.

Keywords: GARCH; heavy tails; QML; robust inference; tail trimming

1. Introduction

It is now widely accepted that log-returns of many macroeconomic and financial time series are heavy tailed, exhibit clustering of large values, and are asymmetrically distributed. In broader contexts extremes are encountered in actuarial, meteorological, and telecommunication network data (e.g., Leadbetter *et al.* [38], Embrehts *et al.* [21], Davis [17]), while GARCH-type clustering alone implies higher moments do not exist due to Pareto-like distribution tails (e.g., Basrak *et al.* [4], Liu [42]).

We develop new methods of robust estimation for a general class of GARCH(1, 1) models:

$$y_t = \sigma_t \epsilon_t \qquad \text{with } \sigma_t^2 = g\left(y_{t-1}, \sigma_{t-1}^2, \theta^0\right) \geq 0 \text{ a.s.}, \tag{1}$$

where $g(y, \sigma^2, \theta)$ is a known mapping $g : \mathbb{R} \times [0, \infty) \times \Theta \to [0, \infty)$ and Θ is a compact subset of $\mathbb{R}^q$ for some finite $q \geq 1$. We assume there exists a unique point θ^0 in the interior of Θ such that $\epsilon_t = y_t / \sigma_t$ is i.i.d. with a non-degenerate absolutely continuous distribution with support $(-\infty, \infty)$, $E[\epsilon_t] = 0$ and $E[\epsilon_t^2] = 1$. Further, $\{y_t, \sigma_t\}$ are stationary and geometrically β-mixing. We avoid well known boundary problems by assuming θ^0 lies in the interior of Θ and σ_t^2 has a non-degenerate distribution, hence (1) is a non-trivial GARCH process. In Bollerslev's [7] classic GARCH model $\sigma_t^2 = \omega^0 + \alpha^0 y_{t-1}^2 + \beta^0 \sigma_{t-1}^2$, with $\omega^0 > 0$ and $\alpha^0, \beta^0 \geq 0$ this requires $\alpha^0 + \beta^0 > 0$, cf. Andrews [3] and Francq and Zakoïan [24].

In order to keep technical arguments brief, we assume $\sigma_t^2(\theta) := g(y_{t-1}, \sigma_{t-1}^2(\theta), \theta)$ has properties similar to a non-trivial classic GARCH model: $\sigma_t^2(\theta)$ is twice continuously differentiable, $E[(\sup_{\theta \in \Theta} |\sigma_t^2 / \sigma_t^2(\theta)|)^p] < \infty$ for any $p > 0$, and $\sup_{\theta \in \mathcal{N}_0} \|(\partial/\partial\theta)^i \ln(\sigma_t^2(\theta))\|$ is $L_{2+\iota}$-bounded for tiny $\iota > 0$ and some compact $\mathcal{N}_0 \subseteq \Theta$ containing θ^0, where $\|\cdot\|$ is the matrix norm (cf. Francq and Zakoïan [24]). Similarly, we impose Lipschitz type bounds on g that ensure an iterated approximation $h_0^2(\theta) = \omega$ and $h_t(\theta) = g(y_{t-1}, h_{t-1}(\theta), \theta)$ for $t = 1, 2, \ldots$ satisfies $\sup_{\theta \in \Theta} |h_t(\theta) - \sigma_t^2(\theta)| \xrightarrow{p} 0$ as $t \to \infty$, a key property for feasible estimation (see Nelson [48], Francq and Zakoïan [24], Straumann and Mikosch [60]). The above properties of $\sigma_t^2(\theta)$ cover at least Threshold GARCH with a known threshold, Asymmetric and Nonlinear Asymmetric GARCH, VGARCH, GJR-GARCH, Smooth Transition GARCH, and Quadratic GARCH. Consult Engle and Ng [22], Carrasco and Chen [12], Francq and Zakoïan [24,25] and Meitz and Saikkonen [44,45]. EGARCH evidently is not included here since it is unknown whether $\sup_{\theta \in \Theta} |h_t(\theta) - \sigma_t^2(\theta)| \xrightarrow{p} 0$ as $t \to \infty$ (see Straumann and Mikosch [60], Meitz and Saikkonen [44,45]).

We are interested in heavy tailed errors or innovation outliers, in particular we allow $E[\epsilon_t^4] = \infty$, while GARCH feedback itself may also prompt heavy tails in y_t due to a stochastic recurrence structure (Basrak *et al.* [4], Liu [42]). In this paper, we negligibly transform QML loss or score equations to obtain asymptotically normal estimators of θ^0 allowing for $E[\epsilon_t^4] = \infty$.

Define $\epsilon_t(\theta) := y_t / \sigma_t(\theta)$ and $\mathfrak{s}_t^2(\theta) := (\partial/\partial\theta) \ln \sigma_t^2(\theta)$, and let $I(\cdot)$ denote the indicator function. In Section 2, we tackle the fact that $\sigma_t^2(\theta)$ is not observed for $t \leq 0$. The first method trims QML criterion equations $p_t(\theta) := \ln(\sigma_t^2(\theta)) + \epsilon_t^2(\theta)$ according to extremes that arise in a first order expansion and therefore the score $\sum_{t=1}^n (\epsilon_t^2(\theta) - 1)\mathfrak{s}_t^2(\theta)$. Since $\mathfrak{s}_t^2(\theta)$ has an L_2-bounded envelope near θ^0 it suffices to minimize $\sum_{t=1}^n p_t(\theta) I(-l \leq \epsilon_t^2(\theta) - 1 \leq u)$ for some positive thresholds $\{l, u\}$ that increase with the sample size n. Identification of θ^0 coupled with asymptotic normality are assured if $\{l, u\}$ are replaced with intermediate order statistics of $\epsilon_t^2(\theta) - 1$. The result is the Quasi-Maximum Tail-Trimmed Estimator (QMTTL), similar to the least tail-trimmed squares estimator for autoregressions in Hill [31].

The second method imbeds negligibly transformed errors in QML score equations $(\epsilon_t^2(\theta) - 1)\mathfrak{s}_t^2(\theta)$. We then re-center the transformed errors to minimize small sample bias and estimate θ^0 by the Method of Negligibly Weighted Moments (MNWM). By re-centering we may simply transform $\epsilon_t(\theta)$ itself symmetrically which requires only one threshold, for example in the simple trimming case we use $\epsilon_t^2(\theta) I(|\epsilon_t(\theta)| \leq c)$ for some $c > 0$. In order to simplify proofs we focus on simple trimming, and related bounded but smooth weighted redescending transforms $\epsilon_t^2(\theta) \varpi(\epsilon_t^2(\theta), c) I(|\epsilon_t(\theta)| \leq c)$ where $\varpi(\cdot, c)$ is continuously differentiable in c, and $\varpi(\epsilon_t^2(\theta), c) \to 1$ a.s. as $c \to \infty$. Weights related to simple indicators include Hampel's three-part function, and smooth transforms include Tukey's bisquare and an exponential version (cf. Andrews *et al.* [1], Hampel *et al.* [27]). See Sections 2 and 3.

We show how trimming and distribution tail parameters impact efficiency, while the negligible amount of trimming never affects the asymptotic covariance matrix when $E[\epsilon_t^4] < \infty$. Fixed quantile trimming or truncation always impact efficiency irrespective of higher moments, and cause bias due to $\epsilon_t^2 - 1$ having an asymmetric distribution in general (Sakata and White [58], Mancini *et al.* [43]). Mancini *et al.* [43] use simulation based methods to solve the bias, but this requires knowledge of the error distribution (see also Cantoni and Ronchetti [11], Ronchetti and Trojani [57]).

The convergence rate of our estimators is o($\sqrt{n}$) when $E[\epsilon_t^4] = \infty$, but can be assured to be $\sqrt{n}/g_n$ for any sequence of positive numbers $\{g_n\}$ that satisfies $g_n \to \infty$ as slowly as we choose by following simple rules of thumb for choosing the threshold c. Thus when $E[\epsilon_t^4] = \infty$ our estimators converge faster than QML (cf. Hall and Yao [26]) but slower than $\sqrt{n}$-convergent estimators in Peng and Yao [52], Berkes and Horvath [5] and Zhu and Ling [61], although the latter two are not for standard GARCH models in which $E[\epsilon_t^2] = 1$ identifies the volatility process. See below for literature details. We do not tackle *optimal* threshold selection in order to conserve space. We do, however, show explicitly how threshold selection impacts the convergence rate which suggests simple rules for trimming. We also discuss practical considerations for trimming in terms of small sample bias control. See Sections 2.3 and 2.4.

In Section 4, we show classic inference applies as long as self-normalization is used, a nice convenience since tail thickness and the precise rate of convergence need never be known. We complete the paper with simulation and empirical studies in Sections 5 and 6. In particular, we give evidently the first comparison of various heavy tail robust estimators for GARCH models, and show our estimators obtain in general lower bias and are closer to normally distributed in small samples and therefore lead to better inference.

A complete theory of QML for a variety of strong-GARCH models is presented in Lee and Hansen [39], Berkes *et al.* [6], Francq and Zakoïan [24], Straumann and Mikosch [60] and Meitz and Saikkonen [45] amongst others, while at least a finite fourth moment $E[\epsilon_t^4] < \infty$ is standard. The allowance of heavier tails $E[\epsilon_t^4] = \infty$, with Gaussian asymptotics, evidently only exists for the classic GARCH model, and in most cases requires a non-Gaussian QML criterion and non-standard moment conditions to ensure Fischer consistency (i.e., consistency for the true parameter θ^0). Peng and Yao [52] propose $\sqrt{n}$-convergent Log-LAD, requiring $\ln \epsilon_t^2$ to have a zero median in order to identify θ^0. Berkes and Horvath [5] characterize a general QML criterion class that potentially allows for Fischer consistency, $\sqrt{n}$-convergence and asymptotic normality even when $E[\epsilon_t^4] = \infty$. They treat Gaussian QML, and various non-Gaussian QML like Laplace QML which requires $E|\epsilon_t| = 1$ and $E[\epsilon_t^2] < \infty$, and Power-Law QML (PQML) with index $\vartheta > 1$ requiring that ϵ_t have an infinitessimal moment and $E[|\epsilon_t|/(1 + |\epsilon_t|)] = 1/\vartheta$. Student's t-QML is Fischer consistent when ϵ_t is t-distributed, and otherwise may only be consistent for some $\tilde{\theta} \neq \theta^0$ (cf. Newey and Steigerwald [49], Sakata and White [58], Fan *et al.* [23]).

Zhu and Ling [61] combine Berkes and Horvath [5] Laplace class with Ling's [41] weighting method for Weighted Laplace QML (WLQML) under the assumptions ϵ_t has a zero median, $E|\epsilon_t| = 1$ and $E[\epsilon_t^2] < \infty$. The estimator is $\sqrt{n}$-convergent and asymptotically normal when $E[\epsilon_t^4] = \infty$, but the suggested weights at time t are based on the infinite past $y_{t-1}, y_{t-2}, \ldots$. Although the authors use a central order statistic for a threshold and fix $y_t = 0$ for $t \leq 0$ in the weights for the sake of simulations, they do not prove either is valid. Indeed, for a GARCH(1, 1) the restriction $y_t = 0$ for $t \leq 0$ in their weight (2.4) does not support asymptotic normality (see Zhu and Ling [61], Assumption 2.4 and the discussion on weight (2.4)). Thus, the estimator is not evidently feasible.

Assumptions like $E|\epsilon_t| = 1$ or $E[|\epsilon_t|/(1 + |\epsilon_t|)] = 1/\vartheta$ replace the usual $E[\epsilon_t^2] = 1$ to identify θ^0. Of course, if $E[\epsilon_t^2] \neq 1$ then model (1) is not a standard GARCH model since $E[y_t^2|y_{t-1}, y_{t-2}, \ldots] \neq \sigma_t^2$ with positive probability is possible, and Gaussian QML leads to asymptotic bias. Thus, asymptotic normality *and* Fischer consistency are assured precisely by changing the criterion *and* model assumptions and therefore the model by imposing a non-standard moment condition. In practice, this may be untenable as many analysts in economics

and finance first impose a version of (1) with $E[\epsilon_t^2] = 1$ and then seek a robust estimator. In order to sidestep such unpleasant moment conditions, Fan *et al.* [23] introduce a three-step non-Gaussian QML method. In the first stage, Gaussian QML residuals are generated. In a second stage, a scale parameter is estimated to ensure identification in the third non-Gaussian QML stage without imposing non-standard moment conditions. See also Newey and Steigerwald [49]. Our QMTTL and MNWM estimators are computed in one-step and are asymptotically normal and Fischer consistent by imposing negligible weighting on extremes couched in a Gaussian QML criterion.

Evidently simulation experiments demonstrating the robustness properties of Peng and Yao's [52] Log-LAD, Berkes and Horvath's [5] non-Gaussian QML and Zhu and Ling's [61] WLQML does not exist, while Fan *et al.* [23] only inspect the root-mean squared error of their estimator which masks possible bias. In general, the empirical bias and approximate normality properties of these estimators, as well as their ability to gain accurate inference in small samples (e.g., Wald tests), are unknown.

In a simulation experiment, we show QMTTL and MNWM trump QML, Log-LAD, WLQML, and PQML in all cases in terms of bias, approximate normality and t-test performance, and has lower mean-squared-error than every estimator except PQML (PQML has higher bias and lower dispersion). Overall QMTTL performs best. The dominant performance of QMTTL and MNWM follows since only they directly counter the influence of large errors in small *and* large samples by trimming observations with an error extreme. We show this matters even when ϵ_t is Gaussian: negligible trimming always improves QML performance, while untrimmed QML, Log-LAD, WLQML and PQML are comparatively more sensitive to large errors. Moreover, even PQML, which we design as in Berkes and Horvath [5] to ensure identification for Paretian errors with an infinite fourth moment, has greater bias and is farther from normality in small samples than QMTTL and MNWM. Thus, the advantages of non-Gaussian QML for GARCH processes with heavy tailed errors are not clear, at least as seen by our controlled experiments. We emphasize this last point by tail-trimming PQML in a way that removes adverse sample extremes and leaves the estimator asymptotically unbiased. We show in most cases tail trimming helps PQML in terms of bias, approximate normality and inference, yet overall QMTTL is still better. Indeed, PQML is infeasible unless the tail index of ϵ_t is known or estimated using some filtration for ϵ_t (e.g., QML residuals), and is not Fischer consistent if ϵ_t has any other distribution.

In the literature on additive outlier robust estimation, negligible trimming is an example of a *redescending* transformation $\psi : \mathbb{R} \rightarrow \mathbb{R}$ where in general $\psi(u) \rightarrow 0$ as $|u| \rightarrow \infty$, and typically $\psi(u) = 0$ when $|u| > c$ for some c as we use here. See Huber [34] and Hampel *et al.* [27]. Evidently a complete theory of redescending M-estimators exists only for estimates of location for i.i.d. data (Shevlyakov and Shurygin [59]). In this paper, our QML estimator has a score equation that effectively uses $\psi(\epsilon_t) = (\epsilon_t^2 - 1)I(-l \leq \epsilon_t^2 - 1 \leq u)$ where $l, u \rightarrow \infty$ as $n \rightarrow \infty$. Our Method of Moments estimator is more generic since it uses either re-centered $\psi(\epsilon_t) = \epsilon_t^2 I(|\epsilon_t| \leq c)$ with $c \rightarrow \infty$ as $n \rightarrow \infty$, or related variants like Hampel's three-part weight, as well as smooth weights like Tukey's bisquare. In all cases, the increasing thresholds ensure bias is eradicated asymptotically.

We ignore additive or isolated outliers, and so-called one-off events in $\{y_t\}$ for the sake of brevity. In this case, we would observe $y_t = y_t^* + x_t$ where y_t^* is generated by (1) and, for example, $x_t = 0$ in most periods t. The challenge here is controlling the propagation of an aberrant

observation due to $x_t \neq 0$ through the volatility mechanism. See, for example, Charles and Darné [14], Muler and Yohai [47], and Boudt *et al.* [8], and see Mendes [18] for anecdotal evidence of QML estimator bias. Incorporating additive outliers in (1) with innovation outliers would require additional robustness techniques like those employed in these and related papers (e.g., Muler *et al.* [46]). Some methods, however, are proposed to detect outliers in a GARCH process under the assumption of thin tailed errors: a few large values are simply assumed to be due to a non-heavy tailed outlier.[1] Other estimators, contrary to claims, do not identify θ^0 and/or are not robust to heavy tailed errors.[2] Further, all such robust estimators are proposed for the classic GARCH model, hence existing theory does not necessarily extend to the broader model class (1).

Finally, our methods can be easily extended to higher order GARCH models, GARCH-in-Mean, and models of the conditional mean and variance like nonlinear ARMA–GARCH, as well as other estimators like non-Gaussian QML (Berkes and Horvath [5], Zhu and Ling [61], Fan *et al.* [23]), LAD (Peng and Yao [52]), etc. We show trimming matters for PQML in our simulation study, and we expect negligible trimming to improve upon non-Gaussian QML estimators in general, provided they are Fischer consistent in the first place.

We use the following notation conventions. The indicator function $I(\cdot)$ is $I(a) = 1$ if a is true, and otherwise $I(a) = 0$. The spectral norm of matrix A is $\|A\| = \lambda_{\max}(A'A)^{1/2}$ with $\lambda_{\max}(\cdot)$ the maximum eigenvalue. If z is a scalar, we write $(z)_+ := \max\{0, z\}$. K denotes a positive finite constant whose value may change from line to line; $\iota > 0$ is an arbitrarily tiny constant. $\overset{p}{\to}$ and $\overset{d}{\to}$ denote probability and distribution convergence. $x_n \sim a_n$ implies $x_n/a_n \to 1$. $L(n)$ is a slowly varying function that may change with the context.

2. Quasi-maximum tail-trimmed likelihood

The observed sample is $\{y_t\}_{t=0}^n$ with sample size $n + 1 \geq 1$. We start at $t = 0$ to simplify notation since we condition on the first observation y_0 and a volatility constant defined below. Estimation requires a volatility function on Θ,

$$\sigma_t^2(\theta) = g\big(y_{t-1}, \sigma_{t-1}^2(\theta), \theta\big),$$

---

[1]Charles and Darné [14] extend ideas developed in [15] to test for, and control, additive and innovation outliers in a GARCH process with Gaussian errors. These papers do not provide asymptotic theory, hence the Gaussian assumption can likely be relaxed. The trimming methods used in the present paper can be extended to their test statistics which involve a residual variance estimator (cf. Hill [31], Hill and Aguilar [33]), but a rigorous theory would need to be developed.

[2]Muler and Yohai [47] present a robust M-estimator $\mathring{\theta}_n = \arg\inf_{\theta \in \Theta}\{\sum_{t=1}^n \rho(\ln(y_t^2/h_t^{*2}(\theta)))\}$ where $h_t^{*2}(\theta)$ is a filtered version of $\sigma_t^2(\theta)$ that restricts the propagation of outliers. They assume ρ is thrice continuously differentiable with bounded derivatives. Although claimed to be heavy tail robust and identify the true θ^0 (see their Theorem 3), they do not prove any such ρ exists. In their simulations, for example, they use truncated QML with $\rho(u) = \psi_c(\exp\{u\} - u)$ where ψ_c truncates at a fixed threshold c: $\psi_c(x) = K$ for all $x > c$. Thus $\rho(u)$ is non-differentiable at $\exp\{u\} - u = c$, and at all other points no derivative is bounded which implies non-robustness to heavy tails. The problem is the QML score is not bounded when $\rho(u)$ is truncated according to its large values. Our approach, however, negligibly trims according to properties of the QML score and therefore ensures heavy tail robustness *and* identification of θ^0.

hence $\sigma_t^2 = \sigma_t^2(\theta^0)$. It is convenient to assume Θ is a compact subset of points θ on which $\sigma_t^2(\theta)$ is stationary:

$$\Theta \subseteq \{\theta \in \mathbb{R}^q : \{\sigma_t^2(\theta)\} \text{ has a stationary solution}\}. \tag{2}$$

In practice $\sigma_t^2(\theta)$ for $t \leq 0$ is not observed, so define an iterated volatility approximation

$$h_0(\theta) = \tilde{\omega} > 0 \quad \text{and} \quad h_t(\theta) = g(y_{t-1}, h_{t-1}(\theta), \theta) \qquad \text{for } t = 1, 2, \ldots, \tag{3}$$

where $\tilde{\omega}$ is not necessarily an element of θ. We initially develop an infeasible robust estimator based on the QML equations $\ln \sigma_t^2(\theta) + y_t^2/\sigma_t^2(\theta)$. We then show a feasible version based on $\ln h_t(\theta) + y_t^2/h_t(\theta)$ has the same limit distribution.

2.1. Tail-trimming

In order to understand when and where trimming should be applied, define the GARCH error function, and a scaled volatility function and its derivative

$$\epsilon_t(\theta) := \frac{y_t}{\sigma_t(\theta)} = \frac{y_t}{g(y_{t-1}, \sigma_{t-1}^2(\theta), \theta)},$$

$$\mathfrak{s}_t(\theta) = [\mathfrak{s}_{i,t}(\theta)]_{i=1}^q := \frac{1}{\sigma_t^2(\theta)} \frac{\partial}{\partial \theta} \sigma_t^2(\theta) \quad \text{and} \quad \mathfrak{d}_t(\theta) = [\mathfrak{d}_{i,j,t}(\theta)]_{i,j=1}^q := \frac{\partial}{\partial \theta} \mathfrak{s}_t(\theta).$$

Throughout, we drop θ^0 and write $\epsilon_t = \epsilon_t(\theta^0)$, $\mathfrak{s}_t = \mathfrak{s}_t(\theta^0)$, $\mathfrak{d}_t = \mathfrak{d}_t(\theta^0)$ and so on. Gaussian asymptotics for QML are grounded on the score equations $m_t(\theta)$ and their Jacobian $G_t(\theta)$:

$$m_t(\theta) := (\epsilon_t^2(\theta) - 1)\mathfrak{s}_t(\theta) \quad \text{and}$$

$$G_t(\theta) := \frac{\partial}{\partial \theta} m_t(\theta) = (\epsilon_t^2(\theta) - 1)\mathfrak{d}_t(\theta) - \epsilon_t^2(\theta)\mathfrak{s}_t(\theta)\mathfrak{s}_t(\theta)'. \tag{4}$$

We assume $\mathfrak{s}_t(\theta)$ and $\mathfrak{d}_t(\theta)$ have $L_{2+\iota}$-bounded envelopes near θ^0 for tiny $\iota > 0$, thus asymptotic normality hinges entirely on $\epsilon_t^2 - 1$. See below for all assumptions. It therefore suffices to trim $\ln \sigma_t^2(\theta) + \epsilon_t^2(\theta)$ negligibly when $\epsilon_t^2(\theta) - 1$ surpasses a large negative or positive threshold. As long as those thresholds represent intermediate order statistics, we can identify θ^0 and have an asymptotically normal estimator. Write

$$\mathcal{E}_t(\theta) := \epsilon_t^2(\theta) - 1,$$

and denote left and right tail observations and their order statistics for $\mathcal{E}_t(\theta)$:

$$\mathcal{E}_t^{(-)}(\theta) := \mathcal{E}_t(\theta)I(\mathcal{E}_t(\theta) < 0) \quad \text{and} \quad \mathcal{E}_{(1)}^{(-)}(\theta) \leq \cdots \leq \mathcal{E}_{(n)}^{(-)}(\theta) \leq 0,$$

$$\mathcal{E}_t^{(+)}(\theta) := \mathcal{E}_t(\theta)I(\mathcal{E}_t(\theta) \geq 0) \quad \text{and} \quad \mathcal{E}_{(1)}^{(+)}(\theta) \geq \cdots \geq \mathcal{E}_{(n)}^{(+)}(\theta) \geq 0.$$

The determination of the number of trimmed large $\mathcal{E}_t(\theta)$ in a sample of size n is made by intermediate order sequences $\{k_{1,n}, k_{2,n}\}$, hence (e.g., Leadbetter *et al.* [38])

$$k_{i,n} \in \{1, \ldots, n-1\}, \qquad k_{i,n} \to \infty \quad \text{and} \quad k_{i,n}/n \to 0.$$

Define an indicator selection function for trimming

$$\hat{I}_{n,t}^{(\mathcal{E})}(\theta) := I\big(\mathcal{E}_{(k_{1,n})}^{(-)}(\theta) \le \mathcal{E}_t(\theta) \le \mathcal{E}_{(k_{2,n})}^{(+)}(\theta)\big).$$

The QMTTL estimator therefore solves

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\arg\min} \left\{ \frac{1}{n} \sum_{t=1}^{n} \big(\ln \sigma_t^2(\theta) + \epsilon_t^2(\theta)\big) \times \hat{I}_{n,t}^{(\mathcal{E})}(\theta) \right\} = \underset{\theta \in \Theta}{\arg\min} \big\{ \hat{Q}_n(\theta) \big\}.$$

Each $k_{i,n}$ represents the number of trimmed $\ln \sigma_t^2(\theta) + \epsilon_t^2(\theta)$ due to large negative or positive $\mathcal{E}_t(\theta) = \epsilon_t^2(\theta) - 1$. We require $k_{i,n} \to \infty$ for asymptotic normality, while negligibility $k_{i,n}/n \to 0$ ensures identification of θ^0 asymptotically. Since $\mathcal{E}_t(\theta)$ in general has an asymmetric distribution, identification of θ^0 is assured asymptotically if we negligibly trim asymmetrically by $\mathcal{E}_t(\theta)$. In a method of moments framework, however, we can re-centered trimmed errors allowing for *symmetric* trimming where negative and positive thresholds are the same: see Section 3.

In practical terms, $\hat{\theta}_n$ can be easily computed using standard iterative optimization routines. In fact, under distribution continuity arguments developed in Cizek [16], Lemma 2.1, page 29, apply for almost sure twice differentiability of the otherwise non-differentiable $\hat{Q}_n(\theta)$. In particular, we have *almost surely* $(\partial/\partial\theta)\hat{Q}_n(\theta) = 1/n \sum_{t=1}^{n} m_t(\theta)\hat{I}_{n,t}^{(\mathcal{E})}(\theta)$ and $(\partial/\partial\theta)^2 \hat{Q}_n(\theta) = 1/n \sum_{t=1}^{n} G_t(\theta)\hat{I}_{n,t}^{(\mathcal{E})}(\theta)$. This implies standard estimation algorithms that exploit the gradient and Hessian apply.

In order to characterize the limit distribution of $\hat{\theta}_n$, we require non-random quantiles which the order statistics $\mathcal{E}_{(k_{1,n})}^{(-)}(\theta)$ and $\mathcal{E}_{(k_{2,n})}^{(+)}(\theta)$ approximate. Define sequences $\{\mathcal{L}_n(\theta), \mathcal{U}_n(\theta)\}$ denoting the lower $k_{1,n}/n$ and upper $k_{2,n}/n$ quantiles of $\mathcal{E}_t(\theta)$:

$$P\big(\mathcal{E}_t(\theta) \le -\mathcal{L}_n(\theta)\big) = \frac{k_{1,n}}{n} \quad \text{and} \quad P\big(\mathcal{E}_t(\theta) \ge \mathcal{U}_n(\theta)\big) = \frac{k_{2,n}}{n}. \tag{5}$$

The selection indicator is then

$$I_{n,t}^{(\mathcal{E})}(\theta) := I\big(-\mathcal{L}_n(\theta) \le \mathcal{E}_t(\theta) \le \mathcal{U}_n(\theta)\big).$$

Notice $\mathcal{E}_t(\theta) \in [-1, \infty)$ and $k_{i,n}/n \to 0$ imply $\mathcal{L}_n(\theta) \to 1$ and $\mathcal{U}_n(\theta) \to \infty$. The quantiles $\{\mathcal{L}_n(\theta), \mathcal{U}_n(\theta)\}$ exist for each θ and any choice of fractiles $\{k_{1,n}, k_{2,n}\}$ since ϵ_t has a smooth distribution. By construction the order statistics $\{\mathcal{E}_{(k_{1,n})}^{(-)}(\theta), \mathcal{E}_{(k_{2,n})}^{(+)}(\theta)\}$ estimate $\{\mathcal{L}_n(\theta), \mathcal{U}_n(\theta)\}$, and are uniformly consistent in view of the β-mixing condition detailed in Assumption 1 below, for example $\sup_{\theta \in \Theta} |\mathcal{E}_{(k_{2,n})}^{(+)}(\theta)/\mathcal{U}_n(\theta) - 1| = O_p(1/k_{2,n}^{1/2})$. See Appendix A.3 for supporting limit theory.

Finally, define equation variances Σ_n and $\mathcal{S}_n$, and a scale $\mathcal{V}_n$ for standardizing $\hat{\theta}_n$:

$$\Sigma_n := E\big[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}\big] \times E\big[\mathfrak{s}_t\mathfrak{s}_t'\big] \quad \text{and} \quad \mathcal{S}_n := E\left[\left(\frac{1}{n^{1/2}}\sum_{t=1}^n m_t I_{n,t}^{(\mathcal{E})}\right)\left(\frac{1}{n^{1/2}}\sum_{t=1}^n m_t I_{n,t}^{(\mathcal{E})}\right)'\right],$$

$$\mathcal{V}_n = [\mathcal{V}_{i,j,n}]_{i,j=1}^q := nE\big[\mathfrak{s}_t\mathfrak{s}_t'\big]\mathcal{S}_n^{-1}E\big[\mathfrak{s}_t\mathfrak{s}_t'\big] \sim \frac{n}{E\big[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}\big]}E\big[\mathfrak{s}_t\mathfrak{s}_t'\big].$$

The scale form $\mathcal{V}_n = nE[\mathfrak{s}_t\mathfrak{s}_t']\mathcal{S}_n^{-1}E[\mathfrak{s}_t\mathfrak{s}_t']$ is standard for M-estimators. In view of identification Assumption 2 and equation (6), below, and independence it is easily verified that the long-run variance satisfies $\mathcal{S}_n = \Sigma_n(1 + \mathrm{o}(1))$. Thus $\mathcal{V}_n \sim n(E[\epsilon_t^4 I_{n,t}^{(\mathcal{E})}] - 1)^{-1}E[\mathfrak{s}_t\mathfrak{s}_t']$, which is positive definite for our data generating process.

2.2. Main results

We require two assumptions concerning the error distribution, properties of the volatility response g, and parameter identification. Let κ denote the moment supremum of ϵ_t:

$$\kappa := \arg\sup\big\{\xi > 0 : E|\epsilon_t|^\xi < \infty\big\} > 2.$$

Assumption 1 (Data generating process).

(a) *There exists a unique point $\theta^0 = [\omega^0, \alpha^0, \beta^0]'$ in the interior of a compact subset Θ of $\mathbb{R}^q$ such that $\epsilon_t = y_t/\sigma_t$ is i.i.d., $E[\epsilon_t] = 0$ and $E[\epsilon_t^2] = 1$.*

(b) *ϵ_t has an absolutely continuous, non-degenerate, and uniformly bounded distribution on $(-\infty, \infty)$: $\sup_{a\in\mathbb{R}}\{(\partial/\partial a)P(\epsilon_t \le a)\} < \infty$. If $E[\epsilon_t^4] = \infty$ then $P(|\epsilon_t| > a) = da^{-\kappa}(1 + \mathrm{o}(1))$, where $d > 0$ and $\kappa \in (2, 4]$.*

(c) *$g(\cdot, \cdot, \theta)$ is twice continuously differentiable in θ; $(\partial/\partial\theta)^i g(\cdot, \cdot, \theta)$ is for each $\theta \in \Theta$ and $i = 0, 1, 2$ Borel measurable; $E[\sup_{\theta\in\Theta}|\sigma_t^2/\sigma_t^2(\theta)|^p] < \infty$ for any $p > 0$; $E[\sup_{\theta\in\mathcal{N}_0}\|(\partial/\partial\theta)^i \ln(\sigma_t^2(\theta))\|^{2+\iota}] < \infty$ for $i = 1, 2$, tiny $\iota > 0$, and some compact $\mathcal{N}_0 \subseteq \Theta$ containing θ^0 and having positive Lebesgue measure.*

(d) *$\{y_t\}$ and $\{\sigma_t^2(\theta)\}$ for $\theta \in \Theta$ are stationary and geometrically β-mixing.*

Remark 1. The tail index κ in (b) is identically the moment supremum (see Resnick [55]). The volatility moment bounds in (c) imply only the tails of ϵ_t matter for Gaussian asymptotics, and can be relaxed at the expense of added notation for trimming also according to $\mathfrak{s}_t$. Verification of (c) for the classic GARCH model is in Francq and Zakoïan [24], and related proofs for asymmetric models are in Francq and Zakoïan [25].

Remark 2. Geometric β-mixing (d) implies mixing in the ergodic sense, hence ergodicity (see Petersen [53]). Lipschitz type conditions on the volatility response g combined with a smooth bounded distribution for ϵ_t suffice, covering a large variety of models (Carrasco and Chen [12], Straumann and Mikosch [60], Meitz and Saikkonen [44], Meitz and Saikkonen [45]). See Theorem 2.3 below for one such set of conditions. In the classic GARCH model $y_t = \sigma_t\epsilon_t$ and $\sigma_t^2(\theta) = \omega + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2(\theta)$, for example, where $\omega > 0$, $\alpha, \beta \ge 0$ and $E[\ln(\alpha^0\epsilon_t^2 + \beta^0)] < 0$ ensure

stationarity and ergodicity, and combined with $E[\epsilon_t^2] = 1$ this allows for IGARCH and mildly explosive cases $\alpha^0 + \beta^0 \geq 1$ (Nelson [48]). If additionally ϵ_t has a continuous distribution that is positive on $(-\infty, \infty)$ then $\{y_t, \sigma_t^2(\theta)\}$ are geometrically β-mixing (Carrasco and Chen [12]).

In the Appendices, we show $\hat{\theta}_n$ obtains the expansion $\mathcal{V}_n^{1/2}(\hat{\theta}_n - \theta^0) = n^{-1/2}\Sigma_n^{-1/2} \times \sum_{t=1}^n m_t I_{n,t}^{(\mathcal{E})}(1 + o_p(1))$, hence $n^{1/2}\Sigma_n^{-1/2}E[m_t I_{n,t}^{(\mathcal{E})}] \to 0$ must hold for asymptotic unbiasedness of $\hat{\theta}_n$. This reduces to assuming $n^{1/2}(E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}])^{-1/2}E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] \to 0$ since by independence $E[m_t I_{n,t}^{(\mathcal{E})}] = E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] \times E[\mathfrak{s}_t]$, while $\Sigma_n = E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \times E[\mathfrak{s}_t \mathfrak{s}_t']$ and $\|E[\mathfrak{s}_t \mathfrak{s}_t']\| \in (0, \infty)$.

Assumption 2 (Identification). *The fractile sequences $\{k_{1,n}, k_{2,n}\}$ satisfy $n^{1/2}(E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}])^{-1/2} \times E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] \to 0$ where $\mathcal{E}_t := \epsilon_t^2 - 1$.*

Remark 3. We do not require $E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] = 0$ for finite n since our results are asymptotic, while $E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] \to E[\epsilon_t^2 - 1] = 0$ automatically holds by dominated convergence and negligibility $k_{i,n}/n = o(1)$. Since $n^{1/2}/(E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}])^{1/2} \to \infty$ as verified in Section 2.4 below, we require $E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] \to 0$ fast enough, else there is asymptotic bias.

Remark 4. There always exists a sequence $\{k_{1,n}, k_{2,n}\}$ such that $E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}]$ is closer to zero than $(E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}])^{1/2}/n^{1/2}$ as n increases. In general $\mathcal{E}_t \in [-1, \infty)$ is skewed right hence, counterintuitively, asymptotic unbiasedness requires $k_{1,n} > k_{2,n}$: a few trimmed large positive values promotes asymptotic normality, but forces us to trim many negative values to ensure identification. See Section 2.3 for discussion and examples. In a method of moments framework, however, identification is assured by re-centering the trimmed errors, hence Assumption 2 is not required. See Section 3.

Remark 5. Define $m_{n,t} := m_t I_{n,t}^{(\mathcal{E})}$. Assumption 2 ensures $E[\{m_{n,s} - E[m_{n,s}]\}\{m_{n,t} - E[m_{n,t}]\}'] = E[m_{n,s}m_{n,t}'] + o(\|\Sigma_n\|/n)$ for all s, t, and $\|\sum_{i=1}^{n-1} E[m_{n,1}m_{n,i+1}']\| \leq n \times o(\|\Sigma_n\|/n) = o(\|\Sigma_n\|)$ by Minkowski and Cauchy–Schwarz inequalities. Hence, Σ_n is asymptotically equal to the long-run covariance matrix $\mathcal{S}_n$ of $n^{-1/2}\sum_{t=1}^n \{m_{n,t} - E[m_{n,t}]\}$ since

$$E\left[\left(\frac{1}{n^{1/2}}\sum_{t=1}^n \{m_{n,t} - E[m_{n,t}]\}\right)\left(\frac{1}{n^{1/2}}\sum_{t=1}^n \{m_{n,t} - E[m_{n,t}]\}\right)'\right]$$

$$= \Sigma_n \times (1 + o(1)) + 2\sum_{i=1}^{n-1}\left(1 - \frac{i}{n}\right)E[m_{n,1}m_{n,i+1}'] = \Sigma_n \times (1 + o(1)). \tag{6}$$

We are now ready to state the main results of this section. The expansion $\mathcal{V}_n^{1/2}(\hat{\theta}_n - \theta^0) = n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^n m_t I_{n,t}^{(\mathcal{E})}(1 + o_p(1))$ requires Jacobian consistency $1/n\sum_{t=1}^n G_t(\hat{\theta}_n)\hat{I}_{n,t}^{(\mathcal{E})}(\hat{\theta}_n) \xrightarrow{p} -E[\mathfrak{s}_t \mathfrak{s}_t']$ and therefore consistency $\hat{\theta}_n \xrightarrow{p} \theta^0$ from first principles. Proofs of main results are contained in Appendices A.1 and A.2.

Theorem 2.1 (QMTTL consistency). *Under Assumptions* 1 *and* 2 $\hat{\theta}_n \xrightarrow{p} \theta^0$.

Theorem 2.2 (QMTTL normality). *Under Assumptions* 1 *and* 2 $\mathcal{V}_n^{1/2}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N(0, I_q)$ *where* $\mathcal{V}_n = nE[\mathfrak{s}_t\mathfrak{s}_t']\mathcal{S}_n^{-1}E[\mathfrak{s}_t\mathfrak{s}_t'] \sim n(E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}])^{-1}E[\mathfrak{s}_t\mathfrak{s}_t']$ *and each* $\mathcal{V}_{i,i,n} \to \infty$.

Now consider feasible QMTTL. Define $\tilde{\epsilon}_t(\theta) := y_t^2/h_t(\theta)$ based on the iterated process $\{h_t(\theta)\}$ in (3), and $\tilde{\mathcal{E}}_t(\theta) := \tilde{\epsilon}_t^2(\theta) - 1$. The feasible estimator is

$$\tilde{\theta}_n = \arg\min_{\theta \in \Theta}\left\{\frac{1}{n}\sum_{t=1}^{n}\left(\ln h_t(\theta) + \tilde{\epsilon}_t^2(\theta)\right) \times I\left(\tilde{\mathcal{E}}_{(k_{1,n})}^{(-)}(\theta) \leq \tilde{\mathcal{E}}_t(\theta) \leq \tilde{\mathcal{E}}_{(k_{2,n})}^{(+)}(\theta)\right)\right\}.$$

Under the following Lipschitz bounds for the response g and its derivatives we show $\tilde{\theta}_n$ has the same limit distribution as the infeasible $\hat{\theta}_n$, cf. Meitz and Saikkonen [44]. Related ideas are contained in Straumann and Mikosch [60].

Drop arguments: $g = g(y, s, \theta)$, and let g_a and $g_{a,b}$ denote first and second derivatives for $a, b \in \{y, s, \theta\}$. We say a matrix function $\xi(y, s, \theta)$ is Lipschitz in s if $\|\xi(y, s_1, \theta) - \xi(y, s_2, \theta)\| \leq K|s_1 - s_2| \,\forall s_1, s_2 \in [0, \infty)$ and $y, \theta \in \mathbb{R} \times \Theta$.

Assumption 3 (Response bounds).

   (a) $g \leq \rho s + K(1 + y^2)$ *for some* $\rho \in (0, 1)$ *and* $\inf_{y\in\mathbb{R}, s\in\mathbb{R}_+, \theta\in\Theta}\{|g|\} > 0$;
   (b) $\|g_a\|$ *and* $\|g_{a,b}\|$ *are bounded by* $K(1 + y^2 + s)$ *for each* $a, b \in \{y, \theta\}$;
   (c) g, g_a *and* $g_{a,b}$ *are Lipschitz in* s, *for each* $a, b \in \{y, s, \theta\}$.

Assumption 3 ensures $h_t(\theta)$, $h_t^\theta(\theta) := (\partial/\partial\theta)h_t(\theta)$ and $h_t^{\theta,\theta}(\theta) := (\partial/\partial\theta)h_t^\theta(\theta)$ have stationary ergodic solutions $\{h_t^*(\theta), h_t^{\theta*}(\theta), h_t^{\theta,\theta*}(\theta)\}$ with the geometric property $E[(\sup_{\theta\in\Theta}|a_t^*(\theta) - a_t(\theta)|)^t] = o(\rho^t)$ for each $a_t(\theta) \in \{h_t(\theta), h_{i,t}^\theta(\theta), h_{i,j,t}^{\theta,\theta}(\theta)\}$ and $a_t^*(\theta) \in \{h_t^*(\theta), h_{i,t}^{\theta*}(\theta), h_{i,j,t}^{\theta,\theta*}(\theta)\}$ and some $\rho \in (0, 1)$. See Lemma A.7 in Appendix A.2. This leads to the next result.

Theorem 2.3 (Feasible QMTTL). *Under Assumptions* 1–3 $\mathcal{V}_n^{1/2}(\tilde{\theta}_n - \hat{\theta}_n) \xrightarrow{p} 0$.

Remark 6. In the remainder of the paper, we focus on the infeasible $\hat{\theta}_n$ for notational economy.

As stated above, we need only trim by error extremes since first order asymptotics rests solely on whether ϵ_t has a fourth moment or not. However, in small samples a large y_{t-1} may cause $\mathfrak{s}_t$ or $\mathfrak{d}_t$ to spike and therefore the score equation to exhibit a sample extreme value. Consider, for example, that in the linear volatility model $\sigma_t^2(\theta) = \omega + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2(\theta)$ the score weight at the origin $\mathfrak{s}_t(\theta)|_{\alpha,\beta=0} = \omega^{-1} \times [1, y_{t-1}^2, \omega]'$ obtains an extreme value *if and only if* $|y_{t-1}|$ does. In general $\mathfrak{s}_t$ exhibits spikes when $|y_{t-1}|$ does for α^0 and β^0 near zero. This same properly applies to a large variety of GARCH models. Thus, although $\hat{\theta}_n$ is consistent and asymptotically normal, for improved small sample performance trimming by large values of y_{t-1} appears to be

highly useful in practice. This is not surprising since true additive outliers render QML biased (see Mendes [18], Muler and Yohai [47], cf. Cavaliere and Georgiev [13], Muler *et al.* [46]).

Let $\{\tilde{k}_n\}$ be an intermediate order sequence and define $\hat{I}_{n,t}^{(y)} := I(|y_t| \leq y_{(\tilde{k}_n)}^{(a)})$ where $y_{(i)}^{(a)}$ are order statistics of $y_t^{(a)} := |y_t|$. The estimator in this case is

$$\hat{\theta}_n^{(y)} = \arg\min_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{t=1}^n \left(\ln \sigma_t^2(\theta) + \epsilon_t^2(\theta) \right) \times \hat{I}_{n,t}^{(\mathcal{E})}(\theta) \hat{I}_{n,t-1}^{(y)} \right\}.$$

Since $\hat{I}_{n,t-1}^{(y)} \xrightarrow{p} 1$, the score equations $\mathfrak{s}_t$ are square integrable, and ϵ_t is i.i.d., asymptotic normality does not depend on whether y_t is heavy tailed. Indeed, it is easy to show $\hat{\theta}_n^{(y)}$ is asymptotically equivalent to $\hat{\theta}_n$. The same property extends to feasible QMTTL with trimming by y_{t-1}, denoted $\tilde{\theta}_n^{(y)}$. We therefore omit the proof of the next result.

Corollary 2.4. *Under Assumptions* 1 *and* 2, *trimming by* y_{t-1} *does not impact the limit distributions of infeasible and feasible QMTTL estimators:* $\mathcal{V}_n^{1/2}(\hat{\theta}_n^{(y)} - \hat{\theta}_n) \xrightarrow{p} 0$ *and* $\mathcal{V}_n^{1/2}(\tilde{\theta}_n^{(y)} - \tilde{\theta}_n) \xrightarrow{p} 0$. *Moreover, infeasible and feasible estimators are asymptotically equivalent:* $\mathcal{V}_n^{1/2}(\tilde{\theta}_n^{(y)} - \hat{\theta}_n^{(y)}) \xrightarrow{p} 0$.

2.3. Verification of identification Assumption 2

We require an explicit model of $P(|\epsilon_t| > c)$ in order to verify Assumption 2. In our simulation study, we use distributions with either power law or exponential tail decay.

2.3.1. *Paretian tails*

In the simulation experiment we use

$$P(|\epsilon_t| > c) = (1 + c)^{-\kappa} \qquad \text{with } \kappa \in (2, 4), \tag{7}$$

hence $\mathcal{E}_t$ has left and right tails:

$$P(\mathcal{E}_t < -c) = P(\epsilon_t^2 < 1 - c) = 0 \qquad \text{if } c \geq 1,$$
$$= 1 - P(\epsilon_t^2 > 1 - c) = 1 - (2 - c)^{-\kappa} \qquad \text{if } c \in [0, 1], \tag{8}$$
$$P(\mathcal{E}_t > c) = P(\epsilon_t^2 > 1 + c) = (2 + c)^{-\kappa/2}.$$

We show below identification $n^{1/2}(E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}])^{-1/2} E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] \to 0$ holds if $k_{1,n} \to \infty$, $k_{1,n}/n \to 0$ and:

$$\left(\frac{k_{2,n}}{n} \right)^{1-2/\kappa} = \frac{\kappa - 2}{2} \left(-1 + \left(\frac{1}{1 - k_{1,n}/n} \right)^{2/\kappa} + \frac{2}{\kappa - 2} \frac{k_{1,n}}{n} \right)$$
$$+ o\left(\left(\frac{n}{k_{1,n}} \right)^{2/\kappa - 1/2} \frac{1}{n^{1/2}} \right). \tag{9}$$

In practice, (9) is greatly simplified asymptotically by noting $(n/k_{1,n})^{2/\kappa-1/2}n^{-1/2} = \mathrm{o}(1)$ and $(1 - k_{1,n}/n)^{-2/\kappa} - 1 \sim (2/\kappa)(k_{1,n}/n)$, hence identification applies if $(k_{2,n}/n)^{1-2/\kappa} \sim ((2\kappa - 2)/\kappa)(k_{1,n}/n)$ or

$$\frac{k_{2,n}}{k_{1,n}^{\kappa/(\kappa-2)}} \sim 2\left(1 - \frac{1}{\kappa}\right)^{\kappa/(\kappa-2)} \frac{1}{n^{\kappa/(\kappa-2)-1}}. \tag{10}$$

A similar condition applies in the second order power law case $P(|\epsilon_t| > c) = dc^{-\kappa}(1 + ec^{-\xi})$ with $d, e > 0, \xi > 0$ and $\kappa \in (2, 4)$, while a less sharp result arises under $P(|\epsilon_t| > c) = dc^{-\kappa}(1 + \mathrm{o}(1))$.

In order to show (9), we must characterize the moments $E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] = E[\mathcal{E}_t I(-\mathcal{L}_n \leq \mathcal{E}_t \leq \mathcal{U}_n)]$ and $E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}]$. Use (8) to deduce $\mathcal{U}_n = (n/k_{2,n})^{2/\kappa} - 2 \to \infty$ and $\mathcal{L}_n = 2 - (n/(n - k_{1,n}))^{2/\kappa} \in [0, 1]$ as $n \to \infty$. Therefore,

$$
\begin{aligned}
&E\big[\mathcal{E}_t I(-\mathcal{L}_n \leq \mathcal{E}_t \leq \mathcal{U}_n)\big]\\
&= -\big\{E\big[\mathcal{E}_t I(\mathcal{E}_t > \mathcal{U}_n)\big] + E\big[\mathcal{E}_t I(\mathcal{E}_t < -\mathcal{L}_n)\big]\big\}\\
&= -\left\{\int_{\mathcal{U}_n}^{\infty} (2 + u)^{-\kappa/2}\,\mathrm{d}u - \int_{\mathcal{L}_n}^{1} \big(1 - (2 - u)^{-\kappa/2}\big)\,\mathrm{d}u\right\}\\
&= -\left\{\frac{2}{\kappa - 2}\left(\frac{k_{2,n}}{n}\right)^{1-2/\kappa} + 1 - \left(\frac{n}{n - k_{1,n}}\right)^{2/\kappa} - \frac{2}{\kappa - 2}\frac{k_{1,n}}{n}\right\}.
\end{aligned}
\tag{11}
$$

Next $E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \sim K(n/k_{2,n})^{4/\kappa-1}$ follows from (15) below. Combined with (11) and by rearranging terms, Assumption 2 holds when $E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] = \mathrm{o}((E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}])^{1/2}/n^{1/2})$, hence when

$$
\begin{aligned}
\left(\frac{k_{2,n}}{n}\right)^{1-2/\kappa} &= \frac{\kappa - 2}{2}\left(-1 + \left(\frac{1}{1 - k_{1,n}/n}\right)^{2/\kappa} + \frac{2}{\kappa - 2}\frac{k_{1,n}}{n}\right)\\
&\quad + \mathrm{o}\left(\left(\frac{n}{k_{2,n}}\right)^{2/\kappa-1/2}\frac{1}{n^{1/2}}\right).
\end{aligned}
\tag{12}
$$

Notice $k_{2,n}$ appears on both sides of the equality. In order to achieve (9), note $k_{2,n}/k_{1,n} \to 0$. This follows since $(n/k_{2,n})^{2/\kappa-1/2}n^{-1/2} = \mathrm{o}(1)$ and by the mean-value-theorem $(1 - k_{1,n}/n)^{-2/\kappa} - 1 \sim (2/\kappa)k_{1,n}/n$ hence $(k_{2,n}/n)^{1-2/\kappa} \sim Kk_{1,n}/n$, therefore

$$\left(\frac{k_{2,n}}{k_{1,n}}\right)^{1-2/\kappa} = \left(\frac{k_{2,n}/n}{k_{1,n}/n}\right)^{1-2/\kappa} \sim K\frac{(k_{1,n}/n)}{(k_{1,n}/n)^{1-2/\kappa}} = K(k_{1,n}/n)^{2/\kappa} \to 0.$$

Now combine $k_{2,n}/k_{1,n} \to 0$ and (12) to deduce (9).

There are several things to note from (10). First, there are arbitrarily many valid $\{k_{1,n}, k_{2,n}\}$. Second, $\{k_{1,n}, k_{2,n}\}$ requires knowledge of κ, which can be consistently estimated for many processes defined by (1) (see Hill [29]). However, the method of moments estimator in Section 3 only requires one two-tailed fractile *without* knowledge of κ.

Third, $k_{2,n}/k_{1,n} \to 0$ since $k_{1,n}/n \to 0$ and $\kappa > 2$. This logically follows since $\mathcal{E}_t$ has support $[-1, \infty)$. The right tail is heavier, hence trimming a positive extreme must be off-set by trimming more negative observations in order to get $E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] \approx 0$.

Fourth, $k_{1,n} \sim n/g_{1,n}$ for slowly varying $g_{2,n} \to \infty$ implies $k_{2,n} \sim n/g_{2,n}$ for slowly varying $g_{2,n}, g_{2,n}/g_{1,n} \to \infty$. Similarly, $k_{1,n} \sim \lambda_1 n^{\delta_1}$ for $\lambda_1 \in (0, 1)$ and $\delta_1 \in (2/\kappa, 1)$ implies $k_{2,n} \sim \lambda_2 n^{\delta_2}$ for $\lambda_2 \in (0, 1)$ and $\delta_2 \in (0, \delta_1)$. Further, slowly varying $k_{1,n} \to \infty$ is not valid since $k_{2,n} \to 0$ is then required which leads to asymptotic non-normality when $E[\epsilon_t^4] = \infty$.

Fifth, we need monotonically larger $k_{1,n}$ as $\kappa \searrow 2$, but always $\limsup_{n\to\infty}(k_{2,n}/k_{1,n}) < 1$. Exponential tails treated in Section 2.3.2 reveals an extreme case: there are no limitations on how we set $\{k_{1,n}, k_{2,n}\}$ outside of an upper bound, although $k_{1,n} > k_{2,n}$ always reduces small sample bias.

Finally, as a numerical example suppose $\kappa = 2.5$ and $n = 100$. If $k_{2,n} = 1$ then $k_{1,n} = 33$ renders (10) a near equality, although any $k_{1,n} \in \{29, \dots, 35\}$ aligns with $k_{2,n} = 1$ by rounding. This is striking: we need to trim roughly 33 times as many negative $\mathcal{E}_t(\theta)$ as positive $\mathcal{E}_t(\theta)$ to approach unbiasedness at $n = 100$. If $n = 800$ then, for example, $k_{2,n} = 2$ aligns with roughly $k_{1,n} = 200$.

2.3.2. *Exponential tails*

Now suppose ϵ_t has a Laplace distribution:

$$P(\epsilon_t \le -c) = \tfrac{1}{2}\exp\{-\sqrt{2}c\} \qquad \text{for } c > 0 \quad \text{and} \quad P(\epsilon_t > c) = \tfrac{1}{2}\exp\{-\sqrt{2}c\} \qquad \text{for } c \ge 0.$$

We use a normal distribution in our simulation study, but the exposition here is greatly simplified under Laplace, while the conclusions are the same.

We have $P(\mathcal{E}_t \le -c) = 1 - \exp\{-\sqrt{2}(1-c)^{1/2}\}$ and $P(\mathcal{E}_t \ge c) = \exp\{-\sqrt{2}(1+c)^{1/2}\}$. The following are then straightforward to verify: $\mathcal{L}_n = 1 - (\ln(n/(n - k_{1,n})))^2$ and $\mathcal{U}_n = (\ln(0.5n/k_{2,n}))^2 - 1$, hence

$$E\big[\mathcal{E}_t I(-\mathcal{L}_n \le \mathcal{E}_t \le \mathcal{U}_n)\big]$$
$$= 2\left(\frac{k_{1,n}}{n} - \frac{k_{2,n}}{n}\right) - \ln\left(\frac{n}{n - k_{1,n}}\right)\left\{-\ln\left(\frac{n}{n - k_{1,n}}\right) + 2\left(\frac{n - k_{1,n}}{n}\right)\right\}.$$

Observe $E[\mathcal{E}_t I(-\mathcal{L}_n \le \mathcal{E}_t \le \mathcal{U}_n)] \approx 0$ when $k_{1,n} > k_{2,n}$, hence if $k_{1,n}/n \to 0$ then $k_{2,n}/n \to 0$ must hold.

Since $E[\epsilon_t^4] < \infty$ we need $E[\mathcal{E}_t I(-\mathcal{L}_n \le \mathcal{E}_t \le \mathcal{U}_n)] = o(1/n^{1/2})$. Notice $\ln(n/(n - k_{1,n})) \sim k_{1,n}/n$. Hence if simply each $k_{i,n} = o(n^{1/2})$, then we achieve $E[\mathcal{E}_t I(-\mathcal{L}_n \le \mathcal{E}_t \le \mathcal{U}_n)] = o(1/n^{1/2})$. This implies that technically we do not even need asymmetric trimming $k_{1,n} > k_{2,n}$ as long as we set $k_{1,n} = k_{2,n} = o(n^{1/2})$. This follows since tails are so thin that in general extremes on $[0, \infty)$ are not much larger than extremes on $[-1, 0)$ in small samples. Similarly, we can use any form of asymmetric trimming that satisfies $k_{i,n} = o(n^{1/2})$. We show by simulation that as n gets large, bias evaporates irrespective of $k_{i,n}$, but $k_{1,n} > k_{2,n}$ always leads to lower small sample bias.

2.3.3. *Remarks*

We demonstrate by simulation in Section 5 that using $k_{1,n} = 10k_{2,n}$ or $k_{1,n} = 35k_{2,n}$ for either $n \in \{100, 800\}$ and either Paretian or Gaussian ϵ_t leads to a superb QMTTL estimator. Indeed, simply using symmetric trimming $k_{1,n} = k_{2,n}$ still leads to a better estimator than Log-LAD and Weighted Laplace QML in terms of small sample bias and approximate normality, although Power-Law QML tends to have lower bias and be closer to normal. In general using bias minimizing fractiles, like $k_{1,n} = 100k_{2,n}$ for Paretian ϵ_t when $n = 800$, is not evidently required for obtaining low bias in finite samples, as long as $k_{1,n}$ is comparatively large relative to $k_{2,n}$ in which case QMTTL trumps Log-LAD, WLQML and PQML.

We also find that our method of moments estimator in Section 3 dominates Log-LAD, WLQML and PQML, although QMTTL with $k_{1,n} = 35k_{2,n}$ leads to smaller bias and is closer to normally distributed in nearly every case. Nevertheless, the method of moments estimator is always asymptotically unbiased and easier to implement because trimming is symmetric. Which estimator is chosen in practice depends on the analyst's preferences: method of moments is guaranteed to be asymptotically unbiased, but QMTTL has superior small sample properties even if $\{k_{1,n}, k_{2,n}\}$ are not chosen to ensure asymptotic unbiasedness.

2.4. QMTTL scale and rate of convergence

The scale $\mathcal{V}_n$ and rate of convergence depend on the error tail index $\kappa > 2$. If $E[\epsilon_t^4] < \infty$ then by dominated convergence $E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] = E[(\epsilon_t^2 - 1)^2 I_{n,t}^{(\mathcal{E})}] \to E[(\epsilon_t^2 - 1)^2] = E[\epsilon_t^4] - 1$, thus $\mathcal{V}_n \sim n(E[\epsilon_t^4] - 1])^{-1}E[\mathfrak{s}_t\mathfrak{s}_t']$, the classic QML asymptotic covariance matrix. This implies trimming does not affect efficiency asymptotically. Hence, we now assume $E[\epsilon_t^4] = \infty$.

Let the intermediate order sequences $\{k_n\}$ and positive thresholds $\{\mathcal{C}_n(\theta)\}$ satisfy

$$P\big(|\mathcal{E}_t(\theta)| \geq \mathcal{C}_n(\theta)\big) = \frac{k_n}{n}.$$

The rate $E[\mathcal{E}_t^2(\theta)I_{n,t}^{(\mathcal{E})}(\theta)] \to \infty$ is logically governed by the right tail of $\mathcal{E}_t(\theta) = \epsilon_t^2(\theta) - 1 \in [-1, \infty)$ since by dominated convergence:

$$E\big[\mathcal{E}_t^2(\theta)I_{n,t}^{(\mathcal{E})}(\theta)\big] = E\big[\mathcal{E}_t^2(\theta)I\big(-\mathcal{L}_n(\theta) \leq \mathcal{E}_t(\theta) \leq \mathcal{U}_n(\theta)\big)\big]$$
$$\sim E\big[\mathcal{E}_t^2(\theta)I\big(|\mathcal{E}_t(\theta)| \leq \mathcal{C}_n(\theta)\big)\big]$$

as though $\mathcal{E}_t(\theta)$ were symmetrically trimmed with thresholds and fractile

$$\mathcal{C}_n(\theta) = \mathcal{U}_n(\theta) \quad \text{and} \quad k_n = k_{2,n}. \tag{13}$$

Note $E[\mathcal{E}_t^2(\theta)I_{n,t}^{(\mathcal{E})}(\theta)] \sim E[\mathcal{E}_t^2(\theta)I(|\mathcal{E}_t(\theta)| \leq \mathcal{C}_n(\theta))]$ is useful for characterizing the convergence rate, but identification Assumption 2 in general requires $k_{1,n} > k_{2,n}$ hence $\mathcal{L}_n(\theta) < \mathcal{U}_n(\theta)$.

As long as $E[\epsilon_t^4] = \infty$, then the rate of convergence is $\mathcal{V}_n^{1/2} = o(n^{1/2})$: heavy tailed errors can only adversely affect the convergence rate. The exact rate can be deduced by observing that from

$P(|\epsilon_t| > a) = da^{-\kappa}(1 + o(1))$ the variable $\mathcal{E}_t = \epsilon_t^2 - 1$ has a tail sum dominated by the right tail:

$$
\begin{aligned}
P(|\mathcal{E}_t| > a) &= P(\epsilon_t^2 > 1 + a) + P(\epsilon_t^2 < 1 - a) \\
&= d(1 + a)^{-\kappa/2}(1 + o(1)) = da^{-\kappa/2}(1 + o(1)) \qquad \text{as } a \to \infty.
\end{aligned}
\tag{14}
$$

Hence, the thresholds $\mathcal{C}_n$ can always be chosen as $\mathcal{C}_n = d^{2/\kappa}(n/k_n)^{2/\kappa}$. Now use an implication of Karamata's theorem to obtain as $n \to \infty$ (e.g., Resnick [55], Theorem 0.6):[3]

$$
\begin{aligned}
&\kappa = 4 : E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \sim d\ln(n), \\
&\kappa \in (2,4) : E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \sim \left(\frac{\kappa}{4 - \kappa}\right)\mathcal{C}_n^2 P(|\mathcal{E}_t| > \mathcal{C}_n) = \left(\frac{\kappa}{4 - \kappa}\right)d^{4/\kappa}\left(\frac{n}{k_n}\right)^{4/\kappa - 1} = o(n).
\end{aligned}
\tag{15}
$$

The following claim summarizes the above details.

Theorem 2.5 (Convergence rate). *Under Assumptions* 1 *and* 2 *if* $\kappa > 4$ *then* $\mathcal{V}_n \sim n(E[\epsilon_t^4] - 1])^{-1}E[\mathfrak{s}_t\mathfrak{s}_t']$. *If* $\kappa \leq 4$ *then for* $i = 1, \ldots, q$

$$
\begin{aligned}
&\kappa = 4 : \mathcal{V}_{i,i,n}^{1/2} \sim \left(\frac{n}{\ln(n)}\right)^{1/2} d^{-1/2}(E[\mathfrak{s}_{i,t}^2])^{1/2}, \\
&\kappa \in (2,4) : \mathcal{V}_{i,i,n}^{1/2} \sim n^{1/2}\left(\frac{k_n}{n}\right)^{2/\kappa - 1/2} d^{-2/\kappa}\left(\frac{4 - \kappa}{\kappa}\right)^{1/2}(E[\mathfrak{s}_{i,t}^2])^{1/2}.
\end{aligned}
\tag{16}
$$

There are several key observations. First, as long as $\kappa \in (2,4)$ then elevating k_n arbitrarily close to a fixed percent of n, that is $k_n \approx \lambda n$ for $\lambda \in (1,0)$, will optimize the convergence rate. This is logical since large errors adversely affect efficiency. In general this implies

$$
k_n \sim n/g_n \qquad \text{for } g_n \to \infty \text{ at a slow rate,}
\tag{17}
$$

ensures $\mathcal{V}_{i,i,n}^{1/2} \sim n^{1/2}/g_n^{2/\kappa - 1/2}$ for any $\kappa \in (2,4]$. Hence, $\mathcal{V}_{i,i,n}^{1/2} \to \infty$ can be driven as close to rate $n^{1/2}$ as we choose by setting $g_n \to \infty$ very slowly (e.g., $g_n = \ln(\ln(n))$). Further, the rate monotonically $n^{1/2}/g_n^{2/\kappa - 1/2} \nearrow n^{1/2}$ as $\kappa \nearrow 4$. Hall and Yao [26] show the QML rate is $n^{1 - 2/\kappa}/L(n)$ for some slowly varying $L(n) \to \infty$ and any $\kappa \in (2,4]$, hence QMTTL can be assured to be faster for every $\kappa \in (2,4)$. Conversely, Peng and Yao's [52] Log-LAD and non-Gaussian QML are $n^{1/2}$-convergent (cf. Berkes and Horvath [5], Zhu and Ling [61]), but the higher rate is not without costs: (i) these estimators are not robust to error extremes in small samples: see Section 5; (ii) Log-LAD requires $\ln \epsilon_t^2$ to have a zero median; and (iii) non-Gaussian QML requires additional moment conditions for Fischer consistency, for example, WLQML requires $E|\epsilon_t| = 1$: see Section 1 for discussion.

---

[3]Note if $\kappa = 4$ then for finite $a > 0$ there exists $K > 0$ such that $E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \sim \int_0^{\mathcal{C}_n^2} P(\mathcal{E}_t > u^{1/2})\,du = K + \int_a^{\mathcal{C}_n^2} P(\mathcal{E}_t > u^{1/2})\,du \sim K + d\int_{a+1}^{\mathcal{C}_n^2} u^{-\kappa/4}\,du \sim K + d\ln(\mathcal{C}_n^4) \sim d\ln(n)$ since $\mathcal{C}_n = K(n/k_n)^{1/4}$.

Second, if $\kappa < 4$ and we use a fractile form $k_n \sim \lambda n / g_n$ for slow $g_n \to \infty$ and $\lambda \in (0, 1]$, then

$$\frac{n^{1/2}}{g_n^{2/\kappa - 1/2}} (\hat{\theta}_n - \theta^0) \xrightarrow{d} N\left(0, \lambda^{-(2/\kappa - 1/2)} \left(\frac{\kappa}{4 - \kappa}\right) d^{4/\kappa} \left(E[\mathfrak{s}_t \mathfrak{s}_t']\right)^{-1}\right)$$

$$= N\big(0, \mathcal{V}(\lambda, \kappa, d)\big). \tag{18}$$

For example, in our simulation study we use $k_n \sim \lambda n / \ln(n)$, hence $\hat{\theta}_n$ is $n^{1/2}/(\ln(n))^{2/\kappa - 1/2}$-convergent with asymptotic variance $\mathcal{V}(\lambda, \kappa, d)$. The asymptotic variance $\mathcal{V}(\lambda, \kappa, d)$ can always by decreased by increasing λ and therefore removing more extremes per sample.

Third, in view of $k_n = k_{2,n}$ by (13), trimming rule (17) only concerns the amount of trimmed positive observations of $\mathcal{E}_t = \epsilon_t^2 - 1$: the left tail of $\mathcal{E}_t$ is bounded, hence only the rate of right tail trimming of $\mathcal{E}_t$ matters for the *convergence rate*. In terms of *identification*, however, as discussed in Section 2.3 the number of trimmed left and right tail observations $k_{1,n}$ and $k_{2,n}$ must be balanced when ϵ_t is governed by a heavy tailed distribution. For example, if $P(|\epsilon_t| > c) = (1 + c)^{-\kappa}$ with $\kappa \in (2, 4)$, and $k_{1,n} \sim \lambda n / \ln(n)$, both as in our simulation study, then Assumption 2 holds when $k_{2,n} \sim K k_{1,n}^{\kappa/(\kappa-2)} / n^{\kappa/(\kappa-2)-1} \sim K n / (\ln(n))^{\kappa/(\kappa-2)}$, hence from (18) the rate of convergence is $n^{1/2}/((\ln(n))^{\kappa/(\kappa-2)})^{2/\kappa - 1/2} = n^{1/2}/(\ln(n))^{(4-\kappa)/[2(\kappa-2)]}$.

As a practical matter, naturally too much trimming in any given sample can lead to small sample bias in $\hat{\theta}_n$. In Section 5, we use $k_n \sim \lambda n / \ln(n)$ with $\lambda = 0.025$ for both very thin and thick tailed error distributions: values much larger than 0.025 (e.g., $\lambda = 0.10$) leads to substantial bias, and values much smaller (e.g., $\lambda = 0.01$) are not effective for rendering $\hat{\theta}_n$ approximately normal in small samples. In general any value $\lambda \in [0.02, 0.05]$ leads to roughly the same results. Similar trimming schemes are found to be highly successful in other robust estimation and inference contexts: see Hill [30,31] and Hill and Aguilar [33].

Last, there are several proposed methods in the robust statistics literature for selecting trimming parameters like λ, but in this literature the seeming universal approach for data transformations involve a fixed quantile threshold hence $k_n \sim \lambda n$ (cf. Huber [34], Hampel *et al.* [27], Jureckova and Sen [37]). Such methods include covariance determinant or asymptotic variance minimization where a unique internal solution for λ exists. These methods are ill posed here since they lead to corner solutions: consider that minimizing $\mathcal{V}(\lambda, \kappa, d)$ above on $\lambda \in [\underline{\lambda}, \bar{\lambda}]$ leads to $\lambda = \bar{\lambda}$. See Hill and Aguilar [33] for references and simulation evidence. In terms of inference more choices exist, including test statistic functionals over λ like the supremum, and empirical process techniques for p-value computation (see Hill [30]).

3. Method of moments with re-centering

Our second estimator uses the method of moments based on negligibly weighted errors imbedded in a QML score equation. This gives us the advantage of re-centering to ensure identification. It therefore allows us to use a greater variety of error transforms, as well as symmetric transforms even if the errors have an asymmetric distribution. Define $\Im_t := \sigma(y_\tau : \tau \le t)$.

The class of transformations we consider have the general form

$$\psi(u, c) := u \times \varpi(u, c) \times I\big(|u| \le c\big), \tag{19}$$

where $\varpi(\cdot, c)$ is for each c a Borel function, and

$$\lim_{c \to \infty} \varpi(u, c) \times I(|u| \leq c) = 1. \tag{20}$$

Thus, $\psi(u, c)$ is a *redescending* function (see Andrews *et al.* [1] and Hampel *et al.* [27]). In the literature typically c is fixed, but the only way we can identify θ^0 and obtain Fischer consistency without an additional simulation step is to enforce $c \to \infty$ as $n \to \infty$.[4] Notice as $c \to \infty$ the transform satisfies $\psi(u, c) \to u$ hence it applies a negligible weight to u. Further, it operates similar to tail-trimming since by (20)

$$\psi(u, c) = uI(|u| \leq c) \times (1 + o(1)) \qquad \text{as } c \to \infty. \tag{21}$$

We focus on two types of weights ϖ. First, the simple trimming case $\psi(u, c) = uI(|u| \leq c)$, hence

$$\varpi(u, c) = 1.$$

The theory developed below easily extends to related redescending functions $\psi(u, c)$, like Hampel's three-part trimming function with thresholds $0 < a < b < c$ (see Andrews *et al.* [1]):

$$\begin{cases} u, & 0 \leq |u| \leq a, \\ a \times \text{sign}(u), & a < |u| \leq b, \\ \dfrac{a \times (c - |u|)}{c - b} \times \text{sign}(u), & b < |u| \leq c, \\ 0, & c < |u|. \end{cases}$$

This can be identically written as (19) with

$$\varpi(u, c) = I(|u| \leq a) + \frac{a}{|u|} \times I(a < |u| \leq b) + \frac{a(c - |u|)}{|u|(c - b)} \times I(b < |u| \leq c). \tag{22}$$

Of course, we abuse notation since there are three thresholds $\{a, b, c\}$. By construction $\varpi(u, c) \in [0, 1]$, while negligibility requires the smallest threshold $a \to \infty$, hence $\varpi(u, c) \to 1$ as $a \to \infty$.

Second, we use smooth weights $\varpi(u, c)$ that are continuously differentiable in c, with

$$\left| \frac{\partial}{\partial c} \varpi(u, c) \right| \times I(|u| \leq c) \leq K \frac{1}{c}. \tag{23}$$

Notice the simple trimming case $\varpi(u, c) = 1$ trivially satisfies (23). Thus, as $c \to \infty$ the transform derivative $(\partial / \partial c)\psi(u, c) \to 0$ at rate $O(1/c)$ for all $|u| \neq c$. An example is Tukey's bisquare $\varpi(u, c) = (1 - (u/c)^2)^2$ with $(\partial / \partial c)\varpi(u, c) = 2(1 - (u/c)^2)u^2/c^3$ hence (23) holds. A second example is the exponential $\varpi(u, c) = \exp\{-|u|/c\}$ with $(\partial / \partial c)\varpi(u, c) = \exp\{-|u|/c\}|u|/c^2$.

Assumption 4 (Redescending transforms). *Let* $\psi(u, c)$ *satisfy* (19), (20) *and* (23).

---

[4]See, for example, Sakata and White [58], Cantoni and Ronchetti [11] and Mancini *et al.* [43].

Now define two-tailed observations $\epsilon_t^{(a)}(\theta) := |\epsilon_t(\theta)|$ and their order statistics $\epsilon_{(1)}^{(a)}(\theta) \geq \epsilon_{(2)}^{(a)}(\theta) \geq \cdots$, and let $\{k_n\}$ be an intermediate order sequence. Write

$$\hat{I}_{n,t}^{(\epsilon)}(\theta) := I\big(\big|\epsilon_t(\theta)\big| \leq \epsilon_{(k_n)}^{(a)}(\theta)\big),$$

$$\hat{\psi}_{n,t}(\theta) := \psi\big(\epsilon_t(\theta), \epsilon_{(k_n)}^{(a)}(\theta)\big) = \epsilon_t(\theta) \times \varpi\big(\epsilon_t(\theta), \epsilon_{(k_n)}^{(a)}(\theta)\big)\hat{I}_{n,t}^{(\epsilon)}(\theta),$$

and define re-centered equations and a Method of Negligibly-Weighted Moments (MNWM) estimator

$$\widehat{m}_{n,t}(\theta) := \left(\hat{\psi}_{n,t}^2(\theta) - \frac{1}{n}\sum_{t=1}^{n}\hat{\psi}_{n,t}^2(\theta)\right) \times \mathfrak{s}_t(\theta) \qquad \text{and}$$

$$\hat{\theta}_n^{(m)} := \arg\min_{\theta \in \Theta}\left(\sum_{t=1}^{n}\widehat{m}_{n,t}(\theta)\right)'\left(\sum_{t=1}^{n}\widehat{m}_{n,t}(\theta)\right).$$

Any positive definite symmetric weight matrix $W \in \mathbb{R}^{q \times q}$ leads to the same solution $\arg\min_{\theta \in \Theta}\sum_{t=1}^{n}\widehat{m}_{n,t}(\theta)' \times W \times \sum_{t=1}^{n}\widehat{m}_{n,t}(\theta)$. Similarly, any $\mathfrak{I}_{t-1}$-measurable uniformly $L_{2+\iota}$-bounded vector $z_t(\theta) \in \mathbb{R}^r$, $r \geq q$, can be used instead of $\mathfrak{s}_t(\theta)$ for a GMM estimator (Hansen [28]). The scaled volatility derivative $\mathfrak{s}_t(\theta)$, however, provides an analogue to QML. Finally, as discussed in Section 2 small sample performance appears to be improved if we also trim by y_{t-1}, while asymptotics are unchanged if trimming is negligible. The estimator in this case uses the transformed error $\epsilon_t(\theta)\varpi(\epsilon_t(\theta), \epsilon_{(k_n)}^{(a)}(\theta))\hat{I}_{n,t}^{(\epsilon)}(\theta)\hat{I}_{n,t-1}^{(y)}$.

Next, for asymptotics let $\{\mathcal{C}_n(\theta)\}$ satisfy

$$P\big(\big|\epsilon_t(\theta)\big| \geq \mathcal{C}_n(\theta)\big) = \frac{k_n}{n},$$

write compactly

$$I_{n,t}^{(\epsilon)}(\theta) := I\big(\big|\epsilon_t(\theta)\big| \leq \epsilon_{(k_n)}^{(a)}(\theta)\big),$$

$$\psi_{n,t}(\theta) := \psi\big(\epsilon_t(\theta), \mathcal{C}_n(\theta)\big) \quad \text{and} \quad \epsilon_{n,t}(\theta) := \epsilon_t(\theta)I_{n,t}^{(\epsilon)}(\theta),$$

and define equations with non-random thresholds

$$\check{m}_{n,t}(\theta) := \big(\psi_{n,t}^2(\theta) - E\big[\psi_{n,t}^2(\theta)\big]\big) \times \big(\mathfrak{s}_t(\theta) - E\big[\mathfrak{s}_t(\theta)\big]\big).$$

In view of re-centering in $\widehat{m}_{n,t}(\theta)$ it can be shown that, asymptotically, $\check{m}_{n,t}$ and $\widehat{m}_{n,t}$ are interchangeable. See the Appendix.

Since ϵ_t is i.i.d. and has a smooth distribution, the transform is negligible in that $\psi_{n,t}(\theta) \overset{\text{a.s.}}{\to} \epsilon_t(\theta)$, and $\mathfrak{s}_t$ is $\mathfrak{I}_{t-1}$-measurable, it follows for all $n \geq N$ and some large $N \in \mathbb{N}$

$$E\big[\check{m}_{n,t}(\theta)|\mathfrak{I}_{t-1}\big] = 0 \quad \text{if and only if} \quad \theta = \theta^0,$$

hence an identification condition like Assumption 2 automatically holds. Similarly, by negligibility $\psi(u,c) = uI(|u| \leq c) \times (1 + o(1))$ as $c \to \infty$ and $E[\epsilon_{n,t}^2] \to 1$, hence by independence of

the errors

$$
\begin{aligned}
E\big[\check{m}_{n,t}\check{m}'_{n,t}\big] &= E\big((\psi(\epsilon_t,\mathcal{C}_n)^2 - E[\psi(\epsilon_t,\mathcal{C}_n)^2])^2\big) \times E\big[(\mathfrak{s}_t - E[\mathfrak{s}_t])(\mathfrak{s}_t - E[\mathfrak{s}_t])'\big] \\
&= E\big((\epsilon_{n,t}^2 - E[\epsilon_{n,t}^2])^2\big) \times E\big[(\mathfrak{s}_t - E[\mathfrak{s}_t])(\mathfrak{s}_t - E[\mathfrak{s}_t])'\big] \times (1 + o(1)) \\
&= (E[\epsilon_{n,t}^4] - 1) \times E\big[(\mathfrak{s}_t - E[\mathfrak{s}_t])(\mathfrak{s}_t - E[\mathfrak{s}_t])'\big] \times (1 + o(1)).
\end{aligned}
$$

The MNWM scale is therefore

$$
\mathring{\mathcal{V}}_n = \frac{n}{E[\epsilon_{n,t}^4] - 1} \times E\big[(\mathfrak{s}_t - E[\mathfrak{s}_t])(\mathfrak{s}_t - E[\mathfrak{s}_t])'\big], \tag{24}
$$

which is positive definite under Assumption 1.

Theorem 3.1 (MNWM). *Under Assumptions* 1 *and* 4 $\mathring{\mathcal{V}}_n^{1/2}(\hat{\theta}_n^{(m)} - \theta^0) \xrightarrow{d} N(0, I_q)$. *Further each* $\mathring{\mathcal{V}}_{i,i,n} \to \infty$ *and* $\mathring{\mathcal{V}}_{i,i,n}/\mathcal{V}_{i,i,n} \to (0,1)$.

Remark 7. In general a direct comparison of QMTTL and MNWM scales $\mathcal{V}_n$ and $\mathring{\mathcal{V}}_n$ is difficult for a particular n due to the different trimming strategies. Notice, however, that $E[\epsilon_{n,t}^4] - 1 = E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \times (1 + o(1))$ if $\mathcal{C}_n = (\mathcal{U}_n + 1)^{1/2}$. This follows by noting $E[\epsilon_t^2] = 1$, $\mathcal{E}_t^2 \in [-1, \infty)$, negligibility and dominated convergence imply

$$
\begin{aligned}
E\big[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}\big] &= E\big[(\epsilon_t^4 - 2\epsilon_t^2 + 1)I(-\mathcal{L}_n \le \epsilon_t^2 - 1 \le \mathcal{U}_n)\big] \\
&= E\big[\epsilon_t^4 I((1 - \mathcal{L}_n)^{1/2} \le |\epsilon_t| \le (\mathcal{U}_n + 1)^{1/2})\big] \times (1 + o(1)) \\
&= E\big[\epsilon_t^4 I(|\epsilon_t| \le (\mathcal{U}_n + 1)^{1/2})\big] \times (1 + o(1)).
\end{aligned}
$$

Thus, $\mathring{\mathcal{V}}_n \times \mathcal{V}_n^{-1} = E[(\mathfrak{s}_t - E[\mathfrak{s}_t])(\mathfrak{s}_t - E[\mathfrak{s}_t])'] \times E[\mathfrak{s}_t\mathfrak{s}'_t]$ as $n \to \infty$. Therefore $\mathring{\mathcal{V}}_n$ is smaller than $\mathcal{V}_n$ due to the centered term $E[(\mathfrak{s}_t - E[\mathfrak{s}_t])(\mathfrak{s}_t - E[\mathfrak{s}_t])']$, hence identification is assured at a cost of efficiency.

Remark 8. Since $\mathring{\mathcal{V}}_n \sim \mathcal{K}_n \mathcal{V}_n$ for some sequence of positive definite matrices $\{\mathcal{K}_n\}$, the Section 2.4 discourse on the QMTTL rate of convergence carries over here.

4. Inference

In view of $\mathcal{V}_n \sim n(E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}])^{-1} E[\mathfrak{s}_t\mathfrak{s}'_t]$, a natural estimator of the QMTTL scale $\mathcal{V}_n$ is

$$
\hat{\mathcal{V}}_n = \hat{\mathcal{V}}_n(\hat{\theta}_n) = n \times \frac{1}{1/n \sum_{t=1}^n 1/n \sum_{t=1}^n \mathcal{E}_t^2(\hat{\theta}_n)\hat{I}_{n,t}^{(\mathcal{E})}(\hat{\theta}_n)} \times \frac{1}{n}\sum_{t=1}^n \mathfrak{s}_t(\hat{\theta}_n)\mathfrak{s}'_t(\hat{\theta}_n). \tag{25}
$$

Theorem 4.1. *Under Assumptions* 1 *and* 2 $\hat{\mathcal{V}}_n = \mathcal{V}_n(1 + o_p(1))$.

Remark 9. Notice $\hat{\mathcal{V}}_n = \mathcal{V}_n(1 + o_p(1))$ only reduces to $\hat{\mathcal{V}}_n = \mathcal{V}_n + o_p(1)$ when $E[\epsilon_t^4] < \infty$. In general classic inference is available without knowing the true rate of convergence, nor even if trimming is required.

Remark 10. A consistent estimator of the MNWM scale $\mathring{\mathcal{V}}_n$ can similarly be constructed.

A Wald statistic naturally follows for a test of (non)linear parameter restrictions $R(\theta^0) = 0$ where $R : \mathbb{R}^q \to \mathbb{R}^J$ and $J \geq 1$. Assume R is differentiable with a gradient $\mathcal{D}(\theta) = (\partial/\partial\theta)R(\theta)$ that is continuous, differentiable and has full column rank. The test statistic with the QMTTL estimator as a plug-in is

$$\mathcal{W}_n = R(\hat{\theta}_n)'\big(\mathcal{D}(\hat{\theta}_n)\hat{\mathcal{V}}_n^{-1}(\hat{\theta}_n)\mathcal{D}(\hat{\theta}_n)'\big)^{-1}R(\hat{\theta}_n).$$

Use Theorems 2.2 and 4.1 to deduce $\mathcal{W}_n \xrightarrow{d} \chi^2(J)$ under the null, and if $R(\theta^0) \neq 0$ then $\mathcal{W}_n \xrightarrow{p} \infty$.

Similarly, the proof of Theorem 2.1 shows the QMTTL first order condition is $1/n \times \sum_{t=1}^n m_t(\hat{\theta}_n) I_{n,t}^{(\mathcal{E})}(\hat{\theta}_n) = 0$ a.s. This naturally suggests the possibility of a score or Lagrange Multiplier test since a QMTTL estimator under the constraint $R(\theta^0) = 0$, denoted $\hat{\theta}_n^{(c)}$, also satisfies $1/n \sum_{t=1}^n m_t(\hat{\theta}_n^{(c)}) I_{n,t}^{(\mathcal{E})}(\hat{\theta}_n^{(c)}) \xrightarrow{p} 0$ if the constraint is true. A heavy tail robust test of $R(\theta^0) = 0$ can therefore be coached as a tail-trimmed moment condition test as in Hill and Aguilar [33].

5. Simulation

We now compare our robust QML and Method of Moments estimators with various estimators in the literature. In order to draw the best comparisons between QMTTL and MNWM, we initially focus on simple trimming for MNWM. We compare our estimators to QML as a benchmark, as well as Log-LAD, Weighted Laplace QML (WLQML) and Power-Law QML (PQML) due to their heavy tail robustness properties. Finally, we investigate other redescending transforms as alternatives for MNWM, and whether tail-trimming can improve the small sample properties of PQML.

5.1. Data generation and estimators

Let P_κ denote a symmetric Pareto distribution: if ϵ_t is distributed P_κ then $P(\epsilon_t \leq -a) = P(\epsilon_t \geq a) = 0.5(1 + a)^{-\kappa}$ for $a > 0$. We draw $20n$ observations for $n \in \{100, 800\}$ from the GARCH process $y_t = \sigma_t\epsilon_t$ and $\sigma_t^2 = 0.05 + 0.05y_{t-1}^2 + 0.90\sigma_{t-1}^2$ with a starting value $\sigma_1^2 = 0.05$, and retain the last n observations for the sample. This is repeated to produce 10,000 samples $\{y_t\}_{t=1}^n$. Our choice of parameter values are indicative of values we obtain in the empirical study below, and frequently encountered in macroeconomic and financial data. The error ϵ_t is i.i.d. $N(0, 1)$, or $P_{2.5}$ standardized such that $E[\epsilon_t^2] = 1$.

We compute the feasible QMTTL and MNWM estimators conditional on the first observation, with parameter space is $\Theta = [\iota, 2] \times [\iota, 1 - \iota] \times [\iota, 1 - \iota]$ where $\iota = 10^{-10}$. The iterated volatility

variable is $h_1(\theta) = \omega$ and $h_t(\theta) = \omega + \alpha y_{t-1}^2 + \beta h_{t-1}(\theta)$ where we initialize $h_1(\theta) = \omega$ for QMTTL and $h_1^\theta(\theta) = [1, 0, 0]'$ for MNWM.

As a benchmark for QMTTL we use *strong asymmetric* trimming with error fractiles $k_{2,n} = \max\{1, [0.025n/\ln(n)]\}$ and $k_{1,n} = 35k_{2,n}$. This equates to $\{k_{1,n}, k_{2,n}\} = \{1, 35\}$ and $\{3, 105\}$ for $n = 100$ and 800. The fractile for trimming by y_{t-1} is $\tilde{k}_n = \max\{1, [0.1\ln(n)]\}$: asymptotics do not require such trimming, while removing a very few criterion equations due to large y_{t-1} appears to improve the estimator's performance. The benchmark for MNWM is simple trimming $\psi(u, c) = uI(|u| \le c)$. The error fractile is as above $k_n = \max\{1, [0.025n/\ln(n)]\}$ and the fractile for trimming by y_{t-1} is again $\tilde{k}_n$.

In addition to the benchmark estimates, we compute MNWM with Tukey's bisquare and exponential transforms. We also compute QMTTL with *weak asymmetric* ($k_{1,n} = 10k_{2,n}$) and *symmetric* ($k_{1,n} = k_{2,n}$) trimming. Recall from Section 2.3 that for QMTTL $k_{1,n} = 35k_{2,n}$ roughly minimizes bias in the Pareto case $P(|\epsilon_t| \ge a) = (1+a)^{-2.5}$ when $n = 100$. We show here that using $k_{1,n} = 35k_{2,n}$ even when $n = 800$ still promotes a sharp estimator. In simulations not reported here, we find that the bias minimizing relation $k_{1,n} = 100k_{2,n}$ when $P(|\epsilon_t| \ge a) = (1+a)^{-2.5}$ and $n = 800$ logically leads to even smaller bias, but bias is still low when $k_{1,n} = 35k_{2,n}$. Recall also that any combination $\{k_{1,n}, k_{2,n}\}$ works in the Gaussian case provided $k_{i,n} = \mathrm{o}(n^{1/2})$. This is violated here since we use $k_{i,n} \sim Kn/\ln(n)$, however this matters only asymptotically, and we demonstrate that using $k_{2,n} \sim Kn/\ln(n)$ and $k_{1,n} = 35k_{2,n}$ for $n = 100$ and 800 in the thin tail case still leads to a competitive estimator in small samples. Indeed, if we use $k_{i,n} \sim Kn^{1/2}/\ln(n)$ then the small sample performance is essentially identical to what we see here.

Peng and Yao's [52] Log-LAD criterion is $\sum_{t=2}^n |\ln y_t^2 - \ln h_t(\theta)|$. The WLQML criterion is $\sum_{t=2}^n \{\ln h_t^{1/2}(\theta) + |y_t/h_t^{1/2}(\theta)|\}w_t$ where we choose the weights $\{w_t\}$ as in Zhu and Ling [61], equation (2.4): $w_t = (\max\{1, C^{-1}\sum_{i=1}^\infty i^{-9}|y_{t-i}I(|y_{t-i}| > C)|\})^{-4}$ where $C = y_{(0.10n)}^{(a)}$ and $y_{t-i} = 0 \; \forall i \ge t$.

The PQML estimator detailed in Berkes and Horvath [5], Example 2.3, is based on the criterion $-\sum_{t=2}^n \ln(h_t^{-1/2}(\theta)f(y_t/h_t^{1/2}(\theta)))$ where $f(u) = K(1 + |u|)^{-\vartheta}$ with tail index $\vartheta > 1$. The value $K > 0$ ensures $\int_{-\infty}^\infty f(u)\,du = 1$ and of course is irrelevant for estimation, hence we simply set $K = 1$. Identification of θ^0 requires $E[|\epsilon_t|/(1 + |\epsilon_t|)] = 1/\vartheta$, while in the Pareto case $P(|\epsilon_t| \ge a) = (1 + a)^{-\kappa}$ it is easily verified that $E[|\epsilon_t|/(1 + |\epsilon_t|)] = 1/(\kappa + 1)$ hence we set $\vartheta = \kappa + 1 = 3.5$ in both Paretian and Gaussian cases.[5] We also set $\vartheta = 3$ as a control case to see if small sample bias increases when ϵ_t is Pareto, as it should.

5.2. Simulation results

Table 1 contains estimator bias, root mse [rmse], and the Kolmogorov–Smirnov statistic scaled by its 5% critical value. We only report results for θ_3^0 in order to conserve space, while the omitted results are qualitatively similar. In Table 2, we report t-test rejection frequencies for tests of the hypotheses $\theta_3^0 = 0.9$, $\theta_3^0 = 0.70$ and $\theta_3^0 = 0.50$, where the first is true. If $\{\hat{\theta}_{3,n}^{(r)}\}_{r=1}^R$ is the sequence

---

[5]Simply note $P(|\epsilon_t| \ge a) = (1 + a)^{-\kappa}$ implies $P(|\epsilon_t|/(1 + |\epsilon_t|) > a) = P(|\epsilon_t| > a/(1 - a)) = (1 - a)^\kappa$ hence $E[|\epsilon_t|/(1 + |\epsilon_t|)] = \int_0^1 P(|\epsilon_t|/(1 + |\epsilon_t|) > a)\,da = \int_0^1 (1 - a)^\kappa\,da = 1/(1 + \kappa)$.

Table 1. Simulation estimation results for θ_3^0

| | $\epsilon_t \sim \bar{P}_{2.5}$ | | | | | | $\epsilon_t \sim N(0,1)$ | | | | | |
| | $n=100$ | | | $n=800$ | | | $n=100$ | | | $n=800$ | | |
	Bias	RMS[a]	KS[b]	Bias	RMS	KS	Bias	RMS	KS	Bias	RMS	KS
QMTTL-SA[c]	−0.010	0.092	1.75	0.008	0.045	1.45	−0.063	0.095	3.87	0.001	0.030	1.07
QMTTL-WA	−0.031	0.102	3.01	0.024	0.038	2.76	−0.060	0.089	4.76	0.003	0.030	1.34
QMTTL-S	−0.041	0.114	4.69	0.016	0.044	4.21	−0.069	0.075	6.30	0.005	0.032	1.89
MNWM-I[d]	−0.023	0.111	2.31	−0.010	0.064	1.56	−0.029	0.108	2.86	−0.008	0.036	1.17
MNWM-T	−0.019	0.103	2.87	−0.012	0.069	1.61	0.021	0.113	3.16	−0.010	0.039	1.20
MNWM-E	−0.025	0.117	3.13	−0.016	0.058	1.50	−0.026	0.097	3.02	−0.013	0.037	1.31
WLQML[e]	−0.063	0.124	5.92	−0.135	0.107	7.64	−0.092	0.082	8.12	−0.088	0.084	6.05
$\text{WLQML}_{E|\epsilon_t|=1}$	−0.082	0.219	8.48	−0.072	0.089	5.64	−0.075	0.078	9.36	−0.065	0.067	3.97
PQML_3	−0.048	0.085	6.17	−0.039	0.059	3.00	−0.065	0.067	9.07	0.005	0.032	1.30
$\text{PQML}_{3.5}$	−0.034	0.083	4.74	−0.018	0.056	3.17	−0.064	0.062	9.54	0.009	0.029	2.75
$\text{PQMTTL}_{3.5}^{\text{SA}}$	−0.054	0.116	6.23	−0.017	0.056	2.23	−0.061	0.074	6.38	0.011	0.028	2.65
$\text{PQMTTL}_{3.5}^{\text{WA}}$	−0.031	0.074	4.28	−0.012	0.046	1.35	−0.051	0.074	8.21	0.008	0.028	2.12
$\text{PQMTTL}_{3.5}^{\text{S}}$	−0.027	0.077	4.05	−0.019	0.057	2.43	−0.055	0.069	8.30	0.011	0.027	2.76
Log-LAD	−0.217	0.165	9.88	−0.253	0.149	9.12	−0.082	0.100	7.01	−0.019	0.046	3.61
QML	−0.073	0.099	6.23	−0.054	0.078	4.65	−0.112	0.089	8.71	−0.013	0.034	1.64

[a]The square root of the empirical mean squared error.

[b]The Kolmogorov–Smirnov statistic divided by the 5% critical value: KS > 1 indicates rejection of normality at the 5% level.

[c]*Benchmark* QMTTL-SA (strong asymmetric) uses fractiles $k_{1,n} = 35k_{2,n}$; QMTTL-WA (weak asymmetric) uses $k_{1,n} = 10k_{2,n}$; QMTTL-S (symmetric) uses $k_{1,n} = k_{2,n}$.

[d]*Benchmark* MNWM-I uses the simple trimming function $\psi(u,c) = uI(|u| \leq c)$; MNWM-T and MNWM-E use Tukey's bisquare and exponential transforms.

[e]WLQML is Weighted Laplace QML. $\text{WLQML}_{E|\epsilon_t|=1}$ is WLQML for processes with $E|\epsilon_t| = 1$. PQML_ϑ is power-law QML with criterion index ϑ. $\text{PQMTTL}_\vartheta^{\text{WA}}$ and $\text{PQMTTL}_\vartheta^{\text{SA}}$ are tail-trimmed PQML with *weak asymmetric* ($k_{1,T} = 5k_{2,T}$) or *strong asymmetric* ($k_{1,T} = 9k_{2,T}$) trimming.

Table 2. Test rejection frequencies[a] at 5% level for θ_3^0

	$\epsilon_t \sim \bar{P}_{2.5}$						$\epsilon_t \sim N(0,1)$					
	$n=100$			$n=800$			$n=100$			$n=800$		
	H_0	H_1^1	H_1^2	H_0	H_1^1	H_1^2	H_0	H_1^1	H_1^2	H_0	H_1^1	H_1^2
QMTTL-SA[b]	0.054	0.694	0.995	0.046	0.951	0.999	0.059	0.664	0.789	0.048	1.00	1.00
QMTTL-WA	0.068	0.431	0.924	0.041	1.00	1.00	0.065	0.067	0.868	0.045	1.00	1.00
QMTTL-S	0.074	0.256	0.880	0.036	1.00	1.00	0.058	0.166	0.942	0.040	1.00	1.00
MNWM-I[c]	0.055	0.521	0.840	0.054	0.899	0.997	0.058	0.716	0.927	0.047	0.998	1.00
MNWM-T	0.043	0.791	0.981	0.055	0.878	0.991	0.031	0.963	0.998	0.054	0.992	1.00
MNWM-E	0.050	0.236	0.907	0.058	0.867	0.982	0.062	0.573	0.981	0.053	0.988	1.00
WLQML[d]	0.058	0.045	0.838	0.038	0.006	0.260	0.047	0.031	0.809	0.043	0.052	0.793
$\text{WLQML}_{E\|\epsilon_t\|=1}$	0.041	0.012	0.493	0.041	0.036	0.436	0.038	0.021	0.771	0.044	0.104	0.965
PQML_3	0.058	0.367	0.955	0.060	0.688	1.00	0.049	0.243	0.980	0.034	1.00	1.00
$\text{PQML}_{3.5}$	0.051	0.578	0.980	0.058	0.891	0.998	0.045	0.246	0.983	0.039	1.00	1.00
$\text{PQMTTL}_{3.5}^{SA}$	0.078	0.122	0.845	0.043	0.863	1.00	0.058	0.260	0.952	0.043	1.00	1.00
$\text{PQMTTL}_{3.5}^{WA}$	0.053	0.600	0.978	0.056	0.938	1.00	0.060	0.281	0.955	0.051	1.00	1.00
$\text{PQMTTL}_{3.5}^{S}$	0.053	0.662	0.978	0.055	0.888	1.00	0.057	0.385	0.972	0.040	1.00	1.00
Log-LAD	0.061	0.009	0.000	0.025	0.000	0.000	0.058	0.046	0.785	0.053	0.951	1.00
QML	0.065	0.109	0.789	0.061	0.342	0.733	0.061	0.000	0.575	0.051	0.997	1.00

[a]The hypotheses are H_0: $\theta_3 = \theta_3^0$, H_1^1: $\theta_3 = \theta_3^0 - 0.2$, and H_1^2: $\theta_3 = \theta_3^0 - 0.4$, where $\theta_3^0 = 0.9$.

[b]*Benchmark* QMTTL-SA (strong asymmetric) uses fractiles $k_{1,n} = 35k_{2,n}$; QMTTL-WA (weak asymmetric) uses $k_{1,n} = 10k_{2,n}$; QMTTL-S (symmetric) uses $k_{1,n} = k_{2,n}$.

[c]*Benchmark* MNWM-I uses the simple trimming function $\psi(u, c) = uI(|u| \leq c)$; MNWM-T and MNWM-E use Tukey's bisquare and exponential transforms.

[d]WLQML is Weighted Laplace QML. $\text{WLQML}_{E|\epsilon_t|=1}$ is WLQML for processes with $E|\epsilon_t| = 1$. PQML_ϑ is power-law QML with criterion index ϑ. $\text{PQMTTL}_\vartheta^{WA}$ and $\text{PQMTTL}_\vartheta^{SA}$ are tail-trimmed PQML with *weak asymmetric* ($k_{1,T} = 5k_{2,T}$) or *strong asymmetric* ($k_{1,T} = 9k_{2,T}$) trimming.

of $R = 10,000$ independent estimates of θ_3^0, we use the empirical variance $1/R \sum_{r=1}^{R} (\hat{\theta}_{3,n}^{(r)} - 1/R \sum_{r=1}^{R} \hat{\theta}_{3,n}^{(r)})^2$ to standardize $\hat{\theta}_{3,n}^{(r)}$ for KS test and t-test computation.

Log-LAD and WLQML perform poorly when $E[\epsilon_t^4] = \infty$: in small samples they are sensitive to large error observations, contrary to their theoretical robustness properties asymptotically. Indeed, Log-LAD leads to exceptionally poor inference when $E[\epsilon_t^4] = \infty$ due to a high degree of bias, and is worst overall. Further, WLQML is sensitive to large errors even in the Gaussian case. It is not surprising that Log-LAD and WLQML are similar since Laplace QML merely generalizes LAD to a likelihood framework (Zhu and Ling [61]). QML performs better than Log-LAD and worse than WLQML when $E[\epsilon_t^4] = \infty$, and is better than both when ϵ_t is normal.

PQML is more promising than QML, Log-LAD and WLQML. It performs better on all measures and in nearly every case: Log-LAD and WLQML are closer to normally distributed for Gaussian ϵ_t with small $n = 100$. In particular, PQML has the smallest rmse of all estimators in this study, suggesting that it exhibits very low empirical variance since it has higher bias than QMTTL and MNWM. Identification is assured in the Pareto case $\kappa = 2.5$ when $\vartheta = 3.5$, so it is not surprising that bias in the Pareto case is higher when $\vartheta = 3$. Further, there should be noticeable bias in the Gaussian case since identification fails, yet bias is actually *smaller* than for Paretian errors when $n = 800$. It is important to stress that PQML with index $\vartheta = 3.5$ is perfectly suited for our Paretian case $P(|\epsilon_t| \geq a) = (1 + a)^{-2.5}$ since this non-Gaussian QML leads to identification and therefore Fischer consistency. However, even this estimator exhibits more bias than QMTTL and MNWM evidently due to the adverse effects of sample error extremes (see Section 5.3).

The best estimators in this study are QMTTL (with strong asymmetric trimming) and MNWM in terms of bias, approximate normality and test performance, while only PQML has a smaller rmse. QMTTL with strong asymmetric trimming ($k_{1,n} = 35k_{2,n}$), as required in the Paretian case when $n = 100$, is superb when ϵ_t is Paretian for either $n \in \{100, 800\}$, and works very well in the Gaussian case with a rmse close to PQML. Overall, *QMTTL with strong asymmetric trimming is the best estimator* since it beats MNWM in terms of bias and approximate normality in nearly every case and has a small rmse in all cases.

QMTTL with weak asymmetric ($k_{1,n} = 10k_{2,n}$) or symmetric ($k_{1,n} = k_{2,n}$) trimming lead to greater bias when ϵ_t is Paretian, and to negligible bias when ϵ_t is Gaussian, in each case as this estimator should. Nevertheless, QMTTL with weak asymmetric or symmetric trimming is superior to QML, Log-LAD, and WLQML by all measure; QMTTL with weak asymmetric trimming beats PQML by all measures except rmse; and QMTTL with symmetric trimming beats PQML when $n = 800$. Our QMTTL simulations strongly point to the use of strong asymmetric trimming in general since it is valid for thin tailed errors, and necessary for heavy tailed errors. They also reveal that using weak asymmetric of symmetric trimming still leads to a competitive estimator.

Further, re-centering after trimming in the MNWM estimator in general leads to higher mean-squared-error than QMTTL. Recall this estimator may be less efficient than QMTTL, and QMTTL with strong asymmetric trimming results in the lowest bias of all estimators in this study. Nevertheless, MNWM works well, with the second smallest bias, and overall is closer to normal than all estimators save QMTTL with strong asymmetric trimming. As discussed in Section 2.3.3, the preferred estimator depends on the analyst's agenda: MNWM is always asymp-

totically unbiased with symmetric trimming which is easy to implement, while QMTTL performs better in small samples.

5.3. Addtional experiments for WLQML and PQML

We now perform two additional experiments. First, recall WLQML requires $E|\epsilon_t| = 1$ which does not hold for either Paretian or Gaussian errors in this study. We now standardize ϵ_t such that $E|\epsilon_t| = 1$ to see if ensuring identification helps in small samples. The results are nevertheless qualitatively similar whether $E[\epsilon_t^2] = 1$ and $E|\epsilon_t| \neq 1$, or $E|\epsilon_t| = 1$, is true. See Tables 1 and 2. In fact, for heavy tailed errors WLQML actually performs *worse* in terms of bias and approximate normality when identification is assured. Further, inference is still quite poor in many cases. This suggests the previous poor performance of WLQML is not due to the identification condition failing to hold.

Second, recall that QMTTL has lower bias and is closer to normally distributed that other estimators whether trimming is needed or not. We therefore tail trim the PQML criterion to see if the benefits of trimming carry over to non-Gaussian QML. Recall PQML with index $\vartheta > 1$ has the identification condition $E[u_t] = 0$ where $u_t := |\epsilon_t|/(1 + |\epsilon_t|) - 1/\vartheta$. Define $u_t^{(-)}(\theta) := u_t(\theta)I(u_t(\theta) < 0)$ and $u_t^{(+)}(\theta) := u_t(\theta)I(u_t(\theta) \geq 0)$ and their order statistics $u_{(1)}^{(-)}(\theta) \leq \cdots \leq u_{(n)}^{(-)}(\theta) \leq 0$ and $u_{(1)}^{(+)}(\theta) \geq \cdots \geq u_{(n)}^{(+)}(\theta) \geq 0$. Let $\{k_{1,n}^{(u)}, k_{2,n}^{(u)}\}$ be intermediate order sequences and let $\{c_{1,n}^{(u)}, c_{2,n}^{(u)}\}$ be positive sequences satisfying $P(u_t(\theta) \leq -c_{1,n}^{(u)}) = k_{1,n}/n$ and $P(u_t(\theta) \geq c_{2,n}^{(u)}) = k_{2,n}/n$. The tail-trimmed PQML (PQMTTL) criterion is $-\sum_{t=2}^{n} \ln(h_t^{-1/2}(\theta)\{1 + |y_t/h_t^{1/2}(\theta)|\}^{-(\kappa+1)})I(u_{(k_{1,n}^{(u)})}^{(-)}(\theta) \leq u_t(\theta) \leq u_{(k_{2,n}^{(u)})}^{(+)}(\theta))$.

If ϵ_t is Paretian $P(|\epsilon_t| \geq a) = (1 + a)^{-2.5}$ it is straightforward to show $k_{1,n}^{(u)} = 5k_{2,n}^{(u)}$ when $n = 100$ and $k_{1,n}^{(u)} = 9k_{2,n}^{(u)}$ when $n = 800$ renders roughly $E[u_t I(-c_{1,n}^{(u)} \leq u_t(\theta) \leq c_{2,n}^{(u)})] = 0$. We therefore set *symmetric* ($k_{1,n}^{(u)} = k_{2,n}^{(u)}$), *weak asymmetric* ($k_{1,n}^{(u)} = 5k_{2,n}^{(u)}$) or *strong asymmetric* ($k_{1,n}^{(u)} = 9k_{2,n}^{(u)}$) trimming with $k_{2,n}^{(u)} = \max\{1, [0.025n/\ln(n)]\}$. Tables 1 and 2 show PQMTTL with weak asymmetric trimming performs better than PQML in all cases. If we use strong asymmetric trimming then the over-trimming for $n = 100$ leads to greater bias, but when $n = 800$ the estimator works well as it should, in particular it is closer to normal and therefore has better inference than PQML. Conversely, symmetric trimming leads to greater bias when $n = 800$ as it should. QMTTL with strong asymmetric trimming and MNWM with simple or exponential trimming are better than PQMTTL in terms of bias and approximate normality in most cases. Consider when $n = 800$ then in the Pareto case PQMTTL with weak asymmetric trimming is marginally closer to normal and slightly more biased than QMTTL, and in the Gaussian case PQMTTL is slightly less biased and farther from normally distributed than QMTTL. Overall tail-trimming seems to matter even for an inherently heavy tail robust non-Gaussian QML estimator.

6. Empirical application

Finally, we apply our estimators to asset returns series generated from the London Stock Exchange (FTSE-100), the NASDAQ composite index (IXIC), and the Hang Seng Index. The period

is Jan. 1, 2008–Dec. 31, 2010, representing 757, 757 and 756 daily observations respectively, net of market closures. We use log-returns $y_t = \ln(x_t/x_{t-1})$ where x_t is the daily open/close average of each index.[6]

As in Section 5, we compute MNWM using simple trimming denoted "I", Tukey's bisquare and exponential transforms, with fractiles $k_n = \max\{1, [0.025n/\ln(n)]\}$ and $\tilde{k}_n = \max\{1, [0.1 \ln(n)]\}$ for trimming by ϵ_t and y_{t-1}, respectively. Similarly, QMTTL is computed using *strong asymmetric* $(k_{1,n} = 35k_{2,n})$, *weak asymmetric* $(k_{1,n} = 10k_{2,n})$, and *symmetric* $(k_{1,n} = k_{2,n})$ error fractiles denoted "SA", "WA" and "S", with $k_{2,n} = \max\{1, [0.025n/\ln(n)]\}$, and $\tilde{k}_n$ for y_{t-1}. The parameter space is $\Theta = [\iota, 2] \times [\iota, 1 - \iota] \times [\iota, 1 - \iota]$ where $\iota = 10^{-10}$.

See Table 3 for estimation details where standard errors are computed using (25) for QMTTL and its logical extension for MNWM. In each case a GARCH model fits well, while QMTTL and MNWM produce qualitatively similar estimates. The various MNWM estimates are similar across transform type, especially exponential and simple trimming versions. The QMTTL estimates are somewhat similar across asymmetric and symmetric trimming. For example, evidence for IGARCH or explosive GARCH $\hat{\alpha}_n + \hat{\beta}_n \geq 1$ exists only for the NASDAQ based on QMTTL-SA and MNWM-I, while QMTTL-WA and QMTTL-S lead to smaller values. However, in all cases $\hat{\beta}_n$ is near 0.9 and $\hat{\alpha}_n$ is near 0.05, in many cases $\hat{\alpha}_n + \hat{\beta}_n \approx 1$, and for each series the various estimates are quite similar. The latter suggests the various asymmetric and symmetric trimming strategies for QMTTL work as well as inherently asymptotically unbiased MNWM. This is matched by our simulations where $n = 800$ aligns with the sample sizes in the present empirical study: strong asymmetric trimming leads to the best QMTTL results when ϵ_t has power law tails with a small index κ, but each trimming strategy leads to similar results, especially when $n = 800$.

7. Conclusion

We develop tail-trimmed QML and Method of Moments estimators for GARCH models with possibly heavy tailed errors ϵ_t that satisfy $E[\epsilon_t^2] = 1$. In the Method of Moments case, the model errors are first negligibly transformed with a redescending function, and then re-centered to control for small sample bias induced by the transform. We show by Monte Carlo experiment that tail-trimming within a QML framework dominates QML, Log-LAD and Weighted Laplace QML based on bias, mean-squared-error, approximate normality, and inference, and trumps Power-Law QML in all aspects except variance (Power-Law QML has higher bias yet lower mean-squared-error). Only QMTTL and MNWM directly counter the negative influence of large errors in small *and* large samples. Indeed, we show trimming leads to a better infeasible Power-Law QML estimator in small samples. The next stage must involve a theoretical development of data-dependent or automatic fractile selection, including possibly bootstrap and covariance determinant methods. This is left for future research.

---

[6]The data were obtained from http://finance.yahoo.com, and the open/close average is computed using the reported adjusted close values.

Table 3. QMTTL and MNWM estimation results for financial returns

	ω	α	β	ω	α	β
	QMTTL-SA[a]: $k_{1,n} = 35k_{2,n}$			MNWM-I[b]		
NASDAQ[c]	0.029 (0.031)[c]	0.113 (0.082)	0.893 (0.069)	0.016 (0.008)	0.117 (0.017)	0.884 (0.018)
HSI[d]	0.058 (0.064)	0.106 (0.151)	0.878 (0.252)	0.020 (0.009)	0.078 (0.013)	0.915 (0.015)
LSE	0.066 (0.083)	0.213 (0.156)	0.743 (0.224)	0.025 (0.006)	0.119 (0.015)	0.822 (0.020)
	QMTTL-WA: $k_{1,n} = 10k_{2,n}$			MNWM-E		
NASDAQ	0.017 (0.038)	0.138 (0.113)	0.849 (0.135)	0.032 (0.010)	0.102 (0.021)	0.886 (0.019)
HSI	0.046 (0.086)	0.082 (0.142)	0.910 (0.211)	0.021 (0.011)	0.078 (0.016)	0.915 (0.028)
LSE	0.022 (0.065)	0.179 (0.134)	0.805 (0.194)	0.030 (0.011)	0.125 (0.019)	0.824 (0.031)
	QMTTL-S $k_{1,n} = k_{2,n}$			MNWM-T		
NASDAQ	0.012 (0.033)	0.171 (0.145)	0.839 (0.125)	0.033 (0.011)	0.095 (0.031)	0.901 (0.034)
HSI	0.065 (0.092)	0.092 (0.123)	0.887 (0.163)	0.039 (0.011)	0.076 (0.017)	0.920 (0.027)
LSE	0.034 (0.076)	0.214 (0.189)	0.752 (0.203)	0.058 (0.016)	0.122 (0.021)	0.839 (0.024)

[a]SA = strong asymmetry; WA = weak asymmetry; S = symmetric.

[b]I = simple trimming, T = Tukey's bisquare, E = exponential.

[c]Standard errors are in parentheses ($\cdot$).

[d]HSI = Hang Seng; LSE = London Stock Exchange.

Appendix: Proofs of main theorems

Recall $\mathcal{E}_t(\theta) := \epsilon_t^2(\theta) - 1$, and score and Jacobian equations are $m_t(\theta) = \mathcal{E}_t(\theta)\mathfrak{s}_t(\theta)$ and $G_t(\theta) = \mathcal{E}_t(\theta)\mathfrak{d}_t(\theta) - \epsilon_t^2(\theta)\mathfrak{s}_t(\theta)\mathfrak{s}_t(\theta)'$ where

$$\mathfrak{s}_t(\theta) := \frac{1}{\sigma_t^2(\theta)}\frac{\partial}{\partial\theta}\sigma_t^2(\theta) \quad \text{and} \quad \mathfrak{d}_t(\theta) := \frac{\partial}{\partial\theta}\mathfrak{s}_t(\theta).$$

Define indicators, trimmed score equations and corresponding covariance and Jacobian matrices:

$$\hat{m}_{n,t}(\theta) := m_t(\theta)\hat{I}_{n,t}^{(\mathcal{E})}(\theta) \quad \text{and} \quad m_{n,t}(\theta) := m_t(\theta)I_{n,t}^{(\mathcal{E})}(\theta),$$

$$\Sigma_n(\theta) := E\big[m_{n,t}(\theta)m_{n,t}(\theta)'\big] \quad \text{and}$$

$$\mathcal{G}(\theta) := -E\big[\mathfrak{s}_t(\theta)\mathfrak{s}_t(\theta)'\big] \quad \text{and} \quad \mathcal{V}_n(\theta) = n\mathcal{G}(\theta)'\Sigma_n^{-1}(\theta)\mathcal{G}(\theta),$$

$$\mathcal{S}_n(\theta) := \frac{1}{n}E\left[\left(\sum_{t=1}^n m_{n,t}(\theta)\right)\left(\sum_{t=1}^n m_{n,t}(\theta)'\right)\right],$$

$$\widehat{\mathcal{G}}_n(\theta) := \frac{1}{n}\sum_{t=1}^n G_t(\theta)\hat{I}_{n,t}^{(\mathcal{E})}(\theta) \quad \text{and} \quad \check{\mathcal{G}}_n(\theta) := \frac{1}{n}\sum_{t=1}^n G_t(\theta)I_{n,t}^{(\mathcal{E})}(\theta).$$

By independence and identification Assumption 2 $\mathcal{S}_n = \Sigma_n \times (1 + \text{o}(1))$.

We implicitly assume all functions in this paper satisfy Pollard's ([54], Appendix C) permissibility criteria, the measure space that governs all random variables in this paper is complete, and therefore all majorants are measurable. Cf. Dudley [19]. Probability statements are therefore with respect to outer probability, and expectations over majorants are outer expectations.

A.1. Theorems 2.1 and 2.2

The proofs of QMTTL consistency and asymptotic normality Theorems 2.1 and 2.2 require supporting lemmas. We state them when required and provide proofs in Appendix A.3. Consistency requires bounding $\sum_{t=1}^n\{\hat{m}_{n,t}(\theta) - m_{n,t}(\theta)\}$, variance bounds, and laws of large numbers. Unless otherwise noted, Assumptions 1 and 2 hold.

Lemma A.1 (Asymptotic approximation). (a) $n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^n\{\hat{m}_{n,t} - m_{n,t}\} = \text{o}_p(1)$; (b) $\sup_{\theta\in\Theta}\|1/n\sum_{t=1}^n\{\hat{m}_{n,t}(\theta) - m_{n,t}(\theta)\}\| = \text{o}_p(\sup_{\theta\in\Theta}E\|m_{n,t}(\theta)\|)$.

Lemma A.2 (Variance bounds). *Under Assumption* 1(a) $\Sigma_n = \text{o}(n/\ln(n))$; (b) $\mathcal{S}_n = \text{o}(n/\ln(n))$.

Remark 11. Under Assumption 2 $\mathcal{S}_n = \Sigma_n(1 + \text{o}(1))$ hence then (b) follows from (a).

Lemma A.3 (LLN and ULLN). (a) $1/n\sum_{t=1}^n m_{n,t} = \text{o}_p(1)$; (b) $\sup_{\theta\in\Theta}\{\|1/n\sum_{t=1}^n m_{n,t}(\theta) - E[m_{n,t}(\theta)]\|\} = \text{o}_p(\sup_{\theta\in\Theta}E\|m_{n,t}(\theta)\|)$.

Asymptotic normality requires an expansion, central limit theorem, and Jacobian consistency.

Lemma A.4 (Asymptotic expansion). *Let* $\{\theta_n\}$ *and* $\{\tilde{\theta}_n\}$ *be any sequences of random variables in* Θ *with probability limit* θ^0. *Let* $\theta_{n,*} \in \Theta$ *satisfy* $\|\theta_{n,*} - \theta_n\| \le \|\theta_n - \tilde{\theta}_n\|$ *which may be different in difference places.* (a) $1/n \sum_{t=1}^n \{m_{n,t}(\theta_n) - m_{n,t}(\tilde{\theta}_n)\} = \breve{\mathcal{G}}_n(\theta_{n,*}) \times (\theta_n - \tilde{\theta}_n) \times (1 + o_p(1))4$; (b) $1/n \sum_{t=1}^n \{\hat{m}_{n,t}(\theta_n) - \hat{m}_{n,t}(\tilde{\theta}_n)\} = \widehat{\mathcal{G}}_n(\theta_{n,*}) \times (\theta_n - \tilde{\theta}_n) \times (1 + o_p(1))$.

Lemma A.5 (CLT). $n^{-1/2} \Sigma_n^{-1/2} \sum_{t=1}^n m_{n,t} \xrightarrow{d} N(0, I_q)$.

Lemma A.6 (Jacobian). (a) $\widehat{\mathcal{G}}_n(\hat{\theta}_n^*) = \mathcal{G} \times (1 + o_p(1))$ *and* $\breve{\mathcal{G}}_n \hat{\theta}_n^* = \mathcal{G} \times (1 + o_p(1))$ *for any* $\hat{\theta}_n^* \xrightarrow{p} \theta^0$; (b) $1/n \sum_{t=1}^n \mathfrak{s}_t(\hat{\theta}_n) \mathfrak{s}_t'(\hat{\theta}_n) = \mathcal{G} \times (1 + o_p(1))$; (c) $(\partial/\partial\theta) E[m_{n,t}(\theta)]|_{\theta^0} = \mathcal{G} \times (1 + o(1))$; (d) $\limsup_{n\to\infty} \sup_{\theta\in\Theta} E\|m_{n,t}(\theta)\| \le K\|\mathcal{G}\| \times (1 + o(1))$.

We are now ready to prove Theorems 2.1 and 2.2.

Proof of Theorem 2.1. Define $\hat{m}_n(\theta) := 1/n \sum_{t=1}^n \hat{m}_{n,t}(\theta)$, $m_n(\theta) := 1/n \sum_{t=1}^n m_{n,t}(\theta)$, $\mathcal{M}_n(\theta) := E[m_{n,t}(\theta)]$ and $\mathfrak{e}_n := \sup_{\theta\in\Theta} E\|m_{n,t}(\theta)\|$. We use an argument in Pakes and Pollard [50], pages 1038–1039.

Step 1. We first prove a required inequality:

$$\epsilon(\delta) := \liminf_{n\to\infty} \inf_{\theta\in\Theta:\|\theta-\theta^0\|>\delta} \{\|\mathcal{M}_n(\theta)\|/\mathfrak{e}_n\} > 0 \qquad \text{for any small } \delta > 0. \qquad (A.1)$$

Note $E[m_{n,t}] \to E[\mathcal{E}_t \mathfrak{s}_t] = 0$ by dominated convergence and independence. By the definition of a derivative and Lemma A.6(c) we have $E[m_{n,t}(\theta)] = \mathcal{G} \times (\theta - \theta^0) \times (1 + o(1))$ where $\mathcal{G} = -E[\mathfrak{s}_t \mathfrak{s}_t']$, and bound Lemma A.6(d) states $\mathfrak{e}_n := \sup_{\theta\in\Theta} E\|m_{n,t}(\theta)\| \le K\|\mathcal{G}\| \times (1 + o(1))$. It therefore follows for every $n \ge N$ and $\delta > 0$

$$\inf_{\|\theta-\theta^0\|>\delta} \{\mathfrak{e}_n^{-1} \|E[m_{n,t}(\theta)]\|\} \ge K \inf_{\|\theta-\theta^0\|>\delta} \left\{ \left\| \frac{\mathcal{G}}{\|\mathcal{G}\|} \times (\theta - \theta^0) \right\| \right\} \times (1 + o(1)) > 0.$$

Step 2. In view of (A.1) we have $P(\|\hat{\theta}_n - \theta^0\| > \delta) \le P(\|\mathcal{M}_n(\hat{\theta}_n)\|/\mathfrak{e}_n > \epsilon(\delta))$, hence it suffices to show $\|\mathcal{M}_n(\hat{\theta}_n)\|/\mathfrak{e}_n = o_p(1)$ in order to prove $\hat{\theta}_n \xrightarrow{p} \theta^0$. By Minkowski's inequality

$$\|\mathcal{M}_n(\hat{\theta}_n)\|/\mathfrak{e}_n \le \|\hat{m}_n(\hat{\theta}_n)\|/\mathfrak{e}_n + \|\hat{m}_n(\hat{\theta}_n) - \mathcal{M}_n(\hat{\theta}_n)\|/\mathfrak{e}_n = \mathcal{A}_{1,n}(\hat{\theta}_n) + \mathcal{A}_{2,n}(\hat{\theta}_n),$$

say. The proof is complete if we show $\mathcal{A}_{1,n}(\hat{\theta}_n)$ and $\mathcal{A}_{2,n}(\hat{\theta}_n)$ are $o_p(1)$.

Consider $\mathcal{A}_{1,n}(\hat{\theta}_n)$. We exploit theory developed in Cizek [16], Lemma 2.1, page 29. By distribution continuity and linearity of the volatility process $\{\sigma_t^2\}$, $\hat{Q}_n(\theta) := 1/n \sum_{t=1}^n (\ln \sigma_t^2(\theta) + y_t^2/\sigma_t^2(\theta)) \hat{I}_{n,t}^{(\mathcal{E})}(\theta)$ is almost surely twice differentiable at $\hat{\theta}_n$. In particular, up to a scalar constant $(\partial/\partial\theta) \hat{Q}_n(\theta)|_{\hat{\theta}_n} = \hat{m}_n(\hat{\theta}_n)$ a.s. By $\hat{\theta}_n$ a minimum $\hat{Q}_n(\hat{\theta}_n) \le \hat{Q}_n(\theta) \forall \in \Theta$ it follows $\|\hat{m}_n(\hat{\theta}_n)\| = 0$ a.s., while $\liminf_{n\to\infty} \mathfrak{e}_n > 0$ by distribution non-degeneracy and trimming negligibility, hence $\mathcal{A}_{1,n}(\hat{\theta}_n) = 0$ a.s.

Next $\mathcal{A}_{2,n}(\hat{\theta}_n)$. By Lemma A.1(b) $\sup_{\theta \in \Theta} \|\hat{m}_n(\theta) - m_n(\theta)\|/\mathfrak{e}_n = o_p(1)$, and $\sup_{\theta \in \Theta} \|m_n(\theta) - \mathcal{M}_n(\theta)\|/\mathfrak{e}_n = o_p(1)$ by ULLN Lemma A.3(b). Hence

$$\sup_{\theta \in \Theta}\{\mathcal{A}_{2,n}(\theta)\} \leq \sup_{\theta \in \Theta}\frac{\|\hat{m}_n(\theta) - m_n(\theta)\|}{\mathfrak{e}_n} + \sup_{\theta \in \Theta}\frac{\|m_n(\theta) - \mathcal{M}_n(\theta)\|}{\mathfrak{e}_n} = o_p(1). \qquad \square$$

Proof of Theorem 2.2. Use $1/n \sum_{t=1}^n \hat{m}_{n,t}(\hat{\theta}_n) = 0$ a.s. by the proof of Theorem 2.1, and expansion Lemma A.4(b) to deduce for some $\hat{\theta}_n^*$, $\|\hat{\theta}_n^* - \theta^0\| \leq \|\hat{\theta}_n - \theta^0\|$:

$$\widehat{\mathcal{G}}_n(\hat{\theta}_n^*)(\hat{\theta}_n - \theta^0)(1 + o_p(1)) + \frac{1}{n}\sum_{t=1}^n \hat{m}_{n,t} = 0 \qquad \text{a.s.} \qquad (A.2)$$

Consistency $\|\hat{\theta}_n^* - \theta^0\| \leq \|\hat{\theta}_n - \theta^0\| \xrightarrow{p} 0$ by Theorem 2.1 ensures $\widehat{\mathcal{G}}_n(\hat{\theta}_n^*) = \mathcal{G}(1 + o_p(1))$ by Lemma A.6(a). Multiply both sides of (A.2) by $n^{1/2}\Sigma_n^{-1/2}$, rearrange terms and use $\mathcal{V}_n = n\mathcal{G}'\Sigma_n^{-1}\mathcal{G}$ to deduce $\mathcal{V}_n^{1/2}(\hat{\theta}_n - \theta^0) = -n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^n \hat{m}_{n,t} \times (1 + o_p(1))$. In view of $n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^n\{\hat{m}_{n,t} - m_{n,t}\} = o_p(1)$ by Lemma A.1(a), we have

$$\mathcal{V}_n^{1/2}(\hat{\theta}_n - \theta^0) = -\Sigma_n^{-1/2}\frac{1}{n^{1/2}}\sum_{t=1}^n m_{n,t} \times (1 + o_p(1)),$$

hence $\mathcal{V}_n^{1/2}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N(0, I_q)$ by Lemma A.5. Finally $\mathcal{V}_{i,i,n} \to \infty$ follows from the fact that $\|\mathcal{G}\| > 0$, and $\|n\Sigma_n^{-1}\| \to \infty$ by Lemma A.2(a). $\qquad \square$

A.2. Remaining theorems

Define $h_t^\theta(\theta) := (\partial/\partial\theta)h_t(\theta)$ and $h_t^{\theta,\theta}(\theta) := (\partial/\partial\theta)h_t^\theta(\theta)$. We require stationary solutions $\{h_t^*(\theta), h_{i,t}^{*\theta}(\theta), h_{i,t}^{*\theta,\theta}(\theta)\}$ of the volatility process $\{h_t(\theta), h_{i,t}^\theta(\theta), h_{i,t}^{\theta,\theta}(\theta)\}$ in order to prove the asymptotic equivalence of the infeasible and feasible QMTTL estimators.

Let $\{\mathfrak{s}_t^*(\theta), \mathfrak{d}_t^*(\theta)\}$ denote $\{\mathfrak{s}_t(\theta), \mathfrak{d}_t(\theta)\}$ evaluated with $\{h_t^*(\theta), h_{i,t}^{*\theta}(\theta), h_{i,t}^{*\theta,\theta}(\theta)\}$. Define error and volatility derivatives evaluated at $\{h_t(\theta), h_t^\theta(\theta), h_t^{\theta,\theta}(\theta)\}$

$$\tilde{\epsilon}_t(\theta) := \frac{y_t}{\sqrt{h_t(\theta)}}, \qquad \tilde{\mathcal{E}}_t(\theta) := \tilde{\epsilon}_t^2(\theta) - 1,$$

$$\tilde{\mathfrak{s}}_t(\theta) := \frac{1}{h_t(\theta)}\frac{\partial}{\partial\theta}h_t(\theta) \quad \text{and} \quad \tilde{\mathfrak{d}}_t(\theta) = \frac{\partial}{\partial\theta}\tilde{\mathfrak{s}}_t(\theta),$$

$$\tilde{m}_t(\theta) := \tilde{\mathcal{E}}_t(\theta)\tilde{\mathfrak{s}}_t(\theta), \qquad \tilde{G}_t(\theta) := \frac{\partial}{\partial\theta}\tilde{m}_t(\theta) \quad \text{and} \quad \tilde{\mathcal{G}} := -E[\tilde{\mathfrak{s}}_t(\theta)\tilde{\mathfrak{s}}_t'(\theta)].$$

Define $\widehat{\tilde{I}}_{n,t}^{(\mathcal{E})}(\theta) := I(\tilde{\mathcal{E}}_{(k_{1,n})}^{(-)}(\theta) \leq \tilde{\mathcal{E}}_t(\theta) \leq \tilde{\mathcal{E}}_{(k_{2,n})}^{(+)}(\theta))$ and let $\{\tilde{\mathcal{L}}_n(\theta), \tilde{\mathcal{U}}_n(\theta)\}$ satisfy $P(\tilde{\mathcal{E}}_t(\theta) \leq -\tilde{\mathcal{L}}_n(\theta)) = k_{1,n}/n$ and $P(\tilde{\mathcal{E}}_t(\theta) \geq \tilde{\mathcal{U}}_n(\theta)) = k_{2,n}/n$. Similarly $\tilde{I}_{n,t}^{(\mathcal{E})}(\theta) := I(-\tilde{\mathcal{L}}_n(\theta) \leq \tilde{\mathcal{E}}_t(\theta) \leq \tilde{\mathcal{U}}_n(\theta))$. Define trimmed variants $\widehat{\tilde{m}}_{n,t}(\theta) := \tilde{m}_t(\theta)\widehat{\tilde{I}}_{n,t}^{(\epsilon)}(\theta)$ and $\tilde{m}_{n,t}(\theta) := \tilde{m}_t(\theta)\tilde{I}_{n,t}^{(\epsilon)}(\theta)$.

Lemma A.7 (Stationary solution). *Let $a_t(\theta) \in \{h_t(\theta), h_{i,t}^\theta(\theta), h_{i,j,t}^{\theta,\theta}(\theta)\}$ and $a_t^*(\theta) \in \{h_t^*(\theta), h_{i,t}^{*\theta}(\theta), h_{i,j,t}^{*\theta,\theta}(\theta)\}$.*

(a) *A stationary and ergodic solution $a_t^*(\theta)$ exists for each $\theta \in \Theta$, it is $\sigma(y_\tau : \tau \leq t - 1)$-measurable, and $\inf_{\theta \in \Theta} a_t^*(\theta) > 0$ a.s. Further, $h_t^*(\theta^0) = \sigma_t^2$ a.s., $h_t^{*\theta}(\theta) = (\partial/\partial\theta)h_t^*(\theta)$ and $h_t^{*\theta,\theta}(\theta) = (\partial/\partial\theta)h_t^{*\theta}(\theta)$ a.s.*

(b) *$E[\sup_{\theta \in \Theta} |a_t^*(\theta)|^\iota] < \infty$ for some tiny $\iota > 0$.*

(c) *If $a_t(\theta)$ is any other stationary solution then $E[(\sup_{\theta \in \Theta} |a_t^*(\theta) - a_t(\theta)|)^\iota] = o(\rho^t)$ for some $\rho \in (0, 1)$.*

(d) *$E[\sup_{\theta \in \Theta} |w_t^*(\theta) - w_t(\theta)|] = o(\rho^t)$ for each $w_t(\theta) \in \{\mathfrak{s}_{i,t}(\theta), \mathfrak{d}_{i,j,t}(\theta)\}$.*

(e) *$1/n \sum_{t=1}^n E[\sup_{\theta \in \Theta} |\tilde{I}_{n,t}^{(\mathcal{E})}(\theta) - I_{n,t}^{(\mathcal{E})}(\theta)|]$ and $1/n \sum_{t=1}^n E[\sup_{\theta \in \Theta} |\widehat{\tilde{I}}_{n,t}^{(\mathcal{E})}(\theta) - \hat{I}_{n,t}^{(\mathcal{E})}(\theta)|]$ are $o(1)$.*

Proof of Theorem 2.3. We first characterize properties of random variables based on $h_t(\theta)$. We then prove consistency of the feasible QMTTL estimator $\tilde{\theta}_n \xrightarrow{p} \theta^0$. Lastly, we prove the claim $\mathcal{V}_n^{1/2}(\tilde{\theta}_n - \hat{\theta}_n) \xrightarrow{p} 0$.

Define $\widehat{\tilde{m}}_n(\theta) := 1/n \sum_{t=1}^n \widehat{\tilde{m}}_{n,t}(\theta)$, $\tilde{m}_n(\theta) := 1/n \sum_{t=1}^n \tilde{m}_{n,t}(\theta)$, $\tilde{\mathcal{M}}_n(\theta) := 1/n \times \sum_{t=1}^n E[\tilde{m}_{n,t}(\theta)]$, and $\tilde{\mathfrak{e}}_n := 1/n \sum_{t=1}^n \sup_{\theta \in \Theta} E\|\tilde{m}_{n,t}(\theta)\|$, and recall $\mathfrak{e}_n := \sup_{\theta \in \Theta} E\|m_{n,t}(\theta)\|$.

Step 1: Use Lemma A.7 to obtain $|\tilde{\mathfrak{e}}_n - \mathfrak{e}_n| \leq \sup_{\theta \in \Theta} 1/n \sum_{t=1}^n \|\tilde{m}_{n,t}(\theta) - m_{n,t}(\theta)\|\| = O_p(1/n) = o_p(1)$. Similarly, $\|1/n \sum_{t=1}^n \|\tilde{G}_t(\theta)\tilde{I}_{n,t}^{(\mathcal{E})}(\theta) - G_t(\theta)I_{n,t}^{(\mathcal{E})}(\theta)\|$, $\|1/n \sum_{t=1}^n \|\tilde{G}_t(\theta) \times \widehat{\tilde{I}}_{n,t}^{(\mathcal{E})}(\theta) - G_t(\theta)\hat{I}_{n,t}^{(\mathcal{E})}(\theta)\|\|$, and $\|\tilde{\mathcal{G}}(\theta) - \mathcal{G}(\theta)\|$ are uniformly $o_p(1)$, and for any sequence of positive numbers $\{g_n\}$, $g_n \to \infty$, $\sup_{\theta \in \Theta}\{1/g_n \sum_{t=1}^n | |\tilde{\epsilon}_t(\theta)| - |\epsilon_t(\theta)| | = o_p(1)$. Use the latter to deduce $\sup_{\theta \in \Theta} k_n^{1/2} |\tilde{\mathcal{E}}_{(k_n)}^{(a)}(\theta) - \mathcal{E}_{(k_n)}^{(a)}(\theta)| \xrightarrow{p} 0$, hence by Lemma B.2 $\sup_{\theta \in \Theta} |\tilde{\mathcal{E}}_{(k_{1,n})}^{(-)}(\theta)/\mathcal{L}_n(\theta) + 1| = O_p(1/k_{1,n}^{1/2})$ and $\sup_{\theta \in \Theta} |\tilde{\mathcal{E}}_{(k_{2,n})}^{(+)}(\theta)/\mathcal{U}_n(\theta) - 1| = O_p(1/k_{1,n}^{1/2})$. By similar arguments and Lemma A.7 it is straightforward to verify Lemmas A.1, A.3 and A.4 extend to $\widehat{\tilde{m}}_n(\theta)$ and $\tilde{m}_{n,t}(\theta)$.

Step 2 ($\tilde{\theta}_n \xrightarrow{p} \theta^0$): We follow the proof of Theorem 2.1. By the Lemma A.6(c), (d) arguments and $\|\tilde{\mathcal{G}} - \mathcal{G}\| = o(1)$ it follows $1/n \sum_{t=1}^n E[\tilde{m}_{n,t}(\theta)] = \mathcal{G} \times (\theta - \theta^0) \times (1 + o(1))$ and $\tilde{\mathfrak{e}}_n \leq K\|\mathcal{G}\|$. Since $\|\mathcal{G}\| > 0$ it follows $\tilde{\epsilon}(\delta) := \liminf_{n \to \infty} \inf_{\|\theta - \theta^0\| > \delta}\{\tilde{\mathfrak{e}}_n^{-1}\|1/n \sum_{t=1}^n E[\tilde{m}_{n,t}(\theta)]\|\} > 0$ for every $n \geq N$ and $\delta > 0$. Therefore $P(\|\tilde{\theta}_n - \theta^0\| > \delta) \leq P(\|\tilde{\mathcal{M}}_n(\tilde{\theta}_n)\|/\tilde{\mathfrak{e}}_n > \tilde{\epsilon}(\delta))$. It remains to show $\|\tilde{\mathcal{M}}_n(\tilde{\theta}_n)\|/\tilde{\mathfrak{e}}_n = o_p(1)$.

Note $\|\tilde{\mathcal{M}}_n(\tilde{\theta}_n)\| \leq \|\widehat{\tilde{m}}_n(\tilde{\theta}_n)\| + \|\widehat{\tilde{m}}_n(\tilde{\theta}_n) - \tilde{\mathcal{M}}_n(\tilde{\theta}_n)\|$, where $\widehat{\tilde{m}}_n(\tilde{\theta}_n) = 0$ a.s. by $\tilde{\theta}_n$ a minimizer. It remains to show $\|\widehat{\tilde{m}}_n(\tilde{\theta}_n) - \tilde{\mathcal{M}}_n(\tilde{\theta}_n)\|/\tilde{\mathfrak{e}}_n = o_p(1)$. Note

$$\sup_{\theta \in \Theta} \|\widehat{\tilde{m}}_n(\theta) - \tilde{\mathcal{M}}_n(\theta)\|$$

$$\leq \sup_{\theta \in \Theta} \|\widehat{\tilde{m}}_n(\theta) - \hat{m}_n(\theta)\| + \sup_{\theta \in \Theta} \|\hat{m}_n(\theta) - \mathcal{M}_n(\theta)\| + \sup_{\theta \in \Theta} \|\tilde{\mathcal{M}}_n(\theta) - \mathcal{M}_n(\theta)\|.$$

The first and third terms on the right-hand side are $o_p(1)$ by Step 1. The second is $o_p(\mathfrak{e}_n)$ by the proof of Theorem 2.1. Since $|\tilde{\mathfrak{e}}_n - \mathfrak{e}_n| = o_p(1)$ we have shown $\sup_{\theta \in \Theta} \|\widehat{m}_n(\theta) - \tilde{\mathcal{M}}_n(\theta)\| = o_p(\tilde{\mathfrak{e}}_n)$ hence $\|\widehat{m}_n(\tilde{\theta}_n) - \tilde{\mathcal{M}}_n(\tilde{\theta}_n)\|/\tilde{\mathfrak{e}}_n = o_p(1)$ as required.

Step 3 $(\mathcal{V}_n^{1/2}(\tilde{\theta}_n - \hat{\theta}_n) \xrightarrow{p} 0)$: The first order conditions are $\sum_{t=1}^n \hat{m}_{n,t}(\hat{\theta}_n) = 0$ a.s. and $\sum_{t=1}^n \widehat{\tilde{m}}_{n,t}(\tilde{\theta}_n) = 0$ a.s. Combine $\tilde{\theta}_n \xrightarrow{p} \theta^0$, $\sup_{\theta \in \Theta} \|1/n \sum_{t=1}^n \|\tilde{G}_t(\theta) \widehat{\tilde{I}}_{n,t}^{(\mathcal{E})}(\theta) - G_t(\theta) \hat{I}_{n,t}^{(\mathcal{E})}(\theta)]\| = o_p(1)$, and $\|\tilde{\mathcal{G}} - \mathcal{G}\| = o(1)$ to deduce by Lemma A.6 $1/n \sum_{t=1}^n \tilde{G}_t(\tilde{\theta}_n) \widehat{\tilde{I}}_{n,t}^{(\mathcal{E})} = \mathcal{G} \times (1 + o_p(1))$. Therefore, in view of consistency of the infeasible estimator $\hat{\theta}_n \xrightarrow{p} \theta^0$, expansion Lemma A.4, and the construction $\mathcal{V}_n = n\mathcal{G}\Sigma_n^{-1}\mathcal{G}$, it follows

$$\frac{1}{n^{1/2}} \Sigma_n^{-1/2} \sum_{t=1}^n \{\widehat{\tilde{m}}_{n,t}(\tilde{\theta}_n) - \hat{m}_{n,t}(\tilde{\theta}_n)\} = \frac{1}{n^{1/2}} \Sigma_n^{-1/2} \sum_{t=1}^n \{\hat{m}_{n,t}(\hat{\theta}_n) - \hat{m}_{n,t}(\tilde{\theta}_n)\}$$
$$= \mathcal{V}_n^{1/2}(\hat{\theta}_n - \tilde{\theta}_n)(1 + o_p(1)). \tag{A.3}$$

Further, by two applications of Lemmas A.1(a), A.4 and A.6, and cancelling the terms $\mathcal{V}_n^{1/2}(\tilde{\theta}_n - \theta^0) = n^{1/2}\Sigma_n^{-1/2}\mathcal{G}(\tilde{\theta}_n - \theta^0)$, we have

$$\frac{1}{n^{1/2}} \Sigma_n^{-1/2} \sum_{t=1}^n \{\widehat{\tilde{m}}_{n,t}(\tilde{\theta}_n) - \hat{m}_{n,t}(\tilde{\theta}_n)\}$$

$$= \frac{1}{n^{1/2}} \Sigma_n^{-1/2} \sum_{t=1}^n \{\widehat{\tilde{m}}_{n,t}(\tilde{\theta}_n) - \tilde{m}_{n,t}\}$$

$$- \frac{1}{n^{1/2}} \Sigma_n^{-1/2} \sum_{t=1}^n \{\hat{m}_{n,t}(\tilde{\theta}_n) - m_{n,t}\} + \frac{1}{n^{1/2}} \Sigma_n^{-1/2} \sum_{t=1}^n \{\tilde{m}_{n,t} - m_{n,t}\}$$

$$= \frac{1}{n^{1/2}} \Sigma_n^{-1/2} \sum_{t=1}^n \{\tilde{m}_{n,t} - m_{n,t}\} \tag{A.4}$$

$$+ n^{1/2}\Sigma_n^{-1/2}\mathcal{G}(\tilde{\theta}_n - \theta^0)(1 + o_p(1)) - n^{1/2}\Sigma_n^{-1/2}\mathcal{G}(\tilde{\theta}_n - \theta^0)(1 + o_p(1))$$

$$= \frac{1}{n^{1/2}} \Sigma_n^{-1/2} \sum_{t=1}^n \{\tilde{m}_{n,t} - m_{n,t}\} + o_p(\|\mathcal{V}_n^{1/2}(\tilde{\theta}_n - \theta^0)\|).$$

Combine (A.3), (A.4) and Theorem 2.2 to obtain $\mathcal{V}_n^{1/2}(\tilde{\theta}_n - \hat{\theta}_n) = n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^n \{\tilde{m}_{n,t} - m_{n,t}\}(1 + o_p(1))$. By Loève's inequality, $\liminf_{n \to \infty} \|\Sigma_n\| > 0$ in view of non-degeneracy and trimming negligibility, and Lemma A.7(d), it follows for tiny $\iota > 0$, $\rho \in (0, 1)$, and sufficiently large n and K

$$E\left|\frac{1}{n^{1/2}} \Sigma_n^{-1/2} \sum_{t=1}^n \{\tilde{m}_{n,t} - m_{n,t}\}\right|^\iota \leq K \frac{1}{n^{\iota/2}} \sum_{t=1}^n E|\tilde{m}_{n,t} - m_{n,t}|^\iota \leq K \frac{1}{n^{\iota/2}} \sum_{t=1}^n \rho^t = o(1).$$

Therefore, $\mathcal{V}_n^{1/2}(\tilde{\theta}_n - \hat{\theta}_n) = \mathrm{o}_p(1)$ by Markov's inequality. $\qquad\square$

Proof of Theorem 3.1. By Assumption 4 $\psi(u,c) = u\varpi(u,c)I(|u| \le c)$ behaves like $uI(|u| \le c)$ as $c \to \infty$. See (21). In the following, we therefore only treat the simple trimming transform $\psi(u,c) = uI(|u| \le c)$. The general case with properties (21) and (23) has a similar proof.

Lemmas A.1–A.6 extend to cover the equations

$$\widehat{m}_{n,t}(\theta) = \left(\epsilon_t^2(\theta)\hat{I}_{n,t}^{(\epsilon)}(\theta) - \frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2(\theta)\hat{I}_{n,t}^{(\epsilon)}(\theta) \right) \times \mathfrak{s}_t(\theta),$$

$$\check{m}_{n,t}(\theta) = \left(\epsilon_t^2(\theta)I_{n,t}^{(\epsilon)}(\theta) - E\left[\epsilon_t^2(\theta)I_{n,t}^{(\epsilon)}(\theta)\right] \right) \times \left(\mathfrak{s}_t(\theta) - E\left[\mathfrak{s}_t(\theta)\right] \right).$$

Consider Lemma A.1(a). By Lemma A.1, it follows

$$\Sigma_n^{-1/2}\frac{1}{n^{1/2}}\sum_{t=1}^{n}\left(\epsilon_t^2\hat{I}_{n,t}^{(\epsilon)} - \frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2\hat{I}_{n,t}^{(\epsilon)} \right) \times \mathfrak{s}_t$$

$$= \Sigma_n^{-1/2}\frac{1}{n^{1/2}}\sum_{t=1}^{n}\left(\epsilon_t^2 I_{n,t}^{(\epsilon)} - \frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2 I_{n,t}^{(\epsilon)} \right) \times \mathfrak{s}_t + \mathrm{o}_p(1),$$

where by independence and dominated convergence $\Sigma_n \sim E[(\epsilon_t^2 I_{n,t}^{(\epsilon)} - E[\epsilon_t^2 I_{n,t}^{(\epsilon)}])^2] \times E[\mathfrak{s}_t\mathfrak{s}_t'] =: \sigma_n^2\mathfrak{S}$. Now add and subtract $E[\epsilon_t^2 I_{n,t}^{(\epsilon)}]$ and $E[\mathfrak{s}_t]$ to deduce

$$\Sigma_n^{-1/2}\frac{1}{n^{1/2}}\sum_{t=1}^{n}\left(\epsilon_t^2\hat{I}_{n,t}^{(\epsilon)} - \frac{1}{n}\sum_{t=1}^{n}\epsilon_t^2\hat{I}_{n,t}^{(\epsilon)} \right) \times \mathfrak{s}_t$$

$$= \Sigma_n^{-1/2}\frac{1}{n^{1/2}}\sum_{t=1}^{n}(\epsilon_t^2 I_{n,t}^{(\epsilon)} - E\left[\epsilon_t^2 I_{n,t}^{(\epsilon)}\right]) \times (\mathfrak{s}_t - E[\mathfrak{s}_t]) + \mathrm{o}_p(1) \qquad (A.5)$$

$$- \frac{1}{\sigma_n n^{1/2}}\sum_{t=1}^{n}(\epsilon_t^2 I_{n,t}^{(\epsilon)} - E\left[\epsilon_t^2 I_{n,t}^{(\epsilon)}\right]) \times \mathfrak{S} \times \left(\frac{1}{n}\sum_{t=1}^{n}\mathfrak{s}_t - E[\mathfrak{s}_t] \right) \times \left(1 + \mathrm{o}_p(1) \right).$$

Under Assumption 1 $\mathfrak{s}_t$ is stationary, ergodic and integrable, hence $1/n\sum_{t=1}^{n}\mathfrak{s}_t - E[\mathfrak{s}_t] = \mathrm{o}_p(1)$, and by a generalization of central limit theorem Lemma A.5 $\sigma_n^{-1}n^{-1/2}\sum_{t=1}^{n}(\epsilon_t^2 I_{n,t}^{(\epsilon)} - E[\epsilon_t^2 I_{n,t}^{(\epsilon)}]) = \mathrm{O}_p(1)$. The second term in (A.5) is therefore $\mathrm{o}_p(1)$, hence $\Sigma_n^{-1/2}n^{-1/2} \times \sum_{t=1}^{n}(\widehat{m}_{n,t} - \check{m}_{n,t}) = \mathrm{o}_p(1)$ which extends Lemma A.1(a) to $\{\widehat{m}_{n,t}, \check{m}_{n,t}\}$. In view of $L_{2+\iota}$-boundedness of $\sup_{\theta \in \mathcal{N}_0} \|\mathfrak{s}_t(\theta)\|$ for some compact $\mathcal{N}_0 \subset \Theta$ with positive Lebesgue measure and containing θ^0, and independence of ϵ_t, the arguments used to prove Lemmas A.1(b), A.2–A.6 carry over with simple modifications to cover $\{\widehat{m}_{n,t}, \check{m}_{n,t}\}$. The claims therefore follow by imitating the proofs of Theorems 2.1 and 2.2, and by the constructions of $\mathring{\mathcal{V}}_n$ and $\mathcal{V}_n$. $\qquad\square$

Lemma A.8. $1/n\sum_{t=1}^{n}\mathcal{E}_t^2(\hat{\theta}_n)\hat{I}_{n,t}^{(\mathcal{E})}(\hat{\theta}_n)/E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \xrightarrow{p} 1.$

Proof of Theorem 4.1. The claim follows from Jacobian consistency Lemma A.6(b) and Lemma A.8. □

A.3. Proofs of supporting lemmas

In order to decrease the number of cases we augment Assumption 1(b) and impose power law tails on ϵ_t in general:

$$P\big(|\epsilon_t| > a\big) = da^{-\kappa}\big(1 + o(1)\big) \qquad \text{where } d \in (0, \infty) \text{ and } \kappa \in (2, \infty). \tag{A.6}$$

Notice $\epsilon_t(\theta)$ is stationary and ergodic on Θ by (2), and also has a power law tail. The latter follows by noting

$$\epsilon_t(\theta) = \frac{y_t}{\sigma_t(\theta)} = \epsilon_t \frac{\sigma_t}{\sigma_t(\theta)} \quad \text{and} \quad \mathcal{E}_t(\theta) = \epsilon_t^2(\theta) - 1,$$

where $E(\sup_{\theta \in \Theta} |\sigma_t^2/\sigma_t^2(\theta)|)^p < \infty$ for any $p > 0$ under Assumption 1. Since ϵ_t is independent of $\sigma_t/\sigma_t(\theta)$ the product convolution $\epsilon_t \times (\sigma_t/\sigma_t(\theta))$ has tail (A.6) with the same index $\kappa > 2$ (Breiman [10]). In general $\lim_{a \to \infty} \sup_{\theta \in \Theta}\{|c^\kappa P(|\epsilon_t(\theta)| > a) - d(\theta)|\} = 0$ and $\inf_{\theta \in \Theta}\{d(\theta)\} > 0$ and $\sup_{\theta \in \Theta}\{d(\theta)\} < \infty$. Hence, in view of (14), $\mathcal{E}_t(\theta) := \epsilon_t^2(\theta) - 1$ also satisfies

$$\lim_{a \to \infty} \sup_{\theta \in \Theta}\big\{\big|a^{\kappa/2}P\big(|\mathcal{E}_t(\theta)| > a\big) - d(\theta)\big|\big\} = 0$$
$$\text{where } \inf_{\theta \in \Theta}\big\{d(\theta)\big\} > 0 \text{ and } \sup_{\theta \in \Theta}\big\{d(\theta)\big\} < \infty. \tag{A.7}$$

Recall $P(|\mathcal{E}_t(\theta)| > \mathcal{C}_n(\theta)) = k_n/n$ holds for $\mathcal{C}_n(\theta) = \mathcal{U}_n(\theta)$ and $k_n = k_{2,n}$. Then by (A.7)

$$\mathcal{C}_n(\theta) = d(\theta)^{2/\kappa}(n/k_n)^{2/\kappa}. \tag{A.8}$$

Further, by (A.7) and an application of Karamata's theorem:

$$\text{if } \kappa = 4: \quad E\big[\mathcal{E}_t^2(\theta)I\big(|\mathcal{E}_t(\theta)| \leq \mathcal{C}_n(\theta)\big)\big] \sim d(\theta)\ln(n),$$

$$\text{if } \kappa < 4: \quad E\big[\mathcal{E}_t^2(\theta)I\big(|\mathcal{E}_t(\theta)| \leq \mathcal{C}_n(\theta)\big)\big] \sim \frac{\kappa}{4 - \kappa}\mathcal{C}_n^2(\theta)P\big(|\mathcal{E}_t(\theta)| > \mathcal{C}_n(\theta)\big) \tag{A.9}$$

$$= \frac{\kappa}{4 - \kappa}d(\theta)^{4/\kappa}(n/k_n)^{4/\kappa - 1}.$$

Uniform bounds are similar given (A.7)–(A.9). For example, when $\kappa < 4$:

$$\sup_{\theta \in \Theta}\left\{\frac{n}{k_n}\frac{\mathcal{C}_n^2(\theta)}{E[\mathcal{E}_t^2(\theta)I(|\mathcal{E}_t(\theta)| \leq \mathcal{C}_n(\theta))]}\right\} \to (0, \infty). \tag{A.10}$$

Unless otherwise noted, and in view of (14) and (A.7), we assume two-tailed trimming to reduce notation, hence thresholds and fractiles are simply $\mathcal{C}_n(\theta)$ and k_n, and order statistics are $\mathcal{E}_{(k_n)}^{(a)}(\theta)$ where $\mathcal{E}_t^{(a)}(\theta) := |\mathcal{E}_t(\theta)|$.

The proofs of Lemmas A.1–A.8 require two supporting results. See the supplementary material Hill [32] for proofs. First, trimming indicators satisfy a uniform CLT.

Lemma B.1 (Uniform indicator CLT). *Define $\mathcal{I}_{n,t}(\theta) := ((n/k_n)^{1/2})\{I(|\mathcal{E}_t(\theta)| \leq \mathcal{C}_n(\theta)) - E[I(|\mathcal{E}_t(\theta)| \leq \mathcal{C}_n(\theta))]\}$. Then $\{n^{-1/2}\sum_{t=1}^n \mathcal{I}_{n,t}(\theta) : \theta \in \Theta\} \Longrightarrow^* \{\mathcal{I}(\theta) : \theta \in \Theta\}$, where $\mathcal{I}(\theta)$ is a Gaussian process with uniformly bounded and uniformly continuous sample paths with respect to L_2-norm, and $\Longrightarrow^*$ denotes weak convergence on a Polish space (Hoffman-Jørgensen [35]).*

Second, intermediate order statistics are uniformly bounded in probability.

Lemma B.2 (Uniform order statistic bound). $\sup_{\theta\in\Theta} |\mathcal{E}_{(k_n)}^{(a)}(\theta)/\mathcal{C}_n(\theta) - 1| = O_p(1/k_n^{1/2})$.

Lemmas A.1, A.3, A.4 and A.6 are similar to results proven in Hill [31], Appendix A, hence their proofs are relegated to the supplementary material Hill [32].

Proof of Lemma A.2.

Claim (a): $\mathfrak{s}_t$ is $L_{2+\iota}$-bounded by Assumption 1, hence by error independence $\Sigma_{i,i,n} \sim E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \times E[\mathfrak{s}_{i,t}^2] \sim K E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}]$. The claim now follows from arguments leading to Theorem 2.5.

Claim (b): We prove the claim for $\mathcal{S}_{i,i,n}$, so let m_t denote $m_{i,t}$, hence $\mathfrak{s}_t$ denotes $\mathfrak{s}_{i,t}$, and express $\mathcal{S}_{i,i,n}$ as $\mathcal{S}_n$. Note $\mathcal{S}_n \sim E[m_{n,t}^2] + 2\sum_{i=1}^{n-1}(1 - i/n)E[m_{n,1}m_{n,i+1}]$. If $E[m_t^2] < \infty$ then $\mathcal{S}_n \sim K = o(n/\ln(n))$ in view of geometric β-mixing (cf. Ibragimov [36]).

Now assume $E[m_t^2] = \infty$. We first characterize the tails of $m_t = (\epsilon_t^2 - 1)\mathfrak{s}_t$, and then bound $\sum_{i=1}^{n-1}|E[m_{n,1}m_{n,i+1}]|$.

Step 1: By Assumption 1(b) and (A.6) independent ϵ_t has a power law tail with index $\kappa \in (2, 4]$, and since $\alpha^0 + \beta^0 > 0$ it follows $E[\mathfrak{s}_t^2] < \infty$. Therefore m_t has a power law tail with index $\kappa_m := \kappa/2 \in (1, 2]$, cf. Breiman [10].

Step 2: Define quantile functions $Q_n(u) = \inf\{m \geq 0 : P(|m_{n,t}| > m) \leq u\}$ and $Q(u) = \inf\{m \geq 0 : P(|m_t| > m) \leq u\}$ for $u \in [0, 1]$, recall geometric β-mixing implies α-mixing with coefficients $\alpha_h \leq K\rho^h$ for $\rho \in (0, 1)$. By Theorem 1.1 of Rio [56]

$$\sum_{i=1}^{n-1}|E[m_{n,1}m_{n,i+1}]| \leq 2\sum_{i=1}^{n-1}\int_0^{2\alpha_i} Q_n^2(u)\, du \leq 2\sum_{i=1}^{n-1}\int_0^{K\rho^i} Q_n^2(u)\, du.$$

Tail-trimming $m_{n,t} = m_t I_{n,t}^{(\mathcal{E})}$ coupled with distribution continuity imply $P(m_{n,t} = 0) = k_n/n$. Thus $Q_n(u) = 0$ for $u \in [0, k_n/n]$ and $Q_n(u) = Q(u)$ for $u \in (k_n/n, 1]$. Further, under the Step 1 power law properties $Q(u) = O(u^{-2/\kappa})$. Therefore

$$\sum_{i=1}^{n-1}|E[m_{n,1}m_{n,i+1}]| \leq K\sum_{i=1}^{n-1}\int_{k_n/n}^{K\rho^i} u^{-4/\kappa}\, du \leq K\sum_{i=1}^{n-1}\max\{0, (n/k_n)^{(4/\kappa-1)} - K\rho^{-i(4/\kappa-1)}\}$$

$$= K\sum_{i=1}^{K\ln(n/k_n)}\{(n/k_n)^{(4/\kappa-1)} - K\rho^{-i(4/\kappa-1)}\}.$$

Moreover $\sum_{i=1}^{K\ln(n/k_n)}\{(n/k_n)^{4/\kappa-1} - K\rho^{-i(4/\kappa-1)}\} = K\ln(n/k_n) \times (n/k_n)^{4/\kappa-1}(1+\mathrm{O}(1))$ and $k_n = \mathrm{o}(n)$ hence

$$\sum_{i=1}^{n-1}\left|E[m_{n,1}m_{n,i+1}]\right| \le K\ln(n/k_n) \times (n/k_n)^{4/\kappa-1}\big(1+\mathrm{O}(1)\big)$$

$$\le K\ln(n/k_n) \times (n/k_n)^{4/\kappa-1} \le K\ln(n)(n/k_n)^{4/\kappa-1}.$$

Further, $\ln(n)(n/k_n)^{4/\kappa-1} = \mathrm{o}(n/\ln(n))$ since $k_n \to \infty$ and $\kappa \in (2,4]$. Finally, by Step 1 and (A.9) $E[m_{n,t}^2] \sim K(n/k_n)^{4/\kappa-1}$ if $\kappa < 4$ and $E[m_{n,t}^2] \sim K\ln(n)$ if $\kappa = 4$. Therefore $\mathcal{S}_n \le K\ln(n)(n/k_n)^{4/\kappa-1} = \mathrm{o}(n/\ln(n))$ which completes the proof. $\qquad\square$

Proof of Lemma A.5. By identification Assumption 2 $n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^n m_{n,t} = n^{-1/2}\Sigma_n^{-1/2} \times \sum_{t=1}^n\{m_{n,t} - E[m_{n,t}]\} + \mathrm{o}(1)$. Define $z_{n,t} := r'\Sigma_n^{-1/2}\{m_{n,t} - E[m_{n,t}]\}$ for any $r \in \mathbb{R}^q$, $r'r = 1$. Note by error independence, dominated convergence and (6): $E(\sum_{t=1}^n z_{n,t})^2 \sim n$. We will prove $1/n^{1/2}\sum_{t=1}^n z_{n,t} \xrightarrow{d} N(0,1)$, hence the claim will follow from the Cramér–Wold Theorem. Define $\mathfrak{I}_t := \sigma(y_\tau : \tau \le t)$.

In view of geometric β-mixing and stationarity under Assumption 1, and $E[z_{n,t}^2] = 1$ it suffices to show the three conditions of Theorem 2.1 in Peligrad [51] hold.[7] The first two are $\sup_{n\ge1} 1/n\sum_{t=1}^n E[z_{n,t}^2] < \infty$ and the Lindeberg condition $1/n\sum_{t=1}^n E[z_{n,t}^2 I(|z_{n,t}| > \varepsilon n^{1/2})] \to 0$ $\forall \varepsilon > 0$. By construction $E[z_{n,t}^2] = 1$ hence $1/n\sum_{t=1}^n E[z_{n,t}^2] = 1 + \mathrm{o}(1)$ which verifies the first.

The Lindeberg condition holds if $\kappa > 4$ since $E|\epsilon_t|^{4+\iota} < \infty$ and $E|\mathfrak{s}_{i,t}|^{2+\iota} < \infty$ for some $\iota > 0$, hence $\limsup_{n\to\infty} E|z_{n,t}|^{2+\iota} < \infty$. Now suppose $\kappa \le 4$, assume $E[\mathcal{E}_t I_{n,t}^{(\mathcal{E})}] = 0$ to simplify notation, and note $z_{n,t} = \mathcal{E}_t I_{n,t}^{(\mathcal{E})} r'\Sigma_n^{-1/2}\mathfrak{s}_t$. By independence and L_2-boundedness of $\mathfrak{s}_t$ it follows $\Sigma_n = E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \times \mathfrak{S}$ where $\mathfrak{S} = E[\mathfrak{s}_t\mathfrak{s}_t']$ is finite and positive definite. By construction $\liminf_{n\to\infty}\inf_{r'r=1}\|\Sigma_n\|(r'\Sigma_n^{-1/2}\mathfrak{s}_t)^2 > 0$ a.s., by independence $\mathcal{E}_t^2 \times r'\Sigma_n^{-1/2}\mathfrak{s}_t\mathfrak{s}_t'\Sigma_n^{-1/2}r \times \|\Sigma_n\|$ has Paretian tails with index $\kappa/4 \le 1$, and by trimming $|\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}| \le K\mathcal{C}_n^2$. Therefore, for finite $K > 0$ that may be different in different places,

$$E[z_{n,t}^2 I(z_{n,t}^2 > \varepsilon^2 n)] \le KE\left((r'\Sigma_n^{-1/2}\mathfrak{s}_t)^2 \times E\left[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})} I\left(\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})} > \frac{\varepsilon^2 n}{(r'\Sigma_n^{-1/2}\mathfrak{s}_t)^2}\right)\Big|\mathfrak{I}_{t-1}\right]\right)$$

$$\le KE\left((r'\Sigma_n^{-1/2}\mathfrak{s}_t)^2 \times E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})} I(\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})} > K\varepsilon^2 n E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}])|\mathfrak{I}_{t-1}]\right)$$

$$\le KE\left((r'\Sigma_n^{-1/2}\mathfrak{s}_t)^2 \times E\left[\int_{K\varepsilon^2 n E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}]}^{K\mathcal{C}_n^2} u^{-\kappa/4}\,\mathrm{d}u\right]\right).$$

---

[7]We require a result like Theorem 2.1 in Peligrad [51] since asymptotically $z_{n,t}$ need not have finite moments higher than two.

In general $\mathcal{C}_n^2 = K(n/k_n)^{4/\kappa}$. If $\kappa = 4$ then $E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \sim K\ln(n))$ hence $\mathcal{C}_n^2 = K(n/k_n) < K\varepsilon^2 n E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}]$ as $n \to \infty$. This implies for some $N \in \mathbb{N}$ and all $n \geq N$ that $\int_{K\varepsilon^2 n E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}]}^{K\mathcal{C}_n^2} u^{-\kappa/4} \, du = 0$. If $\kappa < 4$ then $E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \sim K\mathcal{C}_n^2(k_n/n) = K(n/k_n)^{4/\kappa - 1}$, hence again $\mathcal{C}_n^2 = K(n/k_n) < Kn(n/k_n)^{4/\kappa - 1} = K\varepsilon^2 n E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}]$ as $n \to \infty$. Therefore, $E[z_{n,t}^2 I(z_{n,t}^2 > \varepsilon^2 n)] = 0$ for some $N \in \mathbb{N}$ and all $n \geq N$. This proves $1/n \sum_{t=1}^n E[z_{n,t}^2 I(|z_{n,t}| > \varepsilon n^{1/2})] \to 0 \; \forall \varepsilon > 0$.

The third condition concerns the maximum correlation coefficient $\rho(\mathcal{A}, \mathcal{B}) := \sup_{f \in L_2(\mathcal{A}), g \in L_2(\mathcal{B})} |\mathrm{corr}(f, g)|$ defined on $L_2(\mathfrak{F})$ the space of L_2-bounded $\mathfrak{F}$-measurable random variables. We require the interlaced coefficient $\rho_k^* := \sup_{n \geq 1} \sup_{S_k, T_k} \rho(\sigma(z_{n,i} : i \in T_k), \sigma(z_{n,j} : j \in S_k))$ to satisfy $\lim_{k \to \infty} \rho_k^* < 1$, where $T_k, S_k \subset \{1, \ldots, n\}$ are non-empty subsets with $\inf_{s \in S_k, t \in T_k} \{|s - t|\} \geq k$, and $\sup_{S_k, T_k}$ is taken over all sets $\{S_k, T_k\}$ for a given distance k. See equations (1.2), (1.7) and (1.8) in Peligrad [51]. In view of the GARCH process and Assumption 1, $\{z_{n,t} : 1 \leq t \leq n\}_{n \geq 1}$ is a first order Markov chain that is stationary over $1 \leq t \leq n$, and by geometric β-mixing it is also geometric α-mixing. Since $\rho_1^* < 1$ as a consequence of independence of ϵ_t, it therefore follows $\rho_k^* \to 0$ by an extension of Theorem 3.3 in Bradley [9] to triangular arrays. □

Proof of Lemma A.7. Claims (a)–(c) follow from the Assumption 3 response Lipschitz properties. See Francq and Zakoïan [24,25] and Meitz and Saikkonen [44]. Claim (d) follows from stationarity, independence of ϵ_t, and (b) and (c).

Consider (e). We will prove $1/n \sum_{t=1}^n E[\sup_{\theta \in \Theta} |\tilde{I}_{n,t}^{(\mathcal{E})}(\theta) - I_{n,t}^{(\mathcal{E})}(\theta)|] = o(1)$, the second claim being similar. We can approximate $I(u) := I(u \leq 0)$ with the *regular* sequence $\{\mathfrak{I}_n(u)\}_{n \geq 1}$, defined by $\mathfrak{I}_n(u) := \int_{-\infty}^\infty I(\varpi) S(\mathcal{N}_n(\varpi - u)) \mathcal{N}_n e^{-\varpi^2/\mathcal{N}_n^2} \, d\varpi$ where $S(\xi) = e^{-1/(1-\xi^2)} / \int_{-1}^1 e^{-1/(1-w^2)} \, dw$ if $|\xi| < 1$ and $S(\xi) = 0$ if $|\xi| \geq 1$. Here $\{\mathcal{N}_n\}$ is a sequence of finite positive numbers, $\mathcal{N}_n \to \infty$, the rate to be chosen below. See Lighthill [40]. $\mathfrak{I}_n(u)$ is uniformly bounded in u, continuous and differentiable. Also, $(\partial/\partial u)I(u)$ has a regular sequence $\mathfrak{D}_n(u) := (\mathcal{N}_n/\pi)^{1/2} \exp\{-\mathcal{N}_n u^2\}$.

Define $\mathfrak{e}_t(a) := |\mathcal{E}_t| - a$ and $\tilde{\mathfrak{e}}_t(a) := |\tilde{\mathcal{E}}_t| - a$, and let $\tilde{\mathcal{C}}_n(\theta)$ satisfy $P(|\tilde{\mathcal{E}}_t(\theta)| \geq \tilde{\mathcal{C}}_n(\theta)) = k_n/n$. Hence $\tilde{I}_{n,t}^{(\mathcal{E})}(\theta) = I(\tilde{\mathfrak{e}}_t(\tilde{\mathcal{C}}_n(\theta)))$ and $I_{n,t}^{(\mathcal{E})} = I(\mathfrak{e}_t(\mathcal{C}_n(\theta)))$. Note $\mathcal{N}_n \to \infty$ can be made as fast as we choose such that $\sup_{\theta \in \Theta} |I(\mathfrak{e}_t(\mathcal{C}_n(\theta))) - I(\tilde{\mathfrak{e}}_t(\tilde{\mathcal{C}}_n(\theta)))| \leq K \sup_{\theta \in \Theta} |\mathfrak{I}_n(\mathfrak{e}_t(\mathcal{C}_n(\theta))) - \mathfrak{I}_n(\tilde{\mathfrak{e}}_t(\tilde{\mathcal{C}}_n(\theta)))| + o_p(1)$, and $\mathfrak{D}_n(u) \to 0$ as fast as we choose. Hence, by the mean-value-theorem and boundedness of $\mathfrak{D}_n(u)$ it follows $\sup_{\theta \in \Theta} |I(\mathfrak{e}_t(\mathcal{C}_n(\theta))) - I(\tilde{\mathfrak{e}}_t(\tilde{\mathcal{C}}_n(\theta)))| \leq K \sup_{\theta \in \Theta} |\tilde{\mathcal{E}}_t(\theta) - \mathcal{E}_t(\theta)| + K \sup_{\theta \in \Theta} |\tilde{\mathcal{C}}_n(\theta) - \mathcal{C}_n(\theta)|$. By (iii) $\sup_{\theta \in \Theta} |\tilde{\mathcal{E}}_t(\theta) - \mathcal{E}_t(\theta)| = o_p(\rho^t)$. Similarly $\sup_{\theta \in \Theta} \sum_{t=1}^n ||\tilde{\mathcal{E}}_t(\theta)| - |\mathcal{E}_t(\theta)|| = O_p(1)$ hence $\sup_{\theta \in \Theta} |\tilde{\mathcal{E}}_{(k_n)}^{(a)}(\theta) - \mathcal{E}_{(k_n)}^{(a)}(\theta)| \overset{p}{\to} 0$, hence by Lemma B.2 $\sup_{\theta \in \Theta} |\tilde{\mathcal{C}}_n(\theta) - \mathcal{C}_n(\theta)| \to 0$. Therefore by dominated convergence $1/n \times \sum_{t=1}^n E[\sup_{\theta \in \Theta} |\tilde{I}_{n,t}^{(\mathcal{E})}(\theta) - I_{n,t}^{(\mathcal{E})}(\theta)|] \leq Kn^{-1} \sum_{t=1}^n \rho^t + o(1) = o(1)$. □

Proof of Lemma A.8. Define $\mathfrak{Z}_{n,t}(\theta) := \mathcal{E}_t^2(\theta) I_{n,t}^{(\mathcal{E})}(\theta)/E[\mathcal{E}_t^2(\theta) I_{n,t}^{(\mathcal{E})}(\theta)]$. By the same arguments used to prove approximation Lemma A.1: $1/n \sum_{t=1}^n \mathcal{E}_t^2(\hat{\theta}_n) \hat{I}_{n,t}^{(\mathcal{E})}(\hat{\theta}_n) = 1/n \sum_{t=1}^n \mathcal{E}_t^2(\hat{\theta}_n) \times I_{n,t}^{(\mathcal{E})}(\hat{\theta}_n)(1 + o_p(1))$. Since $\mathfrak{Z}_{n,t}(\theta)$ is uniformly integrable and geometrically β-mixing by

Assumption 1(d), it follows $1/n \sum_{t=1}^{n} \mathfrak{Z}_{n,t}(\theta) \xrightarrow{p} 1$ by Theorem 2 and Example 4 in Andrews [2]. Moreover, since $\mathfrak{Z}_{n,t}(\theta)$ is trivially L_1-bounded uniformly in Θ, $\mathfrak{Z}_{n,t}(\theta)$ belongs to a separable Banach space, hence L_1-bracketing numbers satisfy $N_{[\,]}(\varepsilon, \Theta, \|\cdot\|_1) < \infty$ (Dudley [20], Proposition 7.1.7). Combine the pointwise law and $N_{[\,]}(\varepsilon, \Theta, \|\cdot\|_1) < \infty$ to deduce $\sup_{\theta \in \Theta} |1/n \sum_{t=1}^{n} \mathfrak{Z}_{n,t}(\theta)| \xrightarrow{p} 0$ by Theorem 7.1.5 of Dudley [20]. Therefore $1/n \sum_{t=1}^{n} \mathcal{E}_t^2(\hat{\theta}_n) I_{n,t}^{(\mathcal{E})}(\hat{\theta}_n) / E[\mathcal{E}_t^2(\hat{\theta}_n) I_{n,t}^{(\mathcal{E})}(\hat{\theta}_n)] \xrightarrow{p} 1$. Further, by the definition of a derivative: $|E[\mathcal{E}_t^2(\theta) I_{n,t}^{(\mathcal{E})}(\theta)] - E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}]| \leq \|(\partial/\partial\theta) E[\mathcal{E}_t^2(\theta) I_{n,t}^{(\mathcal{E})}(\theta)]|_{\theta^0}\| \times \|\theta - \theta^0\| \times (1 + \mathrm{o}(1))$. By the same argument as Lemma A.6(c) we can write

$$\frac{\partial}{\partial\theta} E[\mathcal{E}_t^2(\theta) I_{n,t}^{(\mathcal{E})}(\theta)]|_{\theta^0} = E\left[\frac{\partial}{\partial\theta} \mathcal{E}_t^2(\theta)\Big|_{\theta^0} \times I_{n,t}^{(\mathcal{E})}\right] \times (1 + \mathrm{o}(1))$$

$$= -2E[\mathcal{E}_t \epsilon_t^2 I_{n,t}^{(\mathcal{E})} \mathfrak{s}_t] \times (1 + \mathrm{o}(1)),$$

and trivially $E[\mathcal{E}_t \epsilon_t^2 I_{n,t}^{(\mathcal{E})} \mathfrak{s}_t] = E[\mathcal{E}_t^2 \mathfrak{s}_t I_{n,t}^{(\mathcal{E})}] - E[\mathcal{E}_t \mathfrak{s}_t I_{n,t}^{(\mathcal{E})}] = E[\mathcal{E}_t^2 \mathfrak{s}_t I_{n,t}^{(\mathcal{E})}] = E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \times E[\mathfrak{s}_t]$. Therefore $|E[\mathcal{E}_t^2(\theta) I_{n,t}^{(\mathcal{E})}(\theta)] - E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}]| \leq K |E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}]| \times \|\theta - \theta^0\| \times (1 + \mathrm{o}(1))$. Now use $\hat{\theta}_n \xrightarrow{p} \theta^0$ by Theorem 2.1 and $\inf_{n \geq N} E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] > 0$ for some $N \geq 1$ to deduce $E[\mathcal{E}_t^2(\hat{\theta}_n) \times I_{n,t}^{(\mathcal{E})}(\hat{\theta}_n)]/E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \to 1$. This proves $1/n \sum_{t=1}^{n} \mathcal{E}_t^2(\hat{\theta}_n) I_{n,t}^{(\mathcal{E})}(\hat{\theta}_n)/E[\mathcal{E}_t^2 I_{n,t}^{(\mathcal{E})}] \xrightarrow{p} 1$. $\qquad\square$

Acknowledgements

Supplementary Material

Supplement to "Robust estimation and inference for heavy tailed GARCH" (DOI: 10.3150/14-BEJ616SUPP; .pdf). We prove Lemmas A.1, A.3, A.4 and A.6, and Lemmas B.1 and B.2. Assume all functions satisfy Pollard's [54] permissibility criteria, the measure space that governs all random variables in this paper is complete, and therefore all majorants are measurable. Cf. Dudley [19]. Probability statements are therefore with respect to outer probability, and expectations over majorants are outer expectations.

References

[1] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Tukey, J.W. (1972). *Robust Estimates of Location*: *Survey and Advances*. Princeton, NJ: Princeton Univ. Press. MR0331595

[2] Andrews, D.W.K. (1988). Laws of large numbers for dependent nonidentically distributed random variables. *Econometric Theory* **4** 458–467. MR0985156

[3] Andrews, D.W.K. (1999). Estimation when a parameter is on a boundary. *Econometrica* **67** 1341–1383. MR1720781

[4] Basrak, B., Davis, R.A. and Mikosch, T. (2002). Regular variation of GARCH processes. *Stochastic Process. Appl.* **99** 95–115. MR1894253

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

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

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

[8] Boudt, K., Danielsson, J. and Laurent, S. (2011). Robust forecasting of dynamic conditional correlation garch models. *Internat. J. Forecasting* **29** 244–257.

[9] Bradley, R.C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. Update of, and a supplement to, the 1986 original. MR2178042

[10] Breiman, L. (1965). On some limit theorems similar to the arc-sin law. *Teor. Veroyatn. Primen.* **10** 351–360. MR0184274

[11] Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96** 1022–1030. MR1947250

[12] Carrasco, M. and Chen, X. (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* **18** 17–39. MR1885348

[13] Cavaliere, G. and Georgiev, I. (2009). Robust inference in autoregressions with multiple outliers. *Econometric Theory* **25** 1625–1661. MR2557576

[14] Charles, A. and Darné, O. (2005). Outliers and GARCH models in financial data. *Econom. Lett.* **86** 347–352. MR2124418

[15] Chen, C. and Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *J. Amer. Statist. Assoc.* **88** 284–297.

[16] Čížek, P. (2008). General trimmed estimation: Robust approach to nonlinear and limited dependent variable models. *Econometric Theory* **24** 1500–1529. MR2456536

[17] Davis, R.A. (2010). Heavy tails in financial time series. In *Encyclopedia of Quantitative Finance* (R. Cont, ed.). New York: Wiley.

[18] De Melo Mendes, B.V. (2000). Assessing the bias of maximum likelihood estimates of contaminated Garch models. *J. Stat. Comput. Simul.* **67** 359–376. MR1806901

[19] Dudley, R.M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899–929 (1979). MR0512411

[20] Dudley, R.M. (1999). *Uniform Central Limit Theorems. Cambridge Studies in Advanced Mathematics* **63**. Cambridge: Cambridge Univ. Press. MR1720712

[21] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events. For Insurance and Finance. Applications of Mathematics* (*New York*) **33**. Berlin: Springer. MR1458613

[22] Engle, R.F. and Ng, V.K. (1993). Measuring and testing the impact of news on volatility. *J. Finance* **48** 1749–1778.

[23] Fan, J., Qi, L. and Xiu, D. (2014). Quasi maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. *J. Business and Economic Statistics*. To appear.

[24] Francq, C. and Zakoïan, J.-M. (2004). Maximum likelihood estimation of pure GARCH and ARMA–GARCH processes. *Bernoulli* **10** 605–637. MR2076065

[25] Francq, C. and Zakoïan, J.-M. (2010). *GARCH Models*: *Structure, Statistical Inference and Financial Applications*. New York: Wiley.

[26] Hall, P. and Yao, Q. (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica* **71** 285–317. MR1956860

[27] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics. The Approach Based on Influence Functions. Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. New York: Wiley. MR0829458

[28] Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123

[29] Hill, J.B. (2010). On tail index estimation for dependent, heterogeneous data. *Econometric Theory* **26** 1398–1436. MR2684790

[30] Hill, J.B. (2012). Heavy-tail and plug-in robust consistent conditional moment tests of functional form. In *Festschrift in Honor of Hal White* (X. Chen and N. Swanson, eds.) 241–274. New York: Springer.

[31] Hill, J.B. (2013). Least tail-trimmed squares for infinite variance autoregressions. *J. Time Series Anal*. **34** 168–186. MR3028364

[32] Hill, J.B. (2014). Supplement to "Robust estimation and inference for heavy tailed GARCH." DOI:10.3150/14-BEJ616SUPP.

[33] Hill, J.B. and Aguilar, M. (2013). Moment condition tests for heavy tailed time series. *J. Econometrics* **172** 255–274. MR3010616

[34] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann*. *Math*. *Statist*. **35** 73–101. MR0161415

[35] Hoffman-Jørgensen, J. (1984). Convergence of stochastic processes on Polish spaces. Mimeo, Aarhus Univ., Denmark.

[36] Ibragimov, I.A. (1962). Some limit theorems for stationary processes. *Teor. Veroyatn*. *Primen*. **7** 361–392. MR0148125

[37] Jurečková, J. and Sen, P.K. (1996). *Robust Statistical Procedures*. *Asymptotics and Interrelations*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. New York: Wiley. MR1387346

[38] Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. *Springer Series in Statistics*. New York: Springer. MR0691492

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

[40] Lighthill, M.J. (1958). *Introduction to Fourier Analysis and Generalised Functions*. *Cambridge Monographs on Mechanics and Applied Mathematics*. Cambridge: Cambridge Univ. Press. MR0092119

[41] Ling, S. (2007). Self-weighted and local quasi-maximum likelihood estimators for ARMA–GARCH/IGARCH models. *J. Econometrics* **140** 849–873. MR2408929

[42] Liu, J.-C. (2006). On the tail behaviors of a family of GARCH processes. *Econometric Theory* **22** 852–862. MR2291220

[43] Mancini, L., Ronchetti, E. and Trojani, F. (2005). Optimal conditionally unbiased bounded-influence inference in dynamic location and scale models. *J. Amer. Statist. Assoc*. **100** 628–641. MR2160565

[44] Meitz, M. and Saikkonen, P. (2008). Stability of nonlinear AR-GARCH models. *J. Time Ser. Anal*. **29** 453–475. MR2410184

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

[46] Muler, N., Peña, D. and Yohai, V.J. (2009). Robust estimation for ARMA models. *Ann*. *Statist*. **37** 816–840. MR2502652

[47] Muler, N. and Yohai, V.J. (2008). Robust estimates for GARCH models. *J. Statist. Plann. Inference* **138** 2918–2940. MR2442223

[48] Nelson, D.B. (1990). Stationarity and persistence in the GARCH(1, 1) model. *Econometric Theory* **6** 318–334. MR1085577

[49] Newey, W.K. and Steigerwald, D.G. (1997). Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models. *Econometrica* **65** 587–599. MR1445623

[50] Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57** 1027–1057. MR1014540

[51] Peligrad, M. (1996). On the asymptotic normality of sequences of weak dependent random variables. *J. Theoret. Probab.* **9** 703–715. MR1400595

[52] Peng, L. and Yao, Q. (2003). Least absolute deviations estimation for ARCH and GARCH models. *Biometrika* **90** 967–975. MR2024770

[53] Petersen, K. (1983). *Ergodic Theory*. *Cambridge Studies in Advanced Mathematics* **2**. Cambridge: Cambridge Univ. Press. MR0833286

[54] Pollard, D. (1984). *Convergence of Stochastic Processes*. *Springer Series in Statistics*. New York: Springer. MR0762984

[55] Resnick, S.I. (1987). *Extreme Values*, *Regular Variation*, *and Point Processes*. *Applied Probability. A Series of the Applied Probability Trust* **4**. New York: Springer. MR0900810

[56] Rio, E. (1993). Covariance inequalities for strongly mixing processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **29** 587–597. MR1251142

[57] Ronchetti, E. and Trojani, F. (2001). Robust inference with GMM estimators. *J. Econometrics* **101** 37–69. MR1805872

[58] Sakata, S. and White, H. (1998). High breakdown point conditional dispersion estimation with application to s&p 500 daily returns volatility. *Econometrica* **66** 529–567.

[59] Shevlyakov, G., Morgenthaler, S. and Shurygin, A. (2008). Redescending *M*-estimators. *J. Statist. Plann. Inference* **138** 2906–2917. MR2526216

[60] Straumann, D. and Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *Ann. Statist.* **34** 2449–2495. MR2291507

[61] Zhu, K. and Ling, S. (2011). Global self-weighted and local quasi-maximum exponential likelihood estimators for ARMA–GARCH/IGARCH models. *Ann. Statist.* **39** 2131–2163. MR2893864