Original Article

Journal of Time Series Analysis

First vesion received July 2012

Published online in Wiley Online Library: 24 October 2012

(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12005

Least tail-trimmed squares for infinite variance autoregressions

Jonathan B. Hill[a,*,†]

We develop a robust least squares estimator for autoregressions with possibly heavy tailed errors. Robustness to heavy tails is ensured by negligibly trimming the squared error according to extreme values of the error and regressors. Tail-trimming ensures asymptotic normality *and* super-$\sqrt{n}$-convergence with a rate comparable to the highest achieved amongst M-estimators for stationary data. Moreover, tail-trimming ensures robustness to heavy tails in both small and large samples. By comparison, existing robust estimators are not as robust in small samples, have a slower rate of convergence when the variance is infinite, or are not asymptotically normal. We present a consistent estimator of the covariance matrix and treat classic inference without knowledge of the rate of convergence. A simulation study demonstrates the sharpness and approximate normality of the estimator, and we apply the estimator to financial returns data. Finally, tail-trimming can be easily extended beyond least squares estimation for a linear stationary AR model. We discuss extensions to quasi-maximum likelihood for GARCH, weighted least squares for a possibly non-stationary random coefficient autoregression, and empirical likelihood for robust confidence region estimation, in each case for models with possibly heavy tailed errors.

Keywords: Least squares; tail trimming; heavy tails; robust inference.
AMS subject classifications: Primary 62F35; secondary 62F07.

1. INTRODUCTION

We develop the least tail-trimmed squares (LTTS) estimator for a stationary autoregression with an error term that may be very heavy tailed. The model is

$$y_t = \xi + \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t(\theta) = \theta' x_t + \epsilon_t(\theta), \quad p \geq 1, \tag{1}$$

with parameter set $\theta = [\xi, \phi']' \in \mathbb{R}^{p+1}$ and regressors $x_t = [1, y_{t-1}, \ldots, y_{t-p}]'$, and the sample is $\{y_t\}_{t=1}^n$ with sample size $n \leq 1$. We assume there exists a unique point θ^0 such that the error $\epsilon_t = \epsilon_t(\theta^0)$ is independent and identically distributed (i.i.d.) symmetrically distributed about zero, and $E|\epsilon_t|^\iota < \infty$ for some $\iota > 0$. Further, the distribution tails of ϵ_t exhibit power law decay:

$$P(|\epsilon_t| > a) = da^{-\kappa}(1 + o(1)) \quad \text{where } d > 0 \text{ and } \kappa > 0. \tag{2}$$

We are especially interested in the infinite variance case where the tail index $\kappa \leq 2$. Heavy tailed data are widely encountered in financial, macroeconomic, actuarial, telecommunication network traffic, and meteorological time series (see Leadbetter *et al.*, 1983; Embrechts *et al.*, 1997; Finkenstädt and Rootzén, 2001; Davis, 2010, for examples and references).

If the AR errors have an infinite second moment $E[\epsilon_t^2] = \infty$ then M-estimators like least squares (LS) and least absolute deviations (LAD), and Yule–Walker and linear programming estimators are not asymptotically normal, although super-$n^{1/2}$-convergence is achievable. This topic has been thoroughly investigated. Consider Hannan and Kanter (1977) and Gross and Steiger (1979) for classic treatments, and recently An and Chen (1982); Davis and Resnick (1986); Knight (1987); Cline (1989); Davis *et al.* (1992); Feigin and Resnick (1994); Davis (1996) and Davis and Wu (1997). In an important benchmark for model (1), Davis *et al.* (1992) show a large class of smooth M-estimators like LS, as well as LAD, are $n^{1/\kappa}/L(n)$-convergent for some slowly varying function $L(n)$, in particular the LS rate is $(n/\ln(n))^{1/\kappa}$ if the distribution tails of ϵ_t are Paretian (2) (see also An and Chen, 1982; Davis and Resnick, 1986).

Robust M-estimators like least trimmed squares (LTS); (Rousseeuw, 1984; Čižek, 2008), least absolute trimmed deviations (Bassett, 1991; Tableman, 1994), maximum trimmed likelihood (Čižek, 2008) and least weighted absolute deviations (Ling, 2005) are universally based on trimming or weighting by fix quantiles of criterion equations or the data itself. See also Powell (1986); Ling (2007); Agulló *et al.* (2008) and Zhu and Ling (2012) to name a few. In the case of LTS on eqn(1) this entails trimming the squared error $\epsilon_t^2(\theta) = (y_t - \theta' x_t)^2$ by a fixed sample proportion of $\epsilon_t^2(\theta)$, or by residuals $\epsilon_t^2(\hat{\theta}_n)$ from a first-stage regression, or by y_t itself. In its purest one-step form the LTS estimator minimizes $\sum_{t=1}^n \epsilon_t^2(\theta) I(\epsilon_t^2(\theta) \leq \epsilon_{(\lfloor \lambda n \rfloor)}^2(\theta))$ where $I(A)=1$ if A is true and 0 otherwise, $\epsilon_{(1)}^2(\theta) \geq \epsilon_{(2)}^2(\theta) \geq \cdots$ are the criterion order statistics, $\lambda \in (0,1)$ is the chosen quantile and $[\lambda n]$ rounds to an integer. See notation conventions below, and see Čižek (2008) for a review and theory. If the distribution of ϵ_t is sufficiently smooth then asymptotic

[a]University of North Carolina
*Department of Economics, University of North Carolina-Chapel Hill,
†E-mail: jbhill@email.unc.edu

normality rests on each $1/n\sum_{t=1}^{n}\epsilon_t y_{t-i}I(\epsilon_t^2 \leq \epsilon_{([\lambda n])}^2)$ just like LS rests on $1/n\sum_{t=1}^{n}\epsilon_t y_{t-i}$. Since trimming is based only on ϵ_t, the regressor y_{t-i} must have a finite variance for asymptotic normality, ruling out autoregressions with infinite variance errors. The restriction to finite variance data pervades the robust M-estimation literature (e.g. Ruppert and Carroll, 1980; Neykov and Neytchev, 1990; Bassett, 1991; Chen *et al.*, 2001; Agulló *et al.*, 2008).

Ling (2005, 2007; Pan *et al.* (2007) and Zhu and Ling (2011, 2012) solve the problem for LAD, QML and Exponential QML estimation of heavy tailed AR, ARMA and ARMA-GARCH models. In each case criterion equations are weighted by a smooth stochastic function, ostensibly based on the criterion $|y_{t-i}| > c$ for some fixed threshold $c > 0$, and in the case of model (1) for each lag $i = 1, \ldots, p$. Ling (2005), for example, presents the least weighted absolute deviations (LWAD) estimator for eqn(1) under the assumption ϵ_t has a zero median. Ling (2005) uses a fixed quantile order statistic of $|y_t|$ as a plug-in for c in simulations but only proves asymptotic normality for fixed c, while the rate of convergence is $n^{1/2}$ since the weights work like fixed quantile trimming indicators. Further, LWAD does not remove the most damaging observations (those with a very larger error ϵ_t), hence it is sensitive to error extremes in small samples. See the simulation study in Section 6.

In this study, we tail-trim the squared errors $\epsilon_t^2(\theta)$ according to extreme values of the error $\epsilon_t(\theta)$ and regressors y_{t-i}. The resulting LTTS estimator $\hat{\theta}_n$ is consistent for θ^0 and asymptotically normal provided ϵ_t has a smooth and bounded distribution, and otherwise we impose mild regulatory conditions due to trimming (see Section 2).

Tail-trimming to date is used primarily for location estimation (e.g. Csörgo *et al.*, 1986; Hahn *et al.*, 1991) and moment condition tests (Hill and Aguilar, 2012; Hill, 2012). The same methods can lead to massive efficiency gains in regression model parameter estimation. Tail-trimming ensures both asymptotic normality *and* super-$n^{1/2}$-convergence. Moreover, since we trim $\epsilon_t^2(\theta)$ when $\epsilon_t(\theta)$ *or* y_{t-i} is large, our estimator is robust asymptotically *and* in small samples since the damaging effects of large errors are reduced.

Our estimator is $n^{1/\kappa}/g_n$-convergent when $\kappa \in [1,2)$ for any chosen positive sequence $\{g_n\}, g_n \to \infty$ as $n\to\infty$, that depends on the number of trimmed squared errors $\epsilon_t^2(\theta)$. The number trimmed follows simple rules of thumb that come close to LTS: trimming by the error ϵ_t should be optimized to *nearly* a fixed per cent of the sample λn, and trimming by the regressors y_{t-i} should be minimized as long as the amount increases with n. By comparison, under power law (2) with $\kappa < 2$ least squares and linear programming estimators are asymptotically non-Gaussian with respective rates $(n/\ln(n))^{1/\kappa}$ and $n^{1/\kappa}$ (An and Chen, 1982; Davis and Resnick, 1986; Davis *et al.*, 1992; Feigin and Resnick, 1994, 1999). Hence when $\kappa \in [1,2)$ LTTS obtains a rate of convergence $n^{1/\kappa}/g_n$ that is *faster* than LS since $g_n\to\infty$ can be made arbitrarily slow by controlling the amount of trimming as n increases. If $\kappa=2$ then $\hat{\theta}_n$ matches LS with rate $(n/\ln(n))^{1/\kappa}$, and if the variance is finite then $n^{1/2}$-convergence is obtained with the LS asymptotic variance: there is no loss in efficiency asymptotically due to trimming when ϵ_t has a finite variance. By contrast, for technical reasons if $\kappa < 1$ then LTTS has a rate of convergence that is slower than LS but faster than LWAD (see Section 3).

Inference mirrors classic theory although we do not require knowledge of the convergence rate (see Section 4). Although we only consider t-and Wald tests, the same robust methods extend to tests of serial correlation in the errors, and tests of omitted variables, GARCH effects and functional form (see Hill and Aguilar, 2012; Hill, 2012).

We do not maintain that by trimming within a linear model have we a universal solution to modelling heavy tailed time series. Our use of a linear AR model with i.i.d. error and LS is for convenience since they allow us to focus on the pure idea of tail-trimming for asymptotically Gaussian and therefore robust inference. Indeed, eqn(1) will not be appropriate for some heavy tailed financial and telecommunication time series given evidence of asymmetry and clustered volatility (see Resnick, 1997; Feigin and Resnick, 1999; Hall *et al.*, 2002; Tsay, 2002 for references). Nevertheless, causal and recently non-causal AR models are still used to model possibly heavy tailed time series (Ling, 2005; Aue *et al.*, 2006; Lanne *et al.*, 2012); linear vector autoregressions are de rigueur in macroeconomics since Sims (1980) and are recently used to model heavy tailed data (Lanne and Lütkepohl, 2010; Peters *et al.*, 2011); and if ϵ_t is allowed to be a non-i.i.d. martingale difference then eqn(1) also covers ARCH models (Bollerslev, 1986).

There are two major ways to model a heavy tailed stationary time series $\{y_t\}$. The first assumes a model with an additive heavy tailed i.i.d. error ϵ_t. Under linear AR model (1) with power law error (2), for example, y_t has the same tail index as ϵ_t (Brockwell and Cline, 1985). This property extends to finite Volterra expansions like a Bilinear process (Davis and Resnick, 1996) and to Log-AR Stochastic Volatility processes (Breidt and Davis, 1998; Hill, 2011a). The second assumes a nonlinear feedback structure that implies power law tail decay and therefore heavy tails even if an error term has exponentially decaying tails. Consider a stochastic recurrence equation $x_t = A_t x_{t-1} + B_t$ where x_t is a $q \times 1$ vector for $q \leq 1$, and the coefficients $\{A_t, B_t\}$ are non-degenerate, non-negative, stationary and ergodic $q \times q$ matrix processes that may be independent (Kesten, 1973) or dependent (Roitershtein, 2007). Under mild additional assumptions such processes $\{x_t\}$ have power law tails, covering linear and nonlinear GARCH and ARMA-GARCH processes (Basrak *et al.*, 2002; Liu, 2006; Cline, 2007). See also Hall *et al.* (2002) for non-parametric methods for modeling and forecasting heavy tailed time series.

The methods developed here easily generalize to multivariate models like vector autoregressions, to autoregressions with additional regressors, and to Yule–Walker estimation for eqn(1). But they also apply to nonlinear models like a random coefficient autoregression (RCA) and those with a stochastic recurrence representation like GARCH, and in general to nonlinear ARMA models with nonlinear GARCH errors; to non-stationary models like ARIMA; and to other M-estimators like nonlinear least squares, least absolute deviations, quasi-maximum likelihood, and non-Gaussian QML; and to estimators that allow over-identification like empirical likelihood. There is, however, a cost to pay for trimming, even if only a negligible sample portion is actually trimmed. For example, although our LTTS estimator beats LS when $\kappa \in (1,2)$, linear programming has a slightly higher rate, while trimming in other contexts like QML for GARCH results in a diminished rate. Of course, *we gain asymptotic normality* hence inference is simple. See Section 5 for examples concerning Yule-Walker estimation for AR, QML for GARCH, weighted LS for RCA, and Empirical Likelihood for AR, GARCH and RCA.

Finally, we use tail-trimming solely for robust parameter estimation. The observations y_t and postestimation residuals $\hat{\epsilon}_t := y_t - \hat{\theta}_n' x_t$ with the LTTS estimator $\hat{\theta}_n$ are not themselves removed from the sample. All observations are therefore available for

further study, including tail index estimation based on model residuals $\hat{\epsilon}_t$ (Resnick and Stărică, 1997), robust model specification tests that may impose different trimming strategies on $\hat{\epsilon}_t$ (Hill and Hill, 2012; Aguilar, 2012), or no trimming at all like tests of nonlinear dependence based on entropy (Robinson, 1991) and indicators (Baek and Brock, 1992).

We complete the article by studying LTTS in a controlled experiment in Section 6, and apply the estimator to financial asset returns in Section 7. Appendix A contains proofs of the main results, and all tables are relegated to the end.

We use the following notation conventions. The indicator function is $I(A)=1$ if A is true, and otherwise $I(A)=0$. The L_r-norm of a $M \times N$ matrix A is $\|A\|_r = (\sum_{i=1,j=1}^{M,N} E|A_{i,j}|^r)^{1/r}$ and the spectral norm is $\|A\| = \lambda_{\max}(A'A)^{1/2}$ with $\lambda_{\max}(A)$ the maximum eigenvalue. If z is a scalar we write $(z)_+:=\max\{0,z\}$. K denotes a positive finite constant and $\iota > 0$ is a tiny constant, the values of which may change from line to line. $x_n \sim a_n$ denotes $x_n/a_n \to 1$; $x_n = o(a_n)$ denotes $x_n/a_n \to 0$, and $x_n = o_p(a_n)$ means $x_n/a_n \xrightarrow{p} 0$. $\epsilon_t \overset{i.i.d.}{\sim} (0,1)$ states ϵ_t is i.i.d. with zero mean and unit variance. A random variable is *symmetric* if its distribution is symmetric about zero. $L(n)$ is a slowly varying function that may change with the context (i.e. $L(\xi n)/L(n) \to 1$ as $n \to \infty$ for any $\xi > 0$).

2. LEAST TAIL-TRIMMED SQUARES

The score $1/n\sum_{t=1}^n \epsilon_t(\theta)x_t$ governs least squares asymptotics, hence we need to trim $\epsilon_t^2(\theta)$ when $\epsilon_t(\theta)$ or any regressor y_{t-i} surpasses a large threshold. Our estimator is constructed as follows. Represent two-tailed observations and order statistics for any random variable z_t by

$$z_t^{(a)} := |z_t| \quad \text{and} \quad z_{(1)}^{(a)} \geq \cdots \geq z_{(n)}^{(a)} \geq 0.$$

The determination of the number of trimmed large errors $\epsilon_t(\theta)$ and regressors y_{t-i} is made by intermediate order sequences $\{k_n^{(\epsilon)}, k_n^{(y)}\}$. If $\{k_n^{(z)}\}$ denotes either sequence then $1 \leq k_n^{(z)} < n$, $k_n^{(z)} \to \infty$ and $k_n^{(z)}/n \to 0$ (see Leadbetter *et al.*, 1983). Now define a composite selection function $\hat{I}_{n,t}(\theta)$:

$$\hat{I}_{n,t}(\theta) = I\left(|\epsilon_t(\theta)| \leq \epsilon_{(k_n^{(\epsilon)})}^{(a)}(\theta)\right) \times \prod_{i=1}^p I\left(|y_{t-i}| \leq y_{(k_n^{(y)})}^{(a)}\right) = \hat{I}_{n,t}^{(\epsilon)}(\theta) \times \hat{I}_{n,t-1}^{(y)}.$$

The LTTS estimator solves

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t^2(\theta) \hat{I}_{n,t}(\theta) \right\}$$

where $\Theta \subset \mathcal{R}^{p+1}$, is a compact parameter space.

Notice only those observations $\{y_t, x_t\}$ with non-extreme error $\epsilon_t(\theta)$ and regressors y_{t-i} enter the criterion. Further, each $k_n^{(z)}$ represents the number of trimmed criterion equations $\epsilon_t^2(\theta)$ due to large $\epsilon_t(\theta)$ or y_{t-i}, while at most $k_n^{(\epsilon)} + pk_n^{(y)}$ observations are trimmed. Negligibility $k_n^{(z)}/n \to 0$ implies we trim asymptotically a vanishing sample portion of observations.

In post-estimation, however, we have available the *untrimmed* residuals $\epsilon_t(\hat{\theta}_n) = y_t - \hat{\theta}_n' x_t = \epsilon_t(\hat{\theta}_n) = \epsilon_t - (\hat{\theta}_n - \theta^0)' x_t$ which can be used for model specification tests. Thus, trimming is proposed here solely for parameter estimation, although trimming can be used for robust model specification tests as well (see Hill 2012; Hill and Aguilar, 2012).

Asymptotic theory for $\hat{\theta}_n$ requires the non-random sequences $\{c_n^{(\epsilon)}(\theta), c_n^{(y)}\}$ which the order statistics $\{\epsilon_{(k_n^{(\epsilon)})}^{(a)}(\theta), y_{(k_n^{(y)})}^{(a)}\}$ estimate. Let $\{c_n^{(\epsilon)}(\theta), c_n^{(y)}\}$ be the exact quantiles defined by

$$P\left(|\epsilon_t(\theta)| > c_n^{(\epsilon)}(\theta)\right) = \frac{k_n^{(\epsilon)}}{n} \quad \text{and} \quad P(|y_t| > c_n^{(y)}) = \frac{k_n^{(y)}}{n}, \tag{3}$$

and the composite selection function is

$$I_{n,t}(\theta) := I\left(|\epsilon_t(\theta)| \leq c_n^{(\epsilon)}(\theta)\right) \times \prod_{i=1}^p I\left(|y_{t-i}| \leq c_n^{(y)}\right) = I_{n,t}^{(\epsilon)}(\theta) \times I_{n,t-1}^{(y)}.$$

Distribution continuity ensures the existence of such thresholds $\left\{c_n^{(\epsilon)}(\theta), c_n^{(y)}\right\}$ for any $\left\{k_n^{(\epsilon)}, k_n^{(y)}\right\}$. See below for all assumptions. Clearly $\left\{\epsilon_{(k_n^{(\epsilon)})}^{(a)}(\theta), y_{(k_n^{(y)})}^{(a)}\right\}$ estimate $\left\{c_n^{(\epsilon)}(\theta), c_n^{(y)}\right\}$, and under regularity conditions presented below intermediate order statistics are uniformly consistent, e.g. $\sup_{\theta \in \Theta} |\epsilon_{(k_n^{(\epsilon)})}^{(a)}(\theta)/c_n^{(\epsilon)}(\theta) - 1| = O_p(1/(k_n^{(\epsilon)})^{1/2})$ (see Appendix B).

Throughout we drop θ^0 and write $\epsilon_t = \epsilon_t(\theta^0)$, $c_n^{(\epsilon)} = c_n^{(\epsilon)}(\theta^0)$, $I_{n,t} = I_{n,t}(\theta^0)$ and so on.

2.1. Assumptions

The following assumptions ensure stationarity, restrict the distribution of ϵ_t and restrict the amount of trimming.

ASSUMPTION 1 (*stationarity*). *The roots of* $1 - \sum_{i=1}^{p} \phi_i^0 z^i$ *lie outside the unit circle, and* $\theta^0 = [\xi^0, \phi^{0\prime}]'$ *lies in the interior of a compact subset* $\Theta \subset \mathbb{R}^{p+1}$.

ASSUMPTION 2 (*errors*). *The distribution of* ϵ_t *is absolutely continuous with respect to Lebesgue measure, bounded* $\sup_{a\in\mathbb{R}}(\partial/\partial a)P(\epsilon_t \leq a) < \infty$, *symmetric at zero, and exhibits power-law tail decay (2) with finite scale* $d > 0$ *and tail index* $\kappa > 0$.

ASSUMPTION 3 (*fractiles*). *a.* $k_n^{(\epsilon)} k_n^{(y)}/n \to \infty$; *b. if* $\kappa \in (0,1)$ *then* $k_n^{(\epsilon)} k_n^{(y)}/n^{2-\kappa/(2-\kappa)} \to \infty$.

REMARK 1: Power law tail decay, independence, and stationarity imply y_t has a power law tail with the same index $\kappa > 0$ (e.g. Brockwell and Cline, 1985, cf. Embrechts *et al.*, 1997):

$$P(|y_t| > a) = d\sum_{i=0}^{\infty} |\psi_i|^{\kappa} a^{\kappa}(1 + o(1)) \quad \text{as } a \to \infty \tag{4}$$

where $\{\psi_i\}_{i=0}^{\infty}$ satisfies $\sum_{i=0}^{\infty} \psi_i z^i = (1 - \sum_{i=1}^{p} \phi_i^0 z^i)^{-1}$ for complex z, ψ_0=1 and $\psi_i = O(\rho^i)$ for some $\rho \in (0,1)$.

REMARK 2: In order to prove consistency and therefore asymptotic normality we require a law of large numbers for the trimmed least squares score $1/n \sum_{t=1}^{n} \epsilon_t x_t I_{n,t} \xrightarrow{p} 0$ (cf. Pakes and Pollard, 1989). This follows by independence of the error and Chebyshev's inequality if we restrict the tail-trimmed variance $\|E[\epsilon_t^2 x_t x_t' I_{n,t}]\| = o(n)$. The latter holds when the mean is finite $\kappa > 1$ or hairline infinite κ=1 under Assumption 3a, and otherwise holds under the Assumption 3b fractile restriction. Notice 3a implies trimming cannot be too light, for example we cannot have both $k_n^{(z)} \sim \ln(n)$. Further, 3b implies more trimming is required as the error tails become exceptionally thick: if $\kappa < 1$ then as $\kappa \searrow 0$ we require both $k_n^{(z)} \nearrow n$. Although more general fractile conditions can be used, the cost is lengthy proofs of technical results. In practice neither property will reduce generality by much since many time series in economics and finance appear to satisfy $\kappa \geq 1$, and letting $k_n^{(\epsilon)} \sim n/g_n^{(\epsilon)}$ and $k_n^{(y)} \sim g_n^{(y)}$ for sequences $\{g_n^{(\epsilon)}, g_n^{(y)}\}$ that increase $g_n^{(\cdot)} \to \infty$ as slowly as we choose optimizes the LTTS rate of convergence, while setting $g_n^{(\epsilon)}/g_n^{(y)} \to 0$ ensures 3a holds (see Section 3).

2.2. Main results

We state the main results here and relegate proofs to Appendix A. Consistency follows from well known arguments for non-differentiable criteria (e.g. Pakes and Pollard, 1989). We must prove consistency first in order to establish the expansion $\mathcal{V}_n^{1/2}(\hat{\theta}_n - \theta^0) = n^{-1/2}\Sigma_n^{-1/2} \sum_{t=1}^{n} \epsilon_t x_t I_{n,t}$ for asymptotic normality, where $\Sigma_n := E[\epsilon_t^2 x_t x_t' I_{n,t}]$.

THEOREM 1 (consistency). *Under Assumptions* 1–3 $\hat{\theta}_n \xrightarrow{p} \theta^0$.

Asymptotic normality follows by proving the above expansion and showing $n^{-1/2}\Sigma_n^{-1/2} \sum_{t=1}^{n} \epsilon_t x_t I_{n,t}$ satisfies a Gaussian central limit theorem.

THEOREM 2 (*normality*). *Under Assumptions 1–3* $\mathcal{V}_n^{1/2}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N(0, I_{p+1})$ *where* $\mathcal{V}_n = n(E[\epsilon_t^2 I_{n,t}^{(\epsilon)}])^{-1} \times E[x_t x_t' I_{n,t-1}^{(y)}]$.

REMARK 1: Since the error is independent the covariance matrix $\mathcal{V}_n^{-1} = E[\epsilon_t^2 I_{n,t}^{(\epsilon)}] \times (E[x_t x_t' I_{n,t-1}^{(y)}])^{-1}$ has the classic least squares form. The exact form of $\mathcal{V}_n$ in the case of heavy tails is treated in Section 3.

REMARK 2: The matrix $E[x_t x_t' I_{n,t-1}^{(y)}]$ is positive definite and therefore invertible for sufficiently large n since distribution non-degeneracy and trimming negligibility imply $\liminf_{n\to\infty} \inf_{\lambda'\lambda=1} E[(\lambda' x_t)^2 I_{n,t-1}^{(y)}] > 0$ where $\lambda \in \mathbb{R}^{p+1}$.

If the errors have a finite variance $E[\epsilon_t^2] = \sigma^2 < \infty$ then by stationarity and dominated convergence $\mathcal{V}_n \sim nE[x_t x_t']/\sigma^2$, the classic least square result. Tail-trimming has no impact on asymptotics if the variance is finite.

COROLLARY 1 (finite variance). *Under Assumptions 1–3 and* $E[\epsilon_t^2] = \sigma^2 < \infty$ *it follows* $n^{1/2}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N\left(0, \sigma^2 (E[x_t x_t'])^{-1}\right)$.

3. RATE OF CONVERGENCE

Let $\mathcal{V}_{i,j,n}$ denote the $(i,j)^{\text{th}}$ component of the matrix $\mathcal{V}_n$, thus $\mathcal{V}_n = [\mathcal{V}_{i,j,n}]_{i,j=1}^{p+1}$. Similarly, let $\hat{\theta}_{i,n}$ and θ_i^0 be i^{th} components of these vectors. Apply Theorem 2 to deduce under Assumptions 1–3 the component-wise limit

$$\mathcal{V}_{i,i,n}^{1/2}(\hat{\theta}_{i,n} - \theta_i^0) \xrightarrow{d} N(0,1).$$

In view of the construction $\mathcal{V}_n = n(E[\epsilon_t^2 I_{n,t}^{(\epsilon)}])^{-1} \times E[x_t x_t' I_{n,t-1}^{(y)}]$, the rate of convergence $\mathcal{V}_{i,i,n}^{1/2}$ of $\hat{\theta}_{i,n}$ can be easily characterized by exploiting dominated convergence and Karamata's theorem. First, by construction, $I(|y_t| \leq c_n^{(y)}) \to 1$ a.s. and $\prod_{i=1}^p I(|y_{t-i}| \leq c_n^{(y)}) \leq I(|y_{t-j}| \leq c_n^{(y)})$. Hence for any $j=1,\ldots,p$ by dominated convergence and stationarity $E[\prod_{i=1}^p I(|y_{t-i}| \leq c_n^{(y)})] = 1 + o(1)$ and

$$\begin{aligned} &E\left[y_{t-j}^2 \prod_{i=1}^p I\left(|y_{t-i}| \leq c_n^{(y)}\right)\right] \\ &= E\left[y_{t-j}^2 I\left(|y_{t-j}| \leq c_n^{(y)}\right)\right] \times \left(1 - \frac{E\left[y_{t-j}^2\left(\prod_{i=1}^p I\left(|y_{t-i}| \leq c_n^{(y)}\right) - I\left(|y_{t-j}| \leq c_n^{(y)}\right)\right)\right]}{E\left[y_{t-j}^2 I\left(|y_{t-j}| \leq c_n^{(y)}\right)\right]}\right) \\ &= E\left[y_t^2 I\left(|y_t| \leq c_n^{(y)}\right)\right] \times (1 + o(1)). \end{aligned} \tag{5}$$

There are, therefore, two rates of convergence: one $\mathcal{V}_{\xi,n}^{1/2} := \mathcal{V}_{1,1,n}^{1/2}$ for the intercept and one $\mathcal{V}_{\phi,n}^{1/2} := \mathcal{V}_{i,i,n}^{1/2}$ for any slope $i=2,\ldots,p+1$, where by eqn(5) we have

$$\mathcal{V}_{\xi,n}^{1/2} \sim \frac{n^{1/2}}{\left(E\left[\epsilon_t^2 I\left(|\epsilon_t| \leq c_n^{(\epsilon)}\right)\right]\right)^{1/2}} \quad \text{and} \quad \mathcal{V}_{\phi,n}^{1/2} \sim n^{1/2}\left(\frac{E\left[y_t^2 I\left(|y_t| \leq c_n^{(y)}\right)\right]}{E\left[\epsilon_t^2 I\left(|\epsilon_t| \leq c_n^{(\epsilon)}\right)\right]}\right)^{1/2}. \tag{6}$$

Second, by construction of the thresholds (3) and tail decay (2) and (4), the thresholds are

$$c_n^{(\epsilon)} = d^{1/\kappa}\left(n/k_n^{(\epsilon)}\right)^{1/\kappa} \quad \text{and} \quad c_n^{(y)} = d^{1/\kappa}\left(\sum_{i=0}^{\infty}|\psi_i|^\kappa\right)^{1/\kappa}\left(n/k_n^{(y)}\right)^{1/\kappa}. \tag{7}$$

Third, since each $z_t \in \{\epsilon_t, y_t\}$ has tail (2) or (4) with the same index κ and some scale d_z, we have by and (7) and Karamata's theorem (e.g. Resnick, 1987; Theorem 0.6)[1]

$$\text{if } \kappa \in (0,2) \text{ then } E\left[z_t^2 I\left(|z_t| \leq c_n^{(z)}\right)\right] \sim \frac{\kappa}{2-\kappa}\left(c_n^{(z)}\right)^2 P\left(|z_t| > c_n^{(z)}\right) = \frac{\kappa}{2-\kappa} d_z^{2/\kappa}\left(\frac{n}{k_n^{(z)}}\right)^{2/\kappa-1}. \tag{8}$$

$$\text{if } \kappa = 2 \textit{ then } E\left[z_t^2 I\left(|z_t| \leq c_n^{(z)}\right)\right] \sim d_z \ln(n) \textit{for any intermediate order} \{k_n^{(z)}\}.$$

Combine eqns(6)–(8) to obtain a complete characterization of the intercept and slope rates. We drop multiplicative constants since these do not affect the rates.

THEOREM 3 (*RATES OF CONVERGENCE*). *Let Assumptions* 1–2 *hold.*

a. If $\kappa > 2$ then $\mathcal{V}_{\xi,n}^{1/2}, \mathcal{V}_{\phi,n}^{1/2} = n^{1/2}$.

b. If $\kappa=2$ then $\mathcal{V}_{\xi,n}^{1/2}, \mathcal{V}_{\phi,n}^{1/2} \sim (n/\ln(n))^{1/2}$ for any intermediate order sequence $\{k_n^{(z)}\}$.

c. if $\kappa < 2$ then $\mathcal{V}_{\xi,n}^{1/2} = n^{1/2}(k_n^{(\epsilon)}/n)^{1/\kappa-1/2}$ and $\mathcal{V}_{\phi,n}^{1/2} = n^{1/2}(k_n^{(\epsilon)}/k_n^{(y)})^{1/\kappa-1/2}$.

If the error tail index $\kappa < 2$ then the slope rate $\mathcal{V}_{\phi,n}^{1/2} = Kn^{1/2}(k_n^{(\epsilon)}/k_n^{(y)})^{1/\kappa-1/2}$ depends inversely on error and regressor heavy tails, and therefore inversely on $\{k_n^{(\epsilon)}, k_n^{(y)}\}$. Large errors appear as outliers and therefore reduce estimation accuracy, so heavy trimming (fast $k_n^{(\epsilon)} \to \infty$) augments the rate. Leverage points in terms of large regressors y_{t-i}, however, help identify θ^0 and therefore increase the rate, so light trimming by the regressors (slow $k_n^{(y)} \to \infty$) is optimal.

Evidently this two-fold logic has never been exploited for the sake of M-estimation, yet remarkably it allows us to obtain a convergence rate that beats LS and is comparable to the highest possible amongst M-estimators. Keeping in mind that Assumption 3a requires $k_n^{(\epsilon)}k_n^{(y)}/n \to \infty$, in general for any positive sequences $\{g_n^{(\epsilon)}, g_n^{(y)}\}$ where $g_n^{(1)} \leq 1$, $g_n^{(\cdot)} \to \infty$ as slow as we choose and $g_n^{(\epsilon)}/g_n^{(y)} \to 0$, simply put

$$k_n^{(\epsilon)} \sim n/g_n^{(\epsilon)} \quad \text{and} \quad k_n^{(y)} \sim g_n^{(y)}$$

to satisfy Assumption 3a, and when $\kappa < 2$ to achieve a slope rate

$$\mathcal{V}_{\phi,n}^{1/2} = Kn^{1/\kappa}\left(\frac{1}{g_n^{(\epsilon)}g_n^{(y)}}\right)^{1/\kappa-1/2}.$$

Notice we can make $g_n^{(\cdot)} \to \infty$ as slow as we choose and still satisfy Assumption 3a, but not Assumption 3b in the very heavy tail case $\kappa < 1$. Thus, the rate $\mathcal{V}_{\phi,n}^{1/2}$ can be made as close to $n^{1/\kappa}$ as we choose when $\kappa \in [1,2)$, hence $\mathcal{V}_{\phi,n}^{1/2} \to \infty$ can be made faster that the

least squares rate $(n/\ln(n))^{1/\kappa}$ when $\kappa \in [1,2)$, (cf. Davis *et al.*, 1992). Moreover, for very slow $g_n^{(\cdot)} \to \infty$ the number of trimmed squared errors $\epsilon_t^2(\theta)$ is very close to a fixed quantile and governed strongly by error extremes, as in LTS. By trimming slightly less than for LTS and using error *and* regressor extremes to decide which $\epsilon_t^2(\theta)$ to trim, we can achieve asymptotic normality and not only super-$n^{1/2}$-convergence, but a rate that beats least squares in the infinite variance case $\kappa \in [1,2)$.

In practice if it is assumed $\kappa \geq 1$ then we may consider fractile functions of the form $k_n^{(\epsilon)} = [\lambda_\epsilon n/\ln(n)]$ and $k_n^{(y)} = [\lambda_y(\ln(n))^{1+\iota}]$ for any $\lambda_\epsilon, \lambda_y \in (0,1]$ and infinitessimal $\iota > 0$. Assumption 3a holds and if the error variance is infinite $\kappa \in [1,2)$ then

$$\mathcal{V}_{\phi,n}^{1/2} = K\frac{n^{1/\kappa}}{(\ln(n))^{(1+\iota)(1/\kappa-1/2)}}\left(\frac{\lambda_\epsilon}{\lambda_y}\right)^{1/\kappa-1/2} > K\left(\frac{n}{\ln(n)}\right)^{1/\kappa} \quad \text{as } n \to \infty$$

The rate can be forced up by increasing trimming by the error $\lambda_\epsilon\uparrow$ and decreasing trimming by the regressors $\lambda_y\downarrow$. Of course this can similarly be achieved by using $\ln(\ln(n))$ instead of $\ln(n)$, and so on.

If $\kappa < 1$ is possible then the above $k_n^{(\epsilon)}$ and $k_n^{(y)}$ do not satisfy Assumption 3b. However, Assumption 3a and 3b are satisfied if, for example, $k_n^{(\epsilon)} = [\lambda_\epsilon n/(\ln(n))^a]$ and $k_n^{(y)} = [\lambda_y n/(\ln(n))^b]$ for any $0 < a \leq b$. In this case $\mathcal{V}_{\phi,n}^{1/2} = (\lambda_\epsilon/\lambda_y)^{1/\kappa-1/2}n^{1/2}(\ln(n))^{(b-a)(1/\kappa-1/2)}$ which is superior to $n^{1/2}$ when $b > a$ such that trimming by the error is harsher. In general, in keeping with the Assumption 3b fractile bound, LTTS will have a rate slower than LS but higher than LWAD when $\kappa < 1$.

4. INFERENCE

A natural estimator of the scale $\mathcal{V}_n$ is simply

$$\hat{\mathcal{V}}_n = n\left(\frac{1}{n}\sum_{t=1}^n x_t x_t' \hat{l}_{n,t-1}^{(y)}\right) \times \left(\frac{1}{n}\sum_{t=1}^n \epsilon_t^2(\hat{\theta}_n)\hat{l}_{n,t}(\hat{\theta}_n).\right)^{-1}.$$

Notice $1/n\sum_{t=1}^n \epsilon_t^2(\hat{\theta}_n)\hat{l}_{n,t}(\hat{\theta}_n)$ uses the composite trimming indicator $\hat{l}_{n,t}(\hat{\theta}_n)$ rather than the error specific one $\hat{l}_{n,t}^{(\epsilon)}(\hat{\theta}_n)$. The reason is the squared residual expands as $\epsilon_t^2(\hat{\theta}_n) = \epsilon_t^2 + f((\hat{\theta}_n - \theta^0)'x_t)$ where $f\colon \mathbb{R} \to \mathbb{R}$ is polynomial in its argument: in finite samples large values of $\epsilon_t(\hat{\theta}_n)$ are caused by ϵ_t *and* each y_{t-i}.

THEOREM 4 (scale consistency). *Under Assumptions* 1–3 $\mathcal{V}_n^{-1}\hat{\mathcal{V}}_n \xrightarrow{p} I_p$.

A self-normalized t-ratio $\hat{\tau}_i$ for a test of the hypothesis $H_0 : \theta_i^0 = \theta_i^*$ is simply

$$\hat{\tau}_i = \hat{\mathcal{V}}_{i,i,n}^{1/2}\left(\hat{\theta}_{n,i} - \theta_i^*\right).$$

As long as Assumptions 1–3 hold, Theorems 2 and 4 imply under the null $\hat{\tau}_i \xrightarrow{d} N(0,1)$, and $|\hat{\tau}_i| \to \infty$ with probability one if $\theta_i^0 \neq \theta_i^*$.

Similarly, we can construct a Wald statistic for a test of linear $R\theta^0 = q$ and nonlinear $C(\theta^0) = 0$ restrictions. Consider the former with $R \in \mathbb{R}^{J\times(p+1)}$ and $q \in \mathbb{R}^J$ for some $J \leq 1$, where R has linearly independent rows. The statistic is

$$\hat{W} = \left(R\hat{\theta}_n - q\right)'\left(R\hat{\mathcal{V}}_n^{-1}R'\right)^{-1}\left(R\hat{\theta}_n - q\right).$$

Under the null $\hat{W} \xrightarrow{d} \chi_J^2$ a centered chi-squared distribution with J degrees of freedom. A test of white noise $H_0 : \phi^0 = 0$ against AR(p) simply uses $R = \mathbf{0}|I_p$ and $q = \mathbf{0}$ with zero vector $\mathbf{0} \in \mathbb{R}^p$. Denote this Wald statistic $\hat{W}_p$.

5. EXTENSIONS

Tail-trimming can be used for robust estimation of models of nonlinear, non-stationary, and conditionally heteroscedastic processes; it can be used to robustify M-estimators; and can be used for robust specification tests. In the following we present four examples covering Yule–Walker estimation of AR, GARCH estimated by QML, a possibly non-stationary random coefficient autoregression estimated by weighted least squares, and the empirical likelihood method for confidence region computation for AR, GARCH and RCA. Linear and nonlinear GARCH processes have power law tails under mild assumptions and therefore serve as alternatives to eqn(1) as models for heavy tailed time series. See Hill (2012) for a robust asymptotic power one test of functional form based on tail-trimming that can be used to test whether eqn(1) neglects nonlinear traits of the process $\{y_t\}$. Also consult Hill and Aguilar (2012) for tail-trimmed moment condition tests that cover robust tests of linear dependence, lag order and omitted variables.

In each case below ϵ_t is an i.i.d. random variable with a continuous distribution that is positive on $\mathbb{R}$ and symmetric about zero, and $E|\epsilon_t|^\iota < \infty$ for some $\iota > 0$. Symmetry can be easily relaxed at the expense of notation. Let $\{k_n\}$ be an intermediate order sequence and let the sequence $\{c_n\}$ satisfy $P(|\epsilon_t| \geq c_n) = k_n/n \to 0$.

5.1. Yule–Walker Estimation

Assume for simplicity the intercept in eqn(1) is zero: $y_t = \sum_{i=1}^{p} \phi_i^0 y_{t-i} + \epsilon_t$. If $E[\epsilon_t^4] < \infty$ then the sample autocovariance $\hat{\gamma}_h := 1/n \sum_{t=1}^{n} y_t y_{t-h}$ is $n^{1/2}$-convergent and asymptotically normal. If $E[\epsilon_t^4] = \infty$ and ϵ_t^2 belongs to the domain of attraction of a stable law (e.g. eqn (2) holds with $\kappa \in 0,4]$) then under a different scaling $\hat{\gamma}_h$ is asymptotically a ratio of stable laws (Davis and Resnick, 1986), hence Yule–Walker equations can be used for non-Gaussian estimation of the autoregression parameters ϕ^0.

We can, however, negligibly trim y_t itself for Gaussian inference by way of tail-trimmed Yule–Walker equations. Define trimmed variables $\hat{y}_{n,t}^* := y_t I(|y_t| \leq y_{(k_n)}^{(a)})$ and $y_{n,t}^* := y_t I(|y_t| \leq c_n)$, and trimmed covariances $\hat{\gamma}_{n,h}^* := 1/n \sum_{t=1}^{n} \hat{y}_{n,t}^* \hat{y}_{n,t-h}^*$ and $\gamma_{n,h}^* := E[y_{n,t}^* y_{n,t-h}^*]$ with displacement $h \geq 1$. By dominated convergence and negligibility $I_{n,t}^{(\epsilon)} \times \prod_{i=0}^{p} I_{n,t-i}^{(y)} \to 1$ *a.s.* hence

$$E\left[y_t \times I_{n,t}^{(\epsilon)} \times \prod_{i=0}^{p} I_{n,t-i}^{(y)} \times y_{n,t-h}^*\right] = E\left[y_{n,t}^* y_{n,t-h}^*\right] + E\left[y_{n,t}^* y_{n,t-h}^* \left(I_{n,t}^{(\epsilon)} \times \prod_{i=1}^{p} I_{n,t-i}^{(y)} - 1\right)\right] = \gamma_{n,h}^* \times (1 + o(1)).$$

Similarly, by independence and distribution symmetry $E[\epsilon_t I_{n,t}^{(\epsilon)} y_{n,t-h}^*] = 0$ hence by negligibility and dominated convergence $E[\epsilon_t I_{n,t}^{(\epsilon)} \prod_{i=0}^{p} I_{n,t-i}^{(y)} y_{n,t-h}^*] = E[\epsilon_t I_{n,t}^{(\epsilon)} y_{n,t-h}^*] + E[\epsilon_t I_{n,t}^{(\epsilon)} y_{n,t-h}^* (\prod_{i=0}^{p} I_{n,t-i}^{(y)} - 1)] = o(1)$. Now multiply y_t and $\sum_{i=1}^{p} \phi_i^0 y_{t-i} + \epsilon_t$ by $I_{n,t}^{(\epsilon)} \prod_{i=0}^{p} I_{n,t-i}^{(y)} y_{n,t-h}^*$, take expectations and repeat the above logic to deduce $\gamma_{n,h}^* = \sum_{i=1}^{p} \phi_i^0 \gamma_{n,h-i}^* (1 + o(1)) + o(1)$ hence the solution $\phi^0 = ([\gamma_{n,i-j}^*]_{i,j=0}^{p-1})^{-1} [\gamma_{n,i}^*]_{i=1}^{p} (1 + o(1))$. Trimming removes information such that ϕ^0 may not be exactly $([\gamma_{n,i-j}^*]_{i,j=0}^{p-1})^{-1} [\gamma_{n,i}^*]_{i=1}^{p}$, but even if the variance is infinite we have asymptotic identification by tail-trimmed covariances $\phi^0 = \lim_{n\to\infty} ([\gamma_{n,i-j}^*]_{i,j=0}^{p-1})^{-1} [\gamma_{n,i}^*]_{i=1}^{p}$. Notice in the i.i.d. case $\gamma_{n,h}^* = 0$ hence the solution is exactly ϕ^0=0. A valid estimator of the limit ϕ^0 is the tail-trimmed Yule-Walker estimator $\hat{\phi}_n = ([\hat{\gamma}_{n,i-j}^*]_{i,j=0}^{p-1})^{-1} [\hat{\gamma}_{n,i}^*]_{i=1}^{p}$.

Consider the AR(1) case for simplicity: $\hat{\phi}_n = \hat{\gamma}_{n,1}^* / \hat{\gamma}_{n,0}^*$. If $E[\epsilon_t^2] = \infty$ then assume ϵ_t has power law tail (2). The theory developed in the appendices and the fact that $\gamma_{n,1}^* / \gamma_{n,0}^* \sim \phi^0$ can be exploited to show $n^{1/2}(\gamma_{n,0}^* / \mathcal{S}_n)(\hat{\phi}_n - \phi^0) \xrightarrow{d} N(0,1)$ where $\mathcal{S}_n^2 := E(1/n^{1/2} \sum_{t=1}^{n} \{y_{n,t}^* y_{n,t-1}^* - E[y_{n,t}^* y_{n,t-1}^*]\})^2$. The rate of convergence is easy to see for the case ϕ^0=0. Since y_t is i.i.d. we have $\mathcal{S}_n^2 = (\gamma_{n,0}^*)^2$ hence $\hat{\phi}_n$ is $n^{1/2}$-convergent, making it inferior to LTTS when ϵ_t has a tail index $\kappa < 2$.

5.2. GARCH with tail-trimmed QML

The GARCH(1,1) model is $y_t = \sigma_t \epsilon_t$ where $\sigma_t^2 = \omega^0 + \alpha^0 y_{t-1}^2 + \beta^0 \sigma_{t-1}^2$, $\omega^0 > 0, \alpha^0, \beta^0 \in (0,1), E[\epsilon_t] = 0$, $E[\epsilon_t^2] = 1$, and $E[\ln(\alpha^0 \epsilon_t^2 + \beta^0)] < 0$. The variable y_t is stationary and geometrically β-mixing (Nelson, 1990; Carrasco and Chen, 2002), and has a power law tail due to a stochastic recurrence type feedback, irrespective of the tail decay of ϵ_t (Basrak *et al*, 2002, Section 3).

Now define $\sigma_t^2(\theta) = \omega + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2(\theta)$, $\epsilon_t(\theta) := y_t / \sigma_t^2(\theta)$ and $s_t(\theta) := (\partial/\partial\theta)\sigma_t^2(\theta)/\sigma_t^2(\theta)$ for θ in Θ, a compact subset of $(0,\infty) \times (0,1) \times (0,1)$, and assume θ^0 is an interior point of Θ . Write $s_t = s_t(\theta^0)$. Since $\sup_{\theta \in \mathcal{N}_0} \|s_t(\theta)\|$ is square integrable on some neighbourhood $\mathcal{N}_0$ of θ^0, the QML score $1/n^{1/2} \sum_{t=1}^{n} (\epsilon_t^2 - 1) s_t$ will be asymptotically normal if ϵ_t has a finite fourth moment (Francq and Zakoïan, 2004). Although Zhu and Ling's (2011) double exponential weighted QML (DQML) estimator is asymptotically normal even when ϵ_t has an infinite fourth moment, in principle it may be sensitive to error extremes in small samples because large errors enter the criterion. An estimator that is robust in large *and* small samples can be constructed by tail-trimming QML equations according to error extremes, hence $\hat{\theta}_n = \arg\min_{\theta \in \Theta} \{\sum_{t=1}^{n} (\ln \sigma_t^2(\theta) + \epsilon_t^2(\theta)) \times \hat{I}_{n,t}^{(\epsilon)}(\theta)\}$.

Asymptotics are governed by the tail-trimmed squared error $\epsilon_t^2 I_{n,t}^{(\epsilon)}$, thus if $E[\epsilon_t^4] = \infty$ then assume power law (2) applies with index $\kappa \in (2,4]$. Arguments in the appendices can be generalized to show $V_n^{1/2}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N(0, I_3)$ where $V_n = [V_{i,n}] := nE[s_t s_t'] / E[(\epsilon_t^2 - 1)^2 I(|\epsilon_t| \leq c_n)]$. Thus, if ϵ_t has an infinite fourth moment then each $V_{i,n}^{1/2} = o(n^{1/2})$ which is less than the DQML rate $n^{1/2}$ (Zhu and Ling, 2011). Nevertheless, $V_{i,n}^{1/2}$ can be made arbitrarily close to $Kn^{1/2}$ by choosing $k_n \to \infty$ as close to λn as we want, hence tail-trimmed QML can be assured to be faster than QML (see Theorem 2.1 in Hall and Yao, 2003). This is the same error effect we find with LTTS: by pushing k_n arbitrarily close to the fixed quantile rate λn we then optimize the convergence rate by minimizing the negative impact of error extremes. Tail-trimmed QML leads to a better QML estimator just like LTTS trumps LS when tails are heavy. In both cases, however, there is a cost to pay: trimming reduces the rate below some *other* estimator, in this case linear programming for eqn(1) and DQML for GARCH.

5.3. RCA and weighted LTTS

The random coefficient AR(1) model is $y_t = (\phi^0 + b_t) y_{t-1} + \epsilon_t$, where b_t is i.i.d. and may be dependent on ϵ_t, and $E[b_t^2]$ is positive and finite. If $E[\epsilon_t^2] = \infty$ then assume power law (2) applies. It is well known if $E[\ln|\phi^0 + b_t|] < 0$ then a strictly stationary solution exists (see Aue *et al.*, 2006).

The parameter ϕ^0 can be estimated by weighted least squares with criterion function $\sum_{t=1}^{n} \epsilon_t^2(\phi) \mathcal{W}_{t-1}$ and weight $\mathcal{W}_{t-1} := 1/(1 + y_{t-1}^2)$. The weight $\mathcal{W}_{t-1}$ counteracts explosive sample paths hence we only need $E[\ln \max\{0, |\phi^0 + b_t|\}] < \infty$, which allows non-stationary cases (Aue *et al.*, 2006; Chan *et al.*, 2012). The minimand $\hat{\phi}_n = \sum_{t=1}^{n} y_t y_{t-1} \mathcal{W}_{t-1} / \sum_{t=1}^{n} y_{t-1}^2 \mathcal{W}_{t-1}$ is consistent for ϕ^0 and asymptotically normal if ϵ_t is i.i.d. with a zero mean and finite variance since $1/n \sum_{t=1}^{n} y_{t-1}^2 \mathcal{W}_{t-1} \times n^{1/2}(\hat{\phi}_n - \phi^0) = 1/n^{1/2} \sum_{t=1}^{n} (b_t y_{t-1} + \epsilon_t) y_{t-1} \mathcal{W}_{t-1}$.

Clearly $|y_t \mathcal{W}_t|$ is bounded *a.s.* In non-stationary cases $y_t^2 \mathcal{W}_t \xrightarrow{p} 1$ as $t \to \infty$, hence $n^{1/2}(\hat{\phi}_n - \phi^0) = 1/n^{1/2} \sum_{t=1}^{n} b_t + o_p(1)$ is easily verified (Aue *et al.*, 2006; Chan *et al.*, 2012). Thus, $\hat{\phi}_n$ is asymptotically normal even if $E[\epsilon_t^2] = \infty$.

Although non-stationary y_t is allowed, if we want to allow a heavy tailed error ϵ_t and retain Gaussian asymptotic inference in general, then a robust estimation strategy like tail-trimming is required. The weight $\mathcal{W}_{t-1}$ acts to robustify against heavy tailed y_{t-1}, hence we need only trim by ϵ_t. The Weighted LTTS estimator solves $\hat{\theta}_n = \arg\min_{\phi\in\mathbb{R}}\{\sum_{t=1}^n \epsilon_t^2(\phi)\mathcal{W}_{t-1} \times I(|\epsilon_t(\phi)| \leq \epsilon^{(a)}_{(k_n)}(\phi))\}$ where $\epsilon_t(\phi) := y_t - \phi y_{t-1}$. By using arguments in the appendices and in Aue *et al.* (2006) and Chan *et al.* (2012) it is straightforward to show $V_n^{1/2}(\hat{\phi}_n - \phi^0) \xrightarrow{d} N(0,1)$ where $V_n := n(1/n\sum_{t=1}^n y_{t-1}^2\mathcal{W}_{t-1})^2/\{1/n\sum_{t=1}^n y_{t-1}^2\mathcal{W}_{t-1}^2 \times E[\epsilon_t^2 I(|\epsilon_t| \leq c_n)]\}$.

5.4. Empirical Likelihood Method

The empirical likelihood method (ELM) allows for computation of confidence regions without estimating a covariance structure, and allows for over-identifying restrictions. The method requires equations $\mathcal{Z}_t : \Theta \to \mathbb{R}^q$ for $q \geq k$ that identify the parameter of interest $\theta^0 \in \Theta \subset \mathbb{R}^k$ by the moment condition $E[\mathcal{Z}_t(\theta)] = 0$ *if and only if* $\theta = \theta^0$. If $\mathcal{Z}_t := \mathcal{Z}_t(\theta^0)$ is a stationary finite variance martingale difference then under mild additional assumptions Wilks' Theorem applies (see Owen, 2001). In the presence of heavy tails Wilks' theorem will not apply hence the ELM is not valid for asymptotic inference.

Tail-trimming, however, can robustify against heavy tails. In model (1) we may use the LTTS score equation with added instruments, say $\epsilon_t(\theta)\hat{I}^{(\epsilon)}_{n,t}(\theta)z_t\hat{I}^{(z)}_{n,t}$ where $z_t = [1, y_{t-1}, \ldots, y_{t-q-1}]'$ for some $q \geq p+1$. In the GARCH case from Section 2 we may use tail-trimmed QML score equations with added instruments $(\epsilon_t^2(\theta)\hat{I}^{(\epsilon)}_{n,t}(\theta) - 1)z_t(\theta)$ where $z_t(\theta) \in \mathbb{R}^q$ for $q \geq 3$ contains $s_t(\theta)$ and possibly added lags $s_{t-i}(\theta)$, $i=1,\ldots,l$ for finite $l \leq 0$. In the RCA case we may use the WLTTS score equation $\epsilon_t(\phi)y_{t-1}\mathcal{W}_{t-1}I(|\epsilon_t(\phi)| \leq \epsilon^{(a)}_{(k_n)}(\phi))$. (If over-identifying restrictions are imposed then a weight other than $\mathcal{W}_{t-1}$ may be required.)

6. SIMULATION STUDY

We now compare the small sample properties of LS, LTS, LTTS, and Ling's (2005) LWAD. We draw $n \in \{100,400,800\}$ random variables y_t from an AR(2) model $y_t = 0.2 + 0.8y_{t-1} - 0.3y_{t-2} + \epsilon_t$. The errors ϵ_t are i.i.d. symmetric Pareto distributed $P(\epsilon_t > \epsilon) = P(\epsilon_t < -\epsilon) = 0.5 \times (1+\epsilon)^{-\kappa}$ with index $\kappa \in \{0.75,1.5,2.5\}$ spanning infinite mean, finite mean with infinite variance, and finite variance cases. We simulate 10,000 series $\{y_t\}_{t=1}^n$ for each n and κ.

We compute the LTTS estimator with fractiles $k_n^{(\epsilon)} = \max\{1, [.05n/\ln(n)]\}$ and $k_n^{(y)} = \max\{1, [0.01n/(\ln(n))^2]\}$ in order to satisfy Assumption 3, and to achieve a rate $n^{1/2}(\ln(n))^{(1/\kappa-1/2)}$ that is greater than LWAD's rate $n^{1/2}$ when $\kappa < 2$. We do not use convergence rate elevating functions like $k_n^{(\epsilon)} \sim \lambda_\epsilon n/\ln(n)$ and $k_n^{(\epsilon)} \sim \lambda_\epsilon \ln(n)$ since this pair does not satisfy Assumption 3b when κ=0.75. Nevertheless, a pair like $k_n^{(\epsilon)} = [0.05n/\ln(n)]$ and $k_n^{(y)} \sim \max\{1, [0.1\ln(n)]\}$ results in the same trimming amount as our chosen pair above for sample sizes $n \in \{100,400,800\}$ since $k_n^{(y)}$ is very small in both cases.[2]

The LTS estimator is $\arg\min_{\theta\in\Theta}\{\sum_{t=1}^n \epsilon_t^2(\theta)I(\epsilon_t^2(\theta) \leq \epsilon^2_{([\lambda n])}(\theta))\}$ with λ=0.05 as in Čižek (2008). The LWAD estimator is $\arg\min_{\theta\in\Theta}\{\sum_{t=1}^n w_t|\epsilon_t(\theta)|\}$ with Ling's (2005, eqn. 2.3) chosen weight w_t based on Huber's (1977) influence function. Define $a_t := \sum_{i=1}^p |y_{t-i}|I(|y_{t-i}| \leq y^{(a)}_{([\lambda n])})$: the weight is w_t=1 if a_t=0 and $w_t = (y^{(a)}_{([\lambda n])})^3/a_t^3$ if $a_t \neq 0$, and λ=0.05. The parameter space for all estimators is Θ=[−1,1][3].

See Table 1 for the simulation bias, mean-squared-error and Kolmogorov–Smirnov tests of normality for the slope estimator $\hat{\theta}_{n,3}$ of $\theta_3^0 = -0.3$ (the omitted results being similar). The KS test is based on the standardization $(\hat{\theta}_{n,3} - \theta_3^0)/s_n$ where s_n^2 is the empirical variance of $\hat{\theta}_{n,3}$. The empirical mean of each estimator is accurate. LS and LTS estimators fail normality tests when variance is infinite, as expected. LTTS is closest to normal in general, while LWAD is roughly on par with LS or somewhere between LS and LTTS when $\kappa < 2$. Only LTTS is robust in *both* small and large samples by virtue of trimming observations with large errors. Indeed, although LWAD is asymptotically robust to error extremes, it exhibits a larger mean-squared-error and KS statistic than LTTS in most cases, suggesting it is sensitive to large errors in small samples. LS does not weight the errors in any sense, so its mean-squared-error is the greatest when $\kappa < 2$.

In a second experiment we simulate a variety of AR(1) and AR(2) models, estimate AR(2) models, and compute $\hat{W} = (R\hat{\theta}_n - q)'(R\hat{\mathcal{V}}_n^{-1}R')^{-1}(R\hat{\theta}_n - q)$ for tests of AR(1) against AR(2), hence R=[0,0,1] and q=0. We use the covariance estimator $\hat{\mathcal{V}}_n$ from Section 4. See Table 2 for model descriptions and empirical sizes and powers. Wald tests based on a LTTS plug-in perform well under either hypothesis. Empirical sizes are near the nominal level, and empirical powers are predominantly above 90%, and near 100% when the alternative is far from the null or n is large.

7. EMPIRICAL APPLICATION

We now analyze financial returns data. We use the same Hang Seng Index (HSI) stock market data Ling (2005) investigated for the sake of comparison. The period is 3 June, 1996 to 31 May, 1998 representing 491 daily observations, net of market closures.[3] Consult Ling (2005) for details on the HSI.

We generate a log-returns series $y_t : y_t = \ln(x_t/x_{t-1})$ where x_t are daily closing values on the HSI. See Figure 1 for a plot of y_t. Define $y_t^a := |y_t|$. The case for heavy tails can be made by a plots of the Hill (1975) two-tailed tail index estimator $\hat{\kappa}_{r_n} = (1/r_n\sum_{i=1}^{r_n}\ln(y^a_{(i)}/y^a_{(r_n+1)}))^{-1}$ over fractiles $r_n \in \{5,2,\ldots,200\}$. Although we assume an AR model with i.i.d. error (1), in fact as long as $r_n \to \infty$ and $r_n = o(n)$ it is known $\hat{\kappa}_{r_n} \xrightarrow{p} \kappa$ and $r_n^{1/2}(\hat{\kappa}_{r_n} - \kappa) \xrightarrow{d} N(0, v_\kappa^2)$, $v_\kappa^2 < \infty$, for a truly vast array of time series, including

Table 1. AR(2) Estimation Results ($\theta_2^\circ = -0.3$)

	Tail Index $\kappa = 0.75$											
	$n = 100$				$n = 400$				$n = 800$			
	Bias	MSE	KS[a]	%[b]	Bias	MSE	KS	%	Bias	MSE	KS	%
LS	−0.003	0.004	4.75	0.000	−0.001	0.0009	6.50	0.000	0.000	0.0006	6.79	0.000
LTS	0.005	0.001	6.17	0.050	0.003	0.0008	10.0	0.050	0.002	0.0003	10.5	0.050
LTTS	0.000	0.003	3.51	0.030	0.001	0.0016	1.92	0.030	0.001	0.0004	1.08	0.020
LWAD	0.000	0.021	6.17	0.050	0.000	0.0001	2.09	0.050	0.000	0.0001	1.03	0.050
	Tail Index $\kappa = 1.5$											
	Bias	MSE	KS	%	Bias	MSE	KS	%	Bias	MSE	KS	%
LS	−0.007	0.007	2.52	0.000	0.001	0.0015	3.18	0.000	0.000	0.0007	2.68	0.000
LTS	0.009	0.004	2.85	0.050	−0.002	0.0003	2.50	0.050	0.001	0.0001	2.42	0.050
LTTS	−0.000	0.006	1.64	0.030	0.001	0.0019	0.842	0.031	0.000	0.0008	0.763	0.021
LWAD	0.001	0.040	2.31	0.050	0.000	0.0001	1.61	0.050	0.000	0.0004	1.05	0.050
	Tail Index $\kappa = 2.5$											
	Bias	MSE	KS	%	Bias	MSE	KS	%	Bias	MSE	KS	%
LS	−0.008	0.008	1.94	0.000	−0.002	0.0020	1.26	0.000	− 0.001	0.0003	1.13	0.000
LTS	−0.001	0.004	1.32	0.050	0.001	0.0008	0.868	0.050	0.001	0.0002	0.816	0.050
LTTS	−0.001	0.007	0.758	0.030	−0.001	0.0022	0.737	0.032	0.000	0.0008	0.527	0.021
LWAD	−0.001	0.050	1.03	0.050	0.001	0.0001	0.842	0.050	0.000	0.0001	0.632	0.050

LTTS = Least Tail-Trimmed Squares; LTS = Least Trimmed Squares; LS = Least Squares; LWAD = Least Weighted Absolute Deviations.
a. Kolmogorov-Smirnov statistic for a test of standard normality on standardized $\hat{\theta}_{n,2}$, divided by the 5% critical value: values above 1 imply rejection at the 5% level.
b. The total sample proportion trimmed for LTS and LTTS (Tr% = 0.05 for LTS by construction). In the case of LWAD this represents the percentile used in the weight or 0.05, cf. Ling (2005).

Table 2. LTTS Wald-test of AR(1) vs. AR(2)[a]

		Tail index κ=0.75			Tail index κ=1.5			Tail index κ=2.5		
		n=100			n=100			n=100		
θ_1^0	θ_2^0	10%	0.5%	1%	10%	0.5%	1%	10%	0.5%	1%
0.80	0.00	0.058	0.031	0.008	0.110	0.057	0.012	0.011	.053	.012
0.80	−0.20	0.454	0.333	0.165	0.577	0.462	0.225	0.602	0.495	0.283
0.80	−0.30	0.756	0.654	0.442	0.849	0.772	0.571	0.862	0.781	0.582
		n=400			n=400			n=400		
θ_1^0	θ_2^0	10%	0.5%	1%	10%	0.5%	1%	10%	0.5%	1%
0.80	0.00	0.064	0.041	0.021	0.093	0.054	0.010	0.105	0.051	0.013
0.80	−0.20	0.899	0.812	0.613	0.949	0.897	0.760	0.954	0.912	0.813
0.80	−0.30	0.975	0.944	0.915	0.993	0.993	0.982	0.994	0.994	0.991
		n=800			n=800			n=800		
θ_1^0	θ_2^0	10%	0.5%	1%	10%	0.5%	1%	10%	0.5%	1%
0.80	0.00	0.074	0.049	0.019	0.093	0.049	0.009	0.101	0.049	0.010
0.80	−0.20	0.961	0.938	0.853	0.997	0.991	0.983	0.999	0.994	0.991
0.80	−0.30	0.992	0.985	0.971	0.999	0.998	0.997	0.999	0.998	0.996

[a] We simulate AR(2) models $y_t = 0.2 + \theta_1^0 y_{t-1} + \theta_2^0 y_{t-2} + \epsilon_t$ and test the hypothesis $\theta_2^0 = 0$.

AR with linear or nonlinear GARCH shocks with geometric or hyperbolic memory decay (see Hill, 2010, 2011b and the citations therein). Further, Hill (2010, Theorem 3) presents a consistent kernel estimator $\hat{v}_\kappa^2$ of the asymptotic variance v_κ^2 of $\hat{\kappa}_{r_n}^{-1}$:

$$\hat{v}_n^2 = \frac{1}{n}\sum_{s,t=1}^{n} w_{n,s,t}\left\{\ln\left(\frac{y_s^{(a)}}{y_{(r_n+1)}^{(a)}}\right)_+ - \frac{r_n}{n}\hat{\kappa}_{r_n}^{-1}\right\}\times\left\{\ln\left(\frac{y_t^{(a)}}{y_{(r_n+1)}^{(a)}}\right)_+ - \frac{r_n}{n}\hat{\kappa}_{r_n}^{-1}\right\}$$

where $w_{n,s,t}$ is a kernel function. We use a Bartlett kernel $w_{n,s,t} = (1 - |s-t|/\gamma_n)_+$ with bandwidth $\gamma_n = n^{0.225}$. By the mean-value-theorem the asymptotic 95% confidence band for $\hat{\kappa}_{r_n}$ is $\hat{\kappa}_{r_n} \pm 1.96\hat{v}_n\hat{\kappa}_{r_n}^2/r_n^{1/2}$ (see Figure 3). Values of $\kappa \leq 2$ lie in the 95% intervals at every r_n.

Since a benchmark question is whether asset returns are white noise we estimate AR(6), AR(8) and AR(12) models by LTTS and compute Wald statistics $\hat{W}_p$ for tests that all slopes are zero over $p \in \{6,8,12\}$. All AR models in this study include an intercept, and

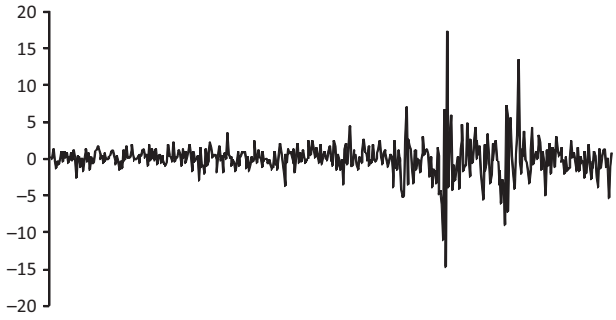

Figure 1. HSI daily log-returns

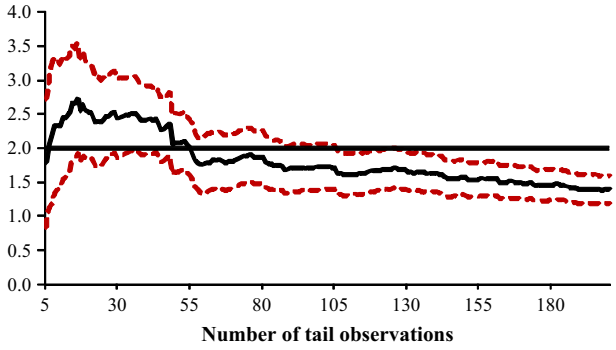

Figure 2. Hill-Plot and robust 95% bands

the LTTS fractiles are $k_n^{(\epsilon)} = [0.05n/\ln(n)]$ and $k_n^{(y)} = \max\{1, [0.01n/(\ln(n))^2]\}$ as in Section 6. The statistics are $\hat{W}_6 = 45.02$ (0.000), $\hat{W}_8 = 43.99$ (0.000), and $\hat{W}_{12} = 52.58$ (0.000) with p-values in parentheses, hence white noise is rejected. We then estimate AR(p) models over p=1,...,12 and test the residuals $\hat{\epsilon}_t = y_t - \hat{\theta}_n' x_t$ for white noise by computing $\hat{W}_{12}$ on $\hat{\epsilon}_t$. All AR(p) models, $p \leq 3$, have white noise residuals, where AR(3) $y_t = \theta_0 + \sum_{i=1}^{3} \theta_i y_{t-i} + \epsilon_t$ results in a residuals test $\hat{W}_{12} = 9.88$ (0.628), while AR(2) $\hat{W}_{12} = 27.62$ (0.006) and AR(1) $\hat{W}_{12} = 18.45$ (0.102). Wald tests of AR(2) against AR(3) and AR(3) against AR(4) lead to the same conclusion: an AR(3) best describes the data, with LTTS estimates (standard errors in parentheses)

$$\hat{y}_t = \underset{(0.08)}{-0.10} + \underset{(0.05)}{0.25} y_{t-1} - \underset{(0.05)}{0.02} y_{t-2} + \underset{(0.06)}{0.11} y_{t-3}\,.$$

The result is robust to higher order specifications. An AR(7), for example, is

$$\hat{y}_t = \underset{(0.08)}{-0.07} + \underset{(0.05)}{0.28} y_{t-1} - \underset{(0.05)}{0.01} y_{t-2} + \underset{(0.05)}{0.13} y_{t-3} + \underset{(0.05)}{0.01} y_{t-4} - \underset{(0.07)}{0.04} y_{t-5} + \underset{(0.06)}{0.04} y_{t-6} + \underset{(0.09)}{0.02} y_{t-7}$$

Finally, in the AR(3) model we test separately whether the first lag y_{t-1}, or second lag y_{t-2}, or both $\{y_{t-1}, y_{t-2}\}$ do not belong. The resulting Wald statistic values are 28.6 (0.000), 0.165 (0.685) and 28.9 (0.000) suggesting the appropriate model is $y_t = \theta_0 + \theta_1 y_{t-1} + \theta_3 y_{t-3} + \epsilon_t$.

Ling's (2005) chosen model by similar Wald tests based on LWAD is also AR(3) but with only the third lag: $y_t = \theta_0 + \theta_3 y_{t-3} + \epsilon_t$. Further, the two sets of estimates are somewhat different: Ling's (2005) AR(7) estimates are

$$\hat{y}_t = \underset{(0.06)}{-0.07} + \underset{(0.04)}{0.07} y_{t-1} - \underset{(0.04)}{0.00} y_{t-2} + \underset{(0.04)}{0.11} y_{t-3} + \underset{(0.04)}{0.03} y_{t-4} - \underset{(0.04)}{0.08} y_{t-5} + \underset{(0.04)}{0.02} y_{t-6} + \underset{(0.04)}{0.09} y_{t-7}$$

By comparison, we obtain a larger and significant estimate of the first order lag y_{t-1}

8. CONCLUSION

We present the LTTS estimator for possibly very heavy tailed autoregressions where the squared errors are negligibly trimmed based on large values of the error and regressors. The estimator is consistent for the true parameter and asymptotically normal, and

super-$n^{1/2}$-convergent for an appropriate choice of the trimming fractiles when the variance is infinite. We can, moreover, always choose the fractiles such that our estimator obtains the highest possible convergence rate for M-estimators of stationary data. A simulation study reveals LTTS dominates LS, LTS and LWAD based on approximate normality, and therefore on small sample inference based on the asymptotic distribution. Tail trimming extends to a variety of linear and nonlinear models of the conditional mean and variance, as well to other criteria like QML and Empirical Likelihood. We present brief examples, but deeper results are left for future research.

APPENDIX A: PROOFS OF MAIN RESULTS

It is helpful to define trimmed normal equations $m_t(\theta)$, their short- and long–run variances $\Sigma_n(\theta)$ and $\mathcal{S}_n(\theta)$, and Jacobian $G_n(\theta)$:

$$
\begin{aligned}
&m_t(\theta) := \epsilon_t(\theta)x_t = (y_t - \theta' x_t)x_t \\
&m_{n,t}(\theta) := m_t(\theta) \times I_{n,t}(\theta) \quad \text{and} \quad \hat{m}_{n,t}(\theta) := m_t(\theta) \times \hat{I}_{n,t}(\theta) \\
&\hat{m}_n(\theta) := \frac{1}{n}\sum_{t=1}^{n} \hat{m}_{n,t}(\theta) \quad \text{and} \quad m_n(\theta) := \frac{1}{n}\sum_{t=1}^{n} m_{n,t}(\theta) \\
&\Sigma_n(\theta) := E\left[m_{n,t}(\theta)m_{n,t}(\theta)'\right] \quad \text{and} \quad G_n := -E\left[x_t x_t' I_{n,t-1}^{(y)}\right] \\
&\mathcal{S}_n(\theta) := \frac{1}{n}\sum_{s,t=1}^{n} E\left[m_{n,s}(\theta)m_{n,t}(\theta)'\right]
\end{aligned}
$$

As usual we drop θ^0. The proofs of consistency and asymptotic normality Theorems 1 and 2 require supporting lemmas. Consistency requires variance bounds, asymptotic bounds on $\hat{m}_{n,t}(\theta) - m_{n,t}(\theta)$, and laws of large numbers.

LEMMA A1 (asymptotic approximation). *(a)* $n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^{n}\{\hat{m}_{n,t} - m_{n,t}\} = o_p(1)$;
(b) $\sup_{\theta\in\Theta}\{\|1/n\sum_{t=1}^{n}(\hat{m}_{n,t}(\theta) - m_{n,t}(\theta))\|\} = o_p(\sup_{\theta\in\Theta} E\|m_{n,t}(\theta)\|)$;
(c) $1/n\sum_{t=1}^{n}\epsilon_t^2\{\hat{I}_{n,t} - I_{n,t}\} = o_p(1)$;
(d) $1/n\sum_{t=1}^{n}x_t x_t'\{\hat{I}_{n,t} - I_{n,t}\} = o_p(1)$.

LEMMA A2 (variance bound). $\Sigma_n = o(n)$.

LEMMA A3 (LLN AND ULLN). *(a)* $1/n\sum_{t=1}^{n} m_{n,t} = o_p(1)$;
(b) $\sup_{\theta\in\Theta}\{\|1/n\sum_{t=1}^{n} m_{n,t}(\theta) - E[m_{n,t}(\theta)]\|\} = o_p(\sup_{\theta\in\Theta} E\|m_{n,t}(\theta)\|)$.

Asymptotic normality requires an asymptotic Taylor expansion, a central limit theorem, and Jacobian consistency. Define

$$\tilde{G}_n(\theta) := -\frac{1}{n}\sum_{t=1}^{n} x_t x_t' I_{n,t}(\theta) \quad \text{and} \quad \hat{G}_n(\theta) := -\frac{1}{n}\sum_{t=1}^{n} x_t x_t' \hat{I}_{n,t}(\theta).$$

LEMMA A4 (asymptotic expansion). *Let* $\theta, \tilde{\theta} \in \Theta$ *be arbitrary:*

(a) $1/n\sum_{t=1}^{n}\{m_{n,t}(\theta) - m_{n,t}(\tilde{\theta})\} = \tilde{G}_n(\theta) \times (\theta - \tilde{\theta}) + o_p(\|G_n\| \times \|\theta - \tilde{\theta}\|)$; and
(b) $1/n\sum_{t=1}^{n}\{\hat{m}_{n,t}(\theta) - \hat{m}_{n,t}(\tilde{\theta})\} = \hat{G}_n(\theta) \times (\theta - \tilde{\theta}) + o_p(\|G_n\| \times \|\theta - \tilde{\theta}\|)$;
(c) $1/n\sum_{t=1}^{n}\epsilon_t^2(\hat{\theta}_n)\{\hat{I}_{n,t}(\hat{\theta}_n) - \hat{I}_{n,t}\} = o_p(1)$;
(d) $1/n\sum_{t=1}^{n}x_t x_t'\{\hat{I}_{n,t}(\hat{\theta}_n) - \hat{I}_{n,t}\} = o_p(1)$.

LEMMA A5 (CLT). $n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^{n} m_{n,t} \xrightarrow{d} N(0, I_{p+1})$.

LEMMA A6 (Jacobian properties). *(a)* $(\partial/\partial\theta)E[m_{n,t}(\theta)]|_{\theta^0} = G_n \times (1 + o(1))$;
(b) $\sup_{\theta\in\Theta} E\|m_{n,t}(\theta)\| \leq K\|G_n\|$;
(c) $\hat{G}_n(\hat{\theta}_n) = G_n \times (1 + o_p(1))$;
(d) $1/n\sum_{t=1}^{n}x_t x_t' I_{n,t-1}^{(y)} = -G_n \times (1 + o_p(1))$.

See Appendix B for proofs of Lemmas A1–A6. We are now ready to prove Theorems 1 and 2.

PROOF OF THEOREM 1. *The proof of consistency follows an argument in Pakes and Pollard* (1989)*: Theorem* 3.1, *Corollary* 3.2). *Define* $\mathcal{M}_n(\theta) := \mathrm{E}[\mathrm{m}_{n,t}(\theta)]$ *and* $\mathfrak{e}_n := \sup_{\theta\in\Theta} E\|m_{n,t}(\theta)\|$. *We will first prove for any small* $\delta > 0$.

$$\epsilon(\delta) := \inf_{n\geq N}\inf_{\theta\in\Theta,\|\theta-\theta^0\|>\delta}\left\{\mathfrak{e}_n^{-1}\times\|\mathcal{M}_n(\theta)\|\right\} > 0. \tag{9}$$

By the definition of a derivative and Lemma A6a $E[m_{n,t}(\theta)] = G_n \times (\theta - \theta^0) \times (1 + o(1))$, and by Lemma A6b the Jacobian G_n satisfies $\sup_{\theta\in\Theta}E\|m_{n,t}(\theta)\| \leq K\|G_n\|$. Further, G_n is non-singular for each $n \leq N$ and some $N \in \mathbb{N}$ since by distribution non-degeneracy and trimming negligibility $I_{n,t-1}^{(y)} \xrightarrow{a.s} 1$ we have

$$\liminf_{n\to\infty}\inf_{r\in\mathbb{R}^{p+1}:r'r=1} r'E\left[x_t x_t' I_{n,t-1}^{(y)}\right]r = \inf_{r'r=1} E\left(\sum_{i=1}^{p} r_i x_{i,t} I_{n,t-1}^{(y)}\right)^2 > 0, \tag{10}$$

hence $\|G_n\| > 0 \forall n \leq N$. This delivers bound (9):

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

In view of eqn(9), since $P(\|\hat\theta_n - \theta^0\| > \delta) \leq P(\mathfrak{e}_n^{-1}\|\mathcal{M}_n(\hat\theta_n)\| > \epsilon(\delta))$ it suffices to show $\|\mathcal{M}_n(\hat\theta_n)\| = o_p(\mathfrak{e}_n)$ to prove $\|\hat\theta_n - \theta^0\| \xrightarrow{p} 0$. By Minkowski's inequality

$$\left\|\mathcal{M}_n(\hat\theta_n)\right\|/\mathfrak{e}_n \leq \left\|\hat m_n(\hat\theta_n)\right\|/\mathfrak{e}_n + \left\|\hat m_n(\hat\theta_n) - \mathcal{M}_n(\hat\theta_n)\right\|/\mathfrak{e}_n = \mathcal{A}_n(\hat\theta_n) + \mathcal{B}_n(\hat\theta_n),$$

say. Consider $\mathcal{A}_n(\hat\theta_n)$. The following utilizes arguments in Čižek (2008, Lemma 2.1 and p. 29). By distribution continuity and linearity, $\hat Q_n(\theta) := 1/n\sum_{t=1}^n \hat\epsilon_{n,t}^2(\theta)$ is differentiable at $\hat\theta_n$ with probability one, hence 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 $\hat Q_n(\hat\theta_n) \leq \hat Q_n(\theta)\forall\theta\in\Theta$ 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 $\mathcal{A}_n(\hat\theta_n) = 0$ *a.s.*

Next, $\mathcal{B}_n(\hat\theta_n) \leq \sup_{\theta\in\Theta}\{\mathcal{B}_n(\theta)\}$. Combine $\sup_{\theta\in\Theta}\{\|\hat m_n(\theta) - m_n(\theta)\|/\mathfrak{e}_n\} = o_p(1)$ by Lemma A1b and $\sup_{\theta\in\Theta}\{\|m_n(\theta) - \mathcal{M}_n(\theta)\|/\mathfrak{e}_n\} = o_p(1)$ by ULLN Lemma A3b to deduce

$$\sup_{\theta\in\Theta}\{\mathcal{B}_n(\theta)\} \leq \sup_{\theta\in\Theta}\left\{\frac{\|\hat m_n(\theta) - m_n(\theta)\|}{\mathfrak{e}_n}\right\} + \sup_{\theta\in\Theta}\left\{\frac{\|m_n(\theta) - \mathcal{M}_n(\theta)\|}{\mathfrak{e}_n}\right\} = o_p(1).$$ □

PROOF OF THEOREM 2. *By the proof of Theorem 1* $\hat\theta_n$ *satisfies* $1/n\sum_{t=1}^n \hat m_{n,t}(\hat\theta_n) = 0$ *a.s. Apply expansion Lemma A4b to deduce*

$$\hat G_n(\hat\theta_n)\left(\hat\theta_n - \theta^0\right) + o_p\left(\|G_n\|\times\left\|\hat\theta_n - \theta^0\right\|\right) + \frac{1}{n}\sum_{t=1}^n \hat m_{n,t}(\theta^0) = 0 a.s. \tag{11}$$

By Lemma A6c $\hat G_n(\hat\theta_n) = G_n(1 + o_p(1)) = -E[x_t x_t' I_{n,t}] \times (1 + o_p(1))$, and by trimming negligibility and error independence $E[x_t x_t' I_{n,t}] = E[x_t x_t' I_{n,t-1}^{(y)}] \times (1 + o(1))$. Further, by error independence $\Sigma_n = E[\epsilon_t^2 I_{n,t}^{(\epsilon)}] \times E[x_t x_t' I_{n,t-1}^{(y)}]$, hence in view of trimming negligibility and (10) it follows Σ_n is non-singular. Mow multiply both sides of (11) by $\Sigma_n^{-1/2}$, rearrange terms and use the fact that $\mathcal{V}_n = nE[x_t x_t' I_{n,t-1}^{(y)}] \times (E[\epsilon_t^2 I_{n,t}^{(\epsilon)}])^{-1} \sim nG_n'\Sigma_n^{-1}G_n$ by error independence to deduce

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

The claim now follows from approximation Lemma A1a and CLT Lemma A5: $\mathcal{V}_n^{1/2}(\hat\theta_n - \theta^0) = -n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^n m_{n,t} \times (1 + o_p(1)) \xrightarrow{d} N(0, I_{p+1})$. □

PROOF OF THEOREM 4. *In view of Jacobian consistency Lemma A6d we only need to show* $1/n\sum_{t=1}^n \epsilon_t^2(\hat\theta_n)\hat I_{n,t}(\hat\theta_n) = E[\epsilon_t^2 I_{n,t}] \times (1 + o_p(1))$. *By Lemmas A1c and A4c* $1/n\sum_{t=1}^n \epsilon_t^2(\hat\theta_n)\hat I_{n,t}(\hat\theta_n) = 1/n\sum_{t=1}^n \epsilon_t^2 I_{n,t} + o_p(1)$, *and by stationarity, ergodicity and the fact that* $\epsilon_t^2 I_{n,t}/E[\epsilon_t^2 I_{n,t}]$ *is integrable we have* $1/n\sum_{t=1}^n \epsilon_t^2 I_{n,t}/E[\epsilon_t^2 I_{n,t}] \xrightarrow{p} 1$.

APPENDIX B: PROOFS OF LEMMAS A1–A6

To reduce notation, in most proofs we only consider the AR(1) case without an intercept. In this case $m_t(\theta) = \epsilon_t(\theta)y_{t-1}$, $\theta=\phi$, $I_{n,t-1}^{(y)} = I(|y_{t-1}| \leq c_n^{(y)})$, $\Sigma_n = E[\epsilon_t^2 I_{n,t}^{(\epsilon)}] \times E[y_{t-1}^2 I_{n,t-1}^{(y)}]$ and $G_n = E[y_{t-1}^2 I_{n,t-1}^{(y)}]$ The general AR(p) case is essentially identical.

Throughout we write $w_t(\theta)$ to denote $\epsilon_t(\theta)$ or y_t, and $c_n(\theta)$ to denote $c_n^{(\epsilon)}(\theta)$ or $c_n^{(y)}$. We repeatedly use the following properties. Under Assumptions 1 and 2 $w_t(\theta)$ is is geometrically β-mixing (Pham and Tran, 1985) and uniformly L_ι-bounded for tiny $\iota > 0$. By Assumption 1 it is easily verified that $\epsilon_t(\theta) = \epsilon_t + (\theta^0 - \theta)y_{t-1} = \sum_{i=0}^{\infty} \tilde{\psi}_i(\theta)\epsilon_{t-i}$ where $\tilde{\psi}_i(\theta)$ is continuous, differentiable, and $\sup_{\theta\in\Theta}|\tilde{\psi}_i(\theta)| = O(\rho^i)$ for some $\rho \in (0,1)$. Therefore by Assumption 2 either $w_t(\theta)$ satisfies (cf. Brockwell and Cline, 1985)

$$\begin{aligned} &\lim_{a\to\infty} \sup_{\theta\in\Theta}\{|c^\kappa P(|w_t(\theta)| > a) - d_w(\theta)|\} = 0 \\ &\inf_{\theta\in\Theta}\{d_w(\theta)\} > 0 \quad \text{and} \quad \sup_{\theta\in\Theta}\{d_w(\theta)\} < \infty, \end{aligned} \tag{12}$$

and $c_n(\theta)$ satisfies

$$c_n(\theta) = d_w(\theta)^{1/\kappa}\left(\frac{n}{k_n}\right)^{1/\kappa}. \tag{13}$$

Therefore, by eqn(12) and Karamata's Theorem

$$\begin{aligned} &\text{if } \kappa = 2 \text{ then } \sup_{\theta\in\Theta}\left\{\frac{\ln(n)}{E[w_t^2(\theta)I(|w_t(\theta)| \le c_n(\theta))]}\right\} \to K \in (0,\infty) \\ &\text{if } \kappa < 2 \text{ then } \sup_{\theta\in\Theta}\left\{\frac{n}{k_n}\frac{c_n^2(\theta)}{E[w_t^2(\theta)I(|w_t(\theta)| \le c_n(\theta))]}\right\} \to K \in (0,\infty). \end{aligned} \tag{14}$$

The proofs of Lemmas A1–A6 require two supporting results. First, trimming indicators satisfy a uniform law.

LEMMA B1 (uniform indicator law). *Define* $\mathcal{I}_{n,t}(\theta) := ((n/k_n)^{1/2})\{I(|w_t(\theta)| \le c_n(\theta)) - E[I(|w_t(\theta)| \le c_n(\theta))]\}$. *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* $\Rightarrow^*$ *denotes weak convergence on a Polish space.*

PROOF. By construction $\mathcal{I}_{n,t}(\theta)$ is L_2-bounded uniformly on $1 \le t \le n$, $n \le 1$, and Θ, and geometrically β-mixing. Further, $\{\mathcal{I}_{n,t}(\theta) : \theta \in \Theta\}$ satisfies the metric entropy with L_2-bracketing bound $\int_0^1 \ln(N_{[\,]}(\varepsilon,\Theta,\|\cdot\|_2))\mathrm{d}\varepsilon < \infty$ with L_2-bracketing numbers $N_{[]}(\varepsilon,\Theta,\|\cdot\|_2)$. This follows since $w_t(\theta)$ have absolutely continuous distributions by linearity and Assumption 2, hence the thresholds $c_n(\theta)$ are continuous. Further, $w_t(\theta)$ have bounded distributions uniformly on Θ by linearity and Assumption 2: $\sup_{\theta\in\Theta}\sup_{a\in\mathbb{R}}|(\partial/\partial\theta)P(w_t(\theta) \le a)| < \infty$. Therefore $\mathcal{I}_{n,t}(\theta)$ is L_2-Lipschitz: $E[(\mathcal{I}_{n,t}(\theta) - \mathcal{I}_{n,t}(\tilde{\theta}))^2] \le K\|\theta - \tilde{\theta}\|$. Proving the L_2-bracketing numbers satisfy $\int_0^1 \ln(N_{[]}(\varepsilon,\Theta,\|\cdot\|_2))\mathrm{d}\varepsilon < \infty$ is then a classic exercise (Giné and Zinn, 1984; Pollard, 1984). We may therefore apply Doukhan *et al.* (1995) Theorem 1; eq. (2.17), Application 4) uniform central limit theorem to deduce $\{1/n^{1/2}\sum_{t=1}^n \mathcal{I}_{n,t}(\theta) : \theta \in \Theta\} \Longrightarrow^* \{\mathcal{I}(\theta) : \theta \in \Theta\}$. □

Second, intermediate order statistics are uniformly bounded in probability.

LEMMA B2 (uniform order statistic). *Write* $w_t^{(a)}(\theta) := |w_t(\theta)|$. *Then* $\sup_{\theta\in\Theta}|w_{(k_n)}^{(a)}(\theta)/c_n(\theta) - 1| = O_p(k_n^{-1/2})$.

PROOF. We first prove a pointwise limit, and then the uniform limit. Assume for notational simplicity $\inf_{\theta\in\Theta} w_t(\theta) \le 0$ hence $w_t^{(a)}(\theta) = w_t(\theta)$.

Step 1 (pointwise): Drop θ and define $\mathcal{I}_n(u/k_n^{1/2}) := 1/k_n\sum_{t=1}^n I(w_t > c_n e^{u/k_n^{1/2}})$ for arbitrary $u \in \mathbb{R}$. In view of geometric β-mixing and power-law tail decay, $\{k_n^{-1/2}I(w_t > c_n e^u)\}$ satisfies the conditions of Hill's (2009: Theorem 2.1, Lemma 3.1) central limit theorem. Therefore point-wise $k_n^{1/2}\{\mathcal{I}_n(u/k_n^{1/2}) - E\{\mathcal{I}_n(u/k_n^{1/2})\} \xrightarrow{d} N(0, v_1^2(u))$, where $v_1 : \mathbb{R}_+ \to \mathbb{R}_+$, and $\sup_{u\ge 0} v_1(u) < \infty$. Since $u/k_n^{1/2} \to 0$ it therefore follows for any u

$$k_n^{1/2}\left\{\mathcal{I}_n(u/k_n^{1/2}) - E\{\mathcal{I}_n(u/k_n^{1/2})\}\right\} \xrightarrow{d} N(0, v_1^2(0)), \qquad \text{where } v_1(0) < \infty \tag{15}$$

. We will show $k_n^{1/2}\ln(w_{(k_n)}/c_n) \xrightarrow{d} N(0, v_2^2)$ for some $v_2^2 > 0$ follows from eqn(15). By construction $k_n^{1/2}\ln(w_{(k_n)}/c_n) \le u$ for $u \in \mathbb{R}$ sufficiently if $\mathcal{I}_n(u/k_n^{1/2}) \le 1$, while $\mathcal{I}_n(u/k_n^{1/2}) \le 1$ if

$$\begin{aligned} k_n^{1/2}\left(\mathcal{I}_n(u/k_n^{1/2}) - E\left[\mathcal{I}_n(u/k_n^{1/2})\right]\right) &\le k_n^{1/2}\left(1 - \frac{n}{k_n}P\left(w_t > c_n e^{u/k_n^{1/2}}\right)\right) \\ &= k_n^{1/2}\left(1 - \frac{P\left(w_t > c_n e^{u/k_n^{1/2}}\right)}{P(w_t > c_n)}\right), \end{aligned}$$

since $(n/k_n)P(w_t > c_n) = 1$. Distribution continuity ensures $f(a) := (\partial/\partial a)P(w_t \le a)$ exists and is uniformly bounded by Assumption 2. Hence by the mean-value-theorem for some $|u^*| \le |u|$

$$k_n^{1/2}\left(\mathcal{I}_n(u/k_n^{1/2}) - E\left[\mathcal{I}_n(u/k_n^{1/2})\right]\right) = k_n^{1/2}\frac{f\left(c_n e^{u^*/k_n^{1/2}}\right)c_n e^{u^*/k_n^{1/2}}u/k_n^{1/2}}{P(w_t > c_n)} = \frac{f\left(c_n e^{u^*/k_n^{1/2}}\right)c_n e^{u^*/k_n^{1/2}}}{P(w_t > c_n)}u.$$

By power law tail decay it follows $P(w_t > c_n e^{u/k_n^{1/2}}) = P(w_t > c_n)e^{-\kappa u/k_n^{1/2}}(1 + o(1))$ and coupled with density boundedness $f(c_n e^{u/k_n^{1/2}})c_n/P(w_t > c_n e^{u/k_n^{1/2}}) \to \xi$ a positive finite constant (cf. Resnick, 1987). Therefore $k_n^{1/2}\ln(w_{(k_n)}/c_n) \le u$ if $\mathcal{I}_n(u/k_n^{1/2}) \le 1$ if $\xi^{-1}k_n^{1/2}(\mathcal{I}_n(u/k_n^{1/2}) - E[\mathcal{I}_n(u/k_n^{1/2})]) = u + o(1)$. Thus since $\xi^{-1}k_n^{1/2}\{\mathcal{I}_n(u/k_n^{1/2}) - E[\mathcal{I}_n(u/k_n^{1/2})]\} \xrightarrow{d} \mathcal{Z}$ a mean-zero normal law with finite variance $v_2^2 := \xi^{-2}v_1^2(0)$,

$$\lim_{n\to\infty} P\left(k_n^{1/2}\ln(w_{(k_n)}/c_n) \le u\right) = \lim_{n\to\infty} P\left(\xi^{-1}k_n^{1/2}\left(\mathcal{I}_n(u/k_n^{1/2}) - E\left[\mathcal{I}_n(u/k_n^{1/2})\right]\right) \le u + o(1)\right) = P(\mathcal{Z} \le u). \tag{16}$$

Therefore $k_n^{1/2}\ln(w_{(k_n)}/c_n) \xrightarrow{d} N(0, v_2^2)$, hence $w_{(k_n)}/c_n = 1 + O_p(k_n^{-1/2})$ by the mean-value-theorem.

Step 2 (uniform): Define $\mathcal{I}_n(u,\theta) := 1/k_n\sum_{t=1}^n I(w_t^{(a)}(\theta) > c_n(\theta)e^{u/k_n^{1/2}})$ and $\mathcal{Z}_n(u,\theta) := k_n(n/k_n)^{1/2}\mathcal{I}_n(u,\theta)$. Invoke uniform tail properties (12)–(14) and repeat the argument leading to eqn(16) to obtain for any $u \in \mathbb{R}$

$$P\left(k_n^{1/2}\ln\left(w_{(k_n)}^{(a)}(\theta)/c_n(\theta)\right) \le u\right) = P\left(\xi^{-1}n^{-1/2}(\mathcal{Z}_n(u,\theta) - E[\mathcal{Z}_n(u,\theta)]) \le u + o(1)\right).$$

The claim now follows from uniform indicator law Lemma B1 and the mapping theorem.

The proofs of Lemmas A1–A6 now follow.

Proof of lemma A1. **Claim (a):** By Minkowski's inequality

$$\left|\sum_{t=1}^n \epsilon_t y_{t-1}\left\{\hat{I}_{n,t}^{(\epsilon)}\hat{I}_{n,t-1}^{(y)} - I_{n,t}^{(\epsilon)}I_{n,t-1}^{(y)}\right\}\right| \le \left|\sum_{t=1}^n \epsilon_t\left(\hat{I}_{n,t}^{(\epsilon)} - I_{n,t}^{(\epsilon)}\right) \times y_{t-1}I_{n,t-1}^{(y)}\right| + \left|\sum_{t=1}^n \epsilon_t I_{n,t}^{(\epsilon)} \times y_{t-1}\left(\hat{I}_{n,t-1}^{(y)} - I_{n,t-1}^{(y)}\right)\right| + \left|\sum_{t=1}^n \epsilon_t\left(\hat{I}_{n,t}^{(\epsilon)} - I_{n,t}^{(\epsilon)}\right) \times y_{t-1}\left(\hat{I}_{n,t-1}^{(y)} - I_{n,t-1}^{(y)}\right)\right|.$$

We will show the first term is $o_p(n^{1/2}\Sigma_n^{1/2})$, the remaining terms being similar.

The indicator $I(u) := I(u \le 0)$ can be approximated by a *regular* sequence $\{\mathfrak{I}_n(u)\}_{n\ge1}$, cf. Lighthill (1958). Let $\{\mathcal{N}_n\}$ be a sequence of finite positive numbers, $\mathcal{N}_n \to \infty$, the rate to be chosen below. Define $\mathfrak{I}_n(u) := \int_{-\infty}^{\infty} I(\varpi)\mathrm{S}(\mathcal{N}_n(\varpi - u))\mathcal{N}_n e^{-\varpi^2/\mathcal{N}_n^2}\mathrm{d}\varpi$ where $\mathrm{S}(\xi) = e^{-1/(1-\xi^2)}/\int_{-1}^1 e^{-1/(1-w^2)}\mathrm{d}w$ if $|\xi| < 1$ and $\mathrm{S}(\xi)=0$ if $|\xi| \le 1$. The function $\mathrm{S}(\mathcal{N}_n(\varpi - u))$ blots out $I(\varpi)$ when ϖ is outside the open interval $(u - 1/\mathcal{N}_n, u + 1/\mathcal{N}_n)$. The function $\mathfrak{I}_n(u)$ is uniformly bounded in u, and continuous and differentiable. Also, $I(u)$ is differentiable except at 0 with derivative $\delta(u) = (\partial/\partial u)I(u) = 0\ \forall u \ne 0$, the Dirac delta function. Therefore $\delta(u)$ has a regular sequence $\mathfrak{D}_n(u) := (\mathcal{N}_n/\pi)^{1/2}\exp\{-\mathcal{N}_n u^2\}$ (see Lighthill, 1958, p. 22).

Write $c_n = c_n^{(\epsilon)}$ and $k_n = k_n^{(\epsilon)}$, define $\mathcal{E}_t(a) := |\epsilon_t| - a$ and notice by our notation $\hat{I}_{n,t}^{(\epsilon)} = I(\mathcal{E}_t(\epsilon_{(k_n)}))$ and $I_{n,t}^{(\epsilon)} = I(\mathcal{E}_t(c_n))$. By the mean value theorem since $\mathcal{N}_n \to \infty$ can be made as fast as we choose, it can be set to ensure for some c_n^*, $|c_n^* - c_n| \le |\epsilon_{(k_n)} - c_n|$,

$$\left|\sum_{t=1}^n \epsilon_t\left(\hat{I}_{n,t}^{(\epsilon)} - I_{n,t}^{(\epsilon)}\right) \times y_{t-1}I_{n,t-1}^{(y)}\right| = \left|\sum_{t=1}^n \left(\mathfrak{I}_n(\mathcal{E}_t(\epsilon_{(k_n)})) - \mathfrak{I}_n(\mathcal{E}_t(c_n))\right) \times \epsilon_t y_{t-1}I_{n,t-1}^{(y)}\right| + o_p(1) \le K\left|\sum_{t=1}^n \mathfrak{D}_n(\mathcal{E}_t(c_n^*)) \times \epsilon_t y_{t-1}I_{n,t-1}^{(y)}\right| \times \left|\epsilon_{(k_n)} - c_n\right| + o_p(1).$$

By Lemma B2 $\epsilon_{(k_n)} - c_n = c_n \times O_p(1/k_n^{1/2})$, hence

$$\left|\sum_{t=1}^n \epsilon_t\left(\hat{I}_{n,t}^{(\epsilon)} - I_{n,t}^{(\epsilon)}\right) \times y_{t-1}I_{n,t-1}^{(y)}\right| \le \left|\sum_{t=1}^n \mathfrak{D}_n(\mathcal{E}_t(c_n^*)) \times \epsilon_t y_{t-1}I_{n,t-1}^{(y)}\right| \times O_p\left(c_n/k_n^{1/2}\right) + o_p(1).$$

Since distribution continuity implies $|\epsilon_t| \neq c_n^*$ *a.s.* it follows $\mathfrak{D}_n(\mathcal{E}_t(c_n^*)) \xrightarrow{p} 0$ as fast as we choose. In particular we always can set $\mathcal{N}_n \to \infty$ sufficiently fast to ensure

$$\begin{aligned}
&\left|\sum_{t=1}^{n} \epsilon_t \mathfrak{D}_n\left(\mathcal{E}_t(c_n^*)\right) \times y_{t-1} I_{n,t-1}^{(y)}\right| \times O_p\left(c_n/k_n^{1/2}\right) \\
&\leq \frac{1}{n}\sum_{t=1}^{n}\left|\epsilon_t y_{t-1} I_{n,t-1}^{(y)}\right| \times O_p\left(\max_{1\leq t\leq n}\left\{\mathfrak{D}_n(\mathcal{E}_t(c_n^*))\right\} c_n n^{1/2}/k_n^{1/2}\right) \times n^{1/2} \\
&\leq \frac{1}{n}\sum_{t=1}^{n}\left|\epsilon_t y_{t-1} I_{n,t}^{(\epsilon)} I_{n,t-1}^{(y)}\right| \times O_p\left(n^{1/2}\right) + o_p(1).
\end{aligned}$$

Further, by stationarity, ergodicity and integrability $1/n\sum_{t=1}^{n}|\epsilon_t y_{t-1} I_{n,t}|/E|\epsilon_t y_{t-1} I_{n,t}| \xrightarrow{p} 1$, and by Lyapunov's inequality $E|\epsilon_t y_{t-1} I_{n,t}| \leq \Sigma_n^{1/2}$. This proves the claim.

Claim (b): Write

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=1}^{n}\left\{\hat{m}_{n,t}(\theta) - m_{n,t}(\theta)\right\}\right| \\
&\leq \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=1}^{n} m_{n,t}(\theta)\left\{1 - \hat{I}_{n,t}^{(\epsilon)}(\theta)\hat{I}_{n,t-1}^{(y)}\right\}\right| + \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=1}^{n} m_t(\theta)\left(1 - I_{n,t}^{(\epsilon)}(\theta) I_{n,t-1}^{(y)}\right)\left\{1 - \hat{I}_{n,t}^{(\epsilon)}(\theta)\hat{I}_{n,t-1}^{(y)}\right\}\right| \\
&+ \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=1}^{n} m_{n,t}(\theta)\left\{1 - I_{n,t}^{(\epsilon)}(\theta) I_{n,t-1}^{(y)}\right\}\right| + \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=1}^{n} m_t(\theta)\left(1 - I_{n,t}^{(\epsilon)}(\theta) I_{n,t-1}^{(y)}\right)^2\right|.
\end{aligned}$$

In view of $\sup_{\theta\in\Theta} I_{n,t}^{(\epsilon)}(\theta) I_{n,t-1}^{(y)} \to^{a.s.} 0$ and $\sup_{\theta\in\Theta} \hat{I}_{n,t}^{(\epsilon)}(\theta)\hat{I}_{n,t-1}^{(y)} \xrightarrow{p} 0$ it follows by dominated convergence with probability approaching one for some large $K > 0$

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=1}^{n} m_t(\theta)\left(1 - I_{n,t}^{(\epsilon)}(\theta) I_{n,t-1}^{(y)}\right)\left\{1 - \hat{I}_{n,t}^{(\epsilon)}(\theta)\hat{I}_{n,t-1}^{(y)}\right\}\right| \leq K\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=1}^{n} m_{n,t}(\theta)\left\{1 - \hat{I}_{n,t}^{(\epsilon)}(\theta)\hat{I}_{n,t-1}^{(y)}\right\}\right| \\
&\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=1}^{n} m_t(\theta)\left(1 - I_{n,t}^{(\epsilon)}(\theta) I_{n,t-1}^{(y)}\right)^2\right| \leq K\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=1}^{n} m_{n,t}(\theta)\left\{1 - I_{n,t}^{(\epsilon)}(\theta) I_{n,t-1}^{(y)}\right\}\right|.
\end{aligned}$$

Therefore, by ULLN Lemma A3b and dominated convergence each term is $o_p\left(\sup_{\theta\in\Theta} E|m_{n,t}(\theta)|\right)$.

Claims (c) and (d): By exploiting the regular sequence $\{\mathfrak{I}_n(u)\}_{n\geq 1}$ the proofs are identical to (a).

Proof Of Lemma A2. *If $E[\epsilon_t^2] < \infty$ then Σ_n=O(1)=o(n). If $E[\epsilon_t^2] = \infty$ consider κ=2 and note by independence and two applications of trimmed moment properties (20) $\Sigma_n = E[\epsilon_t^2 I_{n,t}^{(\epsilon)}] \times E[y_{t-1}^2 I_{n,t-1}^{(y)}] = K(\ln(n))^2 = o(n)$. If $\kappa \in$ [1,2) then by threshold construction (13) and (14) $\Sigma_n = K(n/k_n^{(\epsilon)})^{2/\kappa-1}(n/k_n^{(y)})^{2/\kappa-1}$ and by Assumption 3.a $k_n^{(\epsilon)}k_n^{(y)}/n \to \infty$, hence $\Sigma_n = K(n^2/k_n^{(\epsilon)}k_n^{(y)})^{2/\kappa-1} = o(n^{2/\kappa-1}) = o(n)$. Finally, if $\kappa \in$ (0,1) use the Assumption 3.b implication $n^{2-\kappa/(2-\kappa)} = o(k_n^{(\epsilon)}k_n^{(y)})$ to deduce $\Sigma_n = (n^2/(k_n^{(\epsilon)}k_n^{(y)}))^{2/\kappa-1} = o(n)$.*

Proof Of Lemma A3. **Claim (a):** *$1/n\sum_{t=1}^{n} m_{n,t} = o_p(1)$ follows from E[$m_{n,t}$]=0 by distribution symmetry, the Lemma A2a variance bound Σ_n=o(n) and Chebyshev's inequality.*

Claims (b): Define $h_{n,t}(\theta) := (m_{i,n,t}(\theta) - E[m_{i,n,t}(\theta)])/\sup_{\theta\in\Theta} E\|m_{n,t}(\theta)\|$ for any $i \in \{1,2,3\}$. Observe $h_{n,t}(\theta)$ has a zero mean and is integrable uniformly on Θ. In view of stationarity and ergodicity therefore $1/n\sum_{t=1}^{n} h_{n,t}(\theta) \xrightarrow{p} 0$. Further, by uniform L_1-boundedness $h_{n,t}(\theta)$ it belongs to a separable Banach space, hence the L_1-bracketing numbers satisfy $N_{[]}(\varepsilon, \Theta, \|\cdot\|_1) < \infty$ (Dudley, 1999: Proposition 7.1.7). Now combine the pointwise law and $N_{[]}(\varepsilon, \Theta, \|\cdot\|_1) < \infty$ to deduce $\sup_{\theta\in\Theta}|1/n\sum_{t=1}^{n} h_{n,t}(\theta)| = o_p(1)$ by Theorem 7.1.5 of Dudley (1999). This proves (b). *calQED.*

Proof of Lemma A4. **Claim (a).** *Recall we focus on the AR(1) case without an intercept. Choose any $\theta, \tilde{\theta} \in \Theta$, and define $\tilde{G}_n(\theta) := -1/n\sum_{t=1}^{n} y_{t-1}^2 I_{n,t}(\theta)$. By linearity $m_{n,t}(\theta) = \{m_t(\tilde{\theta}) - y_{t-1}^2(\theta - \tilde{\theta})\} \times I_{n,t}(\theta)$, hence*

$$\begin{aligned}
m_n(\theta) - m_n(\tilde{\theta}) &= \tilde{G}_n(\theta) \times (\theta - \tilde{\theta}) + \frac{1}{n}\sum_{t=1}^{n} m_t(\theta) \times \left\{I_{n,t}^{(\epsilon)}(\theta) - I_{n,t}^{(\epsilon)}(\tilde{\theta})\right\} I_{n,t-1}^{(y)} \\
&= \tilde{G}_n(\theta) \times (\theta - \tilde{\theta}) + \mathfrak{R}_n^*(\theta, \tilde{\theta}),
\end{aligned} \tag{17}$$

say. We must show $\mathfrak{R}_n^*(\theta, \tilde{\theta})$ is $o_p(\|G_n\| \times \|\theta - \tilde{\theta}\|)$.

We exploit the regular sequences $\{\mathfrak{I}_n(u), \mathfrak{D}_n(u)\}_{n\geq 1}$ defined in the proof of Lemma A1 (see that proof for definitions). Write $c_n(\theta) = c_n^{(\epsilon)}(\theta)$ and define $\mathcal{E}_{n,t}(\theta) := |\epsilon_t(\theta)| - c_n(\theta)$. Since $\mathcal{N}_n \to \infty$ can be made as fast as we choose, by the mean value theorem it can be set to ensure for some θ^*, $\|\theta^* - \theta\| \leq \|\theta - \tilde{\theta}\|$

$$\begin{aligned}\mathfrak{M}_n^*(\theta,\tilde{\theta}) &= \frac{1}{n}\sum_{t=1}^n m_t(\theta)\mathfrak{D}_n(\mathcal{E}_{n,t}(\theta^*))\Big(|\epsilon_t(\theta)| - |\epsilon_t(\tilde{\theta})|\Big)I_{n,t-1}^{(y)} \\ &\quad - \frac{1}{n}\sum_{t=1}^n m_t(\theta)\mathfrak{D}_n(\mathcal{E}_{n,t}(\theta^*))I_{n,t-1}^{(y)} \times \Big(c_n(\theta) - c_n(\tilde{\theta})\Big) + o_p(1) \\ &= \mathcal{A}_n(\theta,\theta^*,\tilde{\theta}) + \mathcal{B}_n(\theta,\theta^*,\tilde{\theta}) + o_p(1).\end{aligned}$$

Consider $\mathcal{B}_n(\theta,\theta^*,\tilde{\theta})$. By distribution continuity $|\epsilon_t(\theta)| \neq c_n(\theta)$ *a.s.*, by linearity $\epsilon_t(\theta) = \epsilon_t(\theta^*) - (\theta - \theta^*)y_{t-1}$, and by distribution continuity $c_n(\theta)$ has a derivative $d_n(\theta) := (\partial/\partial\theta)c_n(\theta)$ that is finite for each n. Further $\|\theta^* - \theta\| \leq \|\theta - \tilde{\theta}\|$. Since $\mathcal{N}_n \to \infty$ is arbitrary it can be set to ensure $\sup_{\theta\in\Theta}\|d_n(\theta)\|/\mathcal{N}_n^{1/2} \to 0$ hence by the definition of $\mathfrak{D}_n(\mathcal{E}_{n,t}(\theta^*))$

$$\begin{aligned}\big|\mathcal{B}_n(\theta,\theta^*,\tilde{\theta})\big| \leq & o_p\left(\left|\frac{1}{n}\sum_{t=1}^n \frac{\mathcal{N}_n|\epsilon_t(\theta^*)| \times E\Big|y_{t-1}I_{n,t-1}^{(y)}\Big|}{\exp\Big\{\mathcal{N}_n(|\epsilon_t(\theta^*)| - c_n(\theta^*))^2\Big\}} \frac{\Big|y_{t-1}I_{n,t-1}^{(y)}\Big|}{E\Big|y_{t-1}I_{n,t-1}^{(y)}\Big|}\right| \times \|\theta - \tilde{\theta}\|\right) \\ & + o_p\left(\frac{1}{n}\sum_{t=1}^n \frac{\mathcal{N}_n}{\exp\Big\{\mathcal{N}_n(|\epsilon_t(\theta^*)| - c_n(\theta^*))^2\Big\}} y_{t-1}^2 I_{n,t-1}^{(y)} \times \|\theta - \tilde{\theta}\|\right) \\ = & \mathcal{C}_{1,n}(\theta,\theta^*,\tilde{\theta}) + \mathcal{C}_{2,n}(\theta,\theta^*,\tilde{\theta}).\end{aligned}$$

By stationarity, ergodicity and integrability $1/n\sum_{t=1}^n |y_t I_{n,t}^{(y)}|/E|y_t I_{n,t-1}^{(y)}| \xrightarrow{p} 1$ and $1/n\sum_{t=1}^n y_t^2 I_{n,t}^{(y)}/E[y_t^2 I_{n,t}^{(y)}] \xrightarrow{p} 1$; $E[y_t^2 I_{n,t}^{(y)}] = -G_n$ by construction; by Lyapunov's inequality $E|y_t I_{n,t}^{(y)}| = \|G_n\|^{1/2}$; and by distribution non-degeneracy and trimming negligibility $\liminf_{n\to\infty}\|G_n\| > 0$ hence $\|G_n\|^{1/2} = O(\|G_n\|)$. Further, by distribution continuity $\epsilon_t(\theta^*) \neq c_n(\theta^*)$ *a.s.*, and $E(\sup_{\theta\in\Theta}|\epsilon_t(\theta)|^\iota) < \infty$ for some $\iota > 0$ follows from linearity and $E|\epsilon_t|^\iota < \infty$. Therefore since $\mathcal{N}_n \to \infty$ is arbitrary it follows by Markov's inequality and dominated convergence $\mathcal{N}_n|\epsilon_t(\theta^*)|\exp\{-\mathcal{N}_n(|\epsilon_t(\theta^*)| - c_n(\theta^*))^2\} \xrightarrow{p} 0$ and $\mathcal{N}_n\exp\{-\mathcal{N}_n(|\epsilon_t(\theta^*)| - c_n(\theta^*))^2\} \xrightarrow{p} 0$ as fast as we choose. Therefore for $\mathcal{N}_n \to \infty$ sufficiently fast both $\mathcal{C}_{i,n}(\theta,\theta^*,\tilde{\theta})$ are $o_p(\|G_n\| \times \|\theta - \tilde{\theta}\|)$.

Similarly, use $|\epsilon_t(\theta)| - |\epsilon_t(\tilde{\theta})| \leq \|\theta - \tilde{\theta}\| \times |y_{t-1}|$ to deduce for $\mathcal{N}_n \to \infty$ sufficiently fast

$$\begin{aligned}\big|\mathcal{A}_n(\theta,\theta^*,\tilde{\theta})\big| &\leq \frac{1}{n}\sum_{t=1}^n \frac{|\epsilon_t(\theta)|\mathcal{N}_n^{1/2}}{\exp\Big\{\mathcal{N}_n(|\epsilon_t(\theta^*)| - c_n(\theta^*))^2\Big\}} y_{t-1}^2 I_{n,t-1}^{(y)} \times \|\theta - \tilde{\theta}\| \\ &= o_p\Big(\|G_n\| \times \|\theta - \tilde{\theta}\|\Big).\end{aligned}$$

Claim (b). In view of order statistic consistency Lemma B2, the proof is identical to (a) above, and to Lemma A1a.

Claim (c). Since $\epsilon_t(\hat{\theta}_n) = \epsilon_t - (\hat{\theta}_n - \theta^0)y_{t-1}$ and $\hat{\theta}_n \xrightarrow{p} \theta^0$ by Theorem 2.1, the above argument implies $1/n\sum_{t=1}^n \epsilon_t^2\{\hat{I}_{n,t}(\hat{\theta}_n) - \hat{I}_{n,t}\} = o_p(1)$, hence

$$\frac{1}{n}\sum_{t=1}^n \epsilon_t^2(\hat{\theta}_n)\hat{I}_{n,t}(\hat{\theta}_n) = \frac{1}{n}\sum_{t=1}^n \epsilon_t^2\hat{I}_{n,t} - 2\Big(\hat{\theta}_n - \theta^0\Big)'\frac{1}{n}\sum_{t=1}^n \hat{m}_{n,t} + \Big(\hat{\theta}_n - \theta^0\Big)'\frac{1}{n}\sum_{t=1}^n y_{t-1}^2\hat{I}_{n,t}(\hat{\theta}_n)\Big(\hat{\theta}_n - \theta^0\Big) + o_p(1).$$

By Lemma A1c $1/n\sum_{t=1}^n \epsilon_t^2\hat{I}_{n,t} = 1/n\sum_{t=1}^n \epsilon_t^2 I_{n,t} + o_p(1)$ and by integrability and ergodicity $1/n\sum_{t=1}^n \epsilon_t^2 I_{n,t}/E[\epsilon_t^2 I_{n,t}] \xrightarrow{p} 1$.

We now show the second and third terms are $o_p(1)$. By Lemmas A1a $1/n\sum_{t=1}^n \hat{m}_{n,t} = 1/n\sum_{t=1}^n m_{n,t} + o_p(\|\Sigma_n/n\|^{1/2})$ where $\|\Sigma_n/n\|^{1/2} = o(1)$ by Lemma A2 and $1/n\sum_{t=1}^n m_{n,t} = o_p(1)$ by Lemma A3a, hence the second term is $o_p(1)$. Next, invoke Lemma A6b to obtain $1/n\sum_{t=1}^n y_{t-1}^2\hat{I}_{n,t}(\hat{\theta}_n) = -G_n(1 + o(1))$, note $\hat{\theta}_n - \theta^0 = O_p(\mathcal{V}_n^{-1/2})$ by Theorem 2, and by construction and Lemma A2 $\mathcal{V}_n^{-1/2}G_n\mathcal{V}_n^{-1/2} = \mathcal{V}_n^{-1/2}\Sigma_n^{1/2}/n^{1/2} \to 0$. Hence the third term is $o_p(1)$.

Claim (d). The proof is the same as (c).

Proof of Lemma A5. *Define $z_{n,t} := \Sigma_n^{-1/2}m_{n,t}$. By symmetry and error independence $E[z_{n,t}]=0$ and $E(\sum_{t=1}^n z_{n,t})^2 = n$. Since y_t is stationary and geometrically β-mixing it suffices to verify (2.1) and (2.2) in Peligrad (1996, Theorem 2.1), which are $\sup_{n\geq 1}\{1/n\sum_{t=1}^n E[z_{n,t}^2]\} < \infty$ and $1/n\sum_{t=1}^n E[z_{n,t}^2 I(|z_{n,t}| > \varepsilon n^{1/2})] \to 0 \forall \varepsilon > 0$. By construction $E[z_{n,t}^2] = 1$, hence (2.1).*

The Lindeberg trivially condition (2.2) holds if $\kappa > 2$ since by distribution continuity $E|\epsilon_t|^{2+\iota} < \infty$ for some $\iota > 0$ hence $\limsup_{n\to\infty}E|z_{n,t}|^{2+\iota} < \infty$.

Now suppose $\kappa \leq 2$, and recall $m_t = \epsilon_t y_{t-1}$ has a power law tail with the same index κ, and by trimming $|\epsilon_t y_{t-1}I_{n,t}| \leq c_n^{(\epsilon)}c_n^{(y)}$. Therefore

$$E\left[z_{n,t}^2 I\left(z_{n,t}^2 > \varepsilon^2 n\right)\right] = \frac{1}{\Sigma_n} E\left[\epsilon_t^2 y_{t-1}^2 I_{n,t} I\left(\epsilon_t^2 y_{t-1}^2 I_{n,t} > \varepsilon^2 \Sigma_n n\right)\right] \leq K \frac{1}{\Sigma_n} \int_{\varepsilon^2 \Sigma_n n}^{(c_n^{(\epsilon)} c_n^{(y)})^2} u^{-\kappa/2} du$$

If κ=2 then $\Sigma_n \sim K(\ln(n))^2$ by eqn(14), hence the integral bounds satisfy $(c_n^{(\epsilon)} c_n^{(y)})^2 < \varepsilon^2 \Sigma_n n$ as $n \to \infty$. This follows since by threshold construction (13) we have $(c_n^{(\epsilon)} c_n^{(y)})^2/n = K(n/k_n^{(\epsilon)})^{2/\kappa}(n/k_n^{(y)})^{2/\kappa}/n = Kn/(k_n^{(\epsilon)} k_n^{(y)}) \to 0$ by Assumption 3b. But this implies for some $N \in \mathbb{N}$ and all $n \leq N$ that $\int_{\varepsilon^2 \Sigma_n n}^{(c_n^{(\epsilon)} c_n^{(y)})^2} u^{-\kappa/2} du = 0$.

Finally, if $\kappa < 2$ then $\Sigma_n \sim K(c_n^{(\epsilon)} c_n^{(y)})^2 (k_n^{(\epsilon)}/n)(k_n^{(y)}/n)$ by eqn(14). Hence again the integral bounds $(c_n^{(\epsilon)} c_n^{(y)})^2 < \varepsilon^2 \Sigma_n n$ as $n \to \infty$ since $(c_n^{(\epsilon)} c_n^{(y)})^2/(\Sigma_n n) = K/(k_n^{(\epsilon)} k_n^{(y)}/n) = Kn/(k_n^{(\epsilon)} k_n^{(y)}) \to 0$ by Assumption 3b.

Proof of Lemma A6. **Claim** (*a*): *Recall* $G_n = -E[y_{t-1}^2 I_{n,t-1}^{(y)}]$. *By expansion Lemma A.4.a and Jacobian consistency* (*b*), *we have*

$$\frac{1}{n}\sum_{t=1}^{n}\{m_{n,t}(\theta) - m_{n,t}\} = -\frac{1}{n}\sum_{t=1}^{n} y_{t-1}^2 I_{n,t} \times (\theta - \theta^0) \times (1 + o_p(1)) = G_n \times (\theta - \theta^0) \times (1 + o_p(1)).$$

Invoke dominated convergence and error independence to deduce $E[y_{t-1}^2 I_{n,t}] = E[y_{t-1}^2 I_{n,t-1}^{(y)}] \times (1 + o(1))$ and

$$\frac{E[m_{n,t}(\theta)] - E[m_{n,t}]}{|\theta - \theta^0|} = G_n \times (1 + o(1)).$$

Identically, by the definition of a derivative

$$\frac{E[m_{n,t}(\theta)] - E[m_{n,t}]}{|\theta - \theta^0|} = \frac{\partial}{\partial \theta} E[m_{n,t}(\theta)] \times (1 + o(|\theta - \theta^0|)) + o(\|G_n\|).$$

Equate (18) and (19) and take $|\theta - \theta^0| \to 0$ to prove the claim.

Claim (b): By distribution smoothness there exists a point $\tilde{\theta} \in \Theta$ that satisfies $\sup_{\theta \in \Theta} E|m_{n,t}(\theta)| > E|m_{n,t}|$. Further, since $m_{n,t}(\theta) = m_{n,t}(\tilde{\theta}) - y_{t-1}^2 I_{n,t}(\theta)(\theta - \tilde{\theta}) + m_t\{I_{n,t}^{(\epsilon)}(\theta) - I_{n,t}^{(\epsilon)}(\tilde{\theta})\} I_{n,t-1}^{(y)}$, $I_{n,t}^{(\epsilon)}(\theta) \in \{0, 1\}$, and Θ is compact it follows

$$\sup_{\theta \in \Theta} E|m_{n,t}(\theta)| \leq E|m_{n,t}(\tilde{\theta})| + KE\left[y_{t-1}^2 I_{n,t-1}^{(y)}\right] + \sup_{\theta \in \Theta} E\left|m_t\{I_{n,t}^{(\epsilon)}(\theta) - I_{n,t}^{(\epsilon)}(\tilde{\theta})\} I_{n,t-1}^{(y)}\right|.$$

By construction $E[y_{t-1}^2 I_{n,t-1}^{(y)}] = |G_n|$, and by the proofs of Lemmas A1a and A4a the term $\sup_{\theta \in \Theta} E|m_t\{I_{n,t}^{(\epsilon)}(\theta) - I_{n,t}^{(\epsilon)}(\tilde{\theta})\} I_{n,t-1}^{(y)}|$ can be shown to be $o(|G_n| \times |\theta - \tilde{\theta}|)$ which is $o(|G_n|)$ in view of compactness of Θ. This proves $\sup_{\theta \in \Theta} E|m_{n,t}(\theta)| \leq E|m_{n,t}(\tilde{\theta})| + K|G_n| \times (1 + o(1))$. Since $\sup_{\theta \in \Theta} E|m_{n,t}(\theta)| > E|m_{n,t}|$ it therefore follows $\sup_{\theta \in \Theta} E|m_{n,t}(\theta)| \leq K|G_n| \times (1 + O(1)) \leq K|G_n|$ (recall K may be different in different places).

Claim (c): By Lemma A1d and A4d and consistency $\hat{\theta}_n \xrightarrow{p} \theta^0$ under Theorem 1 it follows $1/n \sum_{t=1}^n y_{t-1}^2 \hat{I}_{n,t}(\hat{\theta}_n) = 1/n \sum_{t=1}^n y_{t-1}^2 I_{n,t} + o_p(1)$. Further, by trimming negligibility $\liminf_{n \to \infty} E[y_{t-1}^2 I_{n,t}] > 0$, and by integrability and ergodicity $1/n \sum_{t=1}^n y_{t-1}^2 I_{n,t}/E[y_{t-1}^2 I_{n,t}] \xrightarrow{p} 1$. Moreover, $E[y_{t-1}^2 I_{n,t}] = E[y_{t-1}^2 I_{n,t-1}^{(y)}] \times (1 + o(1))$ by trimming negligibility. This proves $1/n \sum_{t=1}^n y_{t-1}^2 \hat{I}_{n,t}(\hat{\theta}_n)/E[y_{t-1}^2 I_{n,t-1}^{(y)}] \xrightarrow{p} 1$ which completes the proof.

Claim (d): The proof is the same as (c) since $1/n \sum_{t=1}^n y_{t-1}^2 \hat{I}_{n,t-1}^{(y)} = 1/n \sum_{t=1}^n y_{t-1}^2 I_{n,t-1}^{(y)} + o_p(1)$ can be easily shown by the line of proof of Lemmas A1d and A4d.

NOTES

1. Notice $E[z_t^2 I(|z_t| \leq c_n^{(z)})] = K + \int_a^{(c_n^{(z)})^2} P(|z_t| > u^{1/2}) du$ for finite $a_2 > 0$ and some $K > 0$ that depends on a. If κ=2 then $c_n^{(\epsilon)} = d^{1/2}(n/k_n^{(\epsilon)})^{1/2}$ and therefore $E[z_t^2 I(|z_t| \leq c_n^{(z)})] \sim K + d_z \int_a^{(c_n^{(z)})^2} u^{-1} du \sim 2 d_z \ln(c_n^{(z)}) \sim d_z \ln(n)$.
2. The simulation results for LTTS based on $k_n^{(\epsilon)} = [0.05n/\ln(n)]$ and $k_n^{(y)} \sim \max\{1, [0.1 \ln(n)]\}$ are qualitatively identical to the results reported here, and are available upon request.
3. Our data were taken from finance.yahoo.com, which may be slightly different from Ling's data. Ling reports 497 observations.

Acknowledgement

We thank a referee for constructive comments that lead to a much improved paper.

REFERENCES

Agulló, J., Croux, C. and Van Aelst S. (2008) The multivariate least-trimmed squares estimator. *Journal of Multivariate Analysis* **99**, 311–38.
An, H. Z. and Chen, Z. G. (1982) On convergence of lad estimates in autoregression with infinite variance. *Journal of Multivariate Analysis* **12**, 335–45.

Aue, A., Horváth, L. and Steinbach, J. (2006) Estimation in random coefficient autoregressive models. *Journal of Time Series Analysis* **27**, 61–76.
Baek, E. and Brock, W. (1992) A Nonparameteric Test for Independence of a Multivariate Time Series. *Statistica Sinica* **2**, 137–56.
Basrak, B., Davis, R. A. and Mikosch, T. (2002) Regular variation of GARCH processes. *Stochastic Processes and Their Applications* **99**, 95–115.
Bassett, G. W. (1991) Equivariant, monotonic, 50% breakdown estimators. *American Statistics* **45**, 135–7.
Bollerslev, T. (1986) Generalized conditional autoregressive heteroskedasticity. *Journal of Econometrics* **31**, 307–27.
Breidt, F. J. and Davis, R. A. (1998) Extremes of stochastic volatility models. *The Annals of Applied Probability* **8**, 664–75.
Brockwell, P. and Cline, D. B. H. (1985) Linear Prediction of ARMA Processes with Infinite Variance. *Stochastic Processes and Their Applications* **19**, 281–96.
Carrasco, M. and Chen, X. (2002) Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory* **18**, 17–39.
Chan, N. H., Li, D. and Peng, L. (2012) Toward a unified interval estimation of autoregressions. *Econometric Theory* **28**, 705–17.
Chen, L.-A., Welsh, A. H. and Chan, W. (2001) Estimators for the linear regression model based on winsorized observations. *Statistica Sinica* **11**, 147–72.
Čižek, P. (2008) General trimmed estimation: robust approach to nonlinear and limited dependent variable models. *Econometric Theory* **24**, 1500–29.
Cline, D. B. H. (1989) Consistency for least squares regression estimators with infinite variance data. *Journal of Statistical Planning and Inference* **23**, 163–79.
Cline, D. B. H. (2007) Regular variation of order 1 nonlinear AR-ARCH models. *Stochastic Processes and Their Applications* **117**, 840–61.
Csörgo, S., Horváth, L. and Mason, D. (1986) What portion of the sample makes a partial sum asymptotically stable or normal?. *16. Probability Theory and Related Fields* **72**, 1–16.
Davis, R. A. (1996) Gauss-Newton and M-estimation for ARMA processes with infinite variance. *Stochastic Processes and Their Applications* **63**, 75–95.
Davis, R. A. (2010) Heavy tails in financial time series, In *Encyclopedia Quantitative Finance*. (ed. R. Cont), New York: Wiley.
Davis, R. A. and Resnick, S. I. (1986) Limit theory for sample for the sample covariance and correlation functions of moving averages. *19. Annals of Statistics* **14**, 533–58.
Davis, R. A. and Resnick, S. I. (1996) Limit theory for bilinear processes with heavy-tailed noise. *The Annals of Applied Probability* **6**, 1191–210.
Davis, R. A. and Wu, W. (1997) M-Estimation for linear regression with infinite variance. *Probabilities in Mathematics and Statistics* **17**, 1–20.
Davis, R. A., Knight, K. and Liu, J. (1992) M-Estimation for autoregressions with infinite variance. *Stochastic Processes and Their Applications* **40**, 145–80.
Doukhan, P., Massart, P. and Rio, E. (1995) Invariance principles for absolutely regular empirical processes. *Ann. de' I. H. P. Sec. B* **31**, 393–427.
Dudley, R. M. (1999) *Uniform Central Limit Theorems*. Cambridge: Cambridge University Press.
Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance*. Frankfurt: Springer-Verlag.
Feigin, P. and Resnick, S. I. (1994) Limit distributions for linear programming time series estimators. *Stochastic Processes and Their Applications* **51**, 135–66.
Feigin, P. and Resnick, S. I. (1999) Pitfalls of fitting autoregressive models for heavy tailed times series. *Extremes* **1**, 391–422.
Finkenstädt, B. and Rootzén, H. (2001) *Extreme Values in Finance, Telecommunications, and the Environment*. New York: Chapman and Hall.
Francq, C. and Zakoïan, J-M. (2004) Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* **10**, 605–37.
Giné, E. and Zinn, J. (1984) Some limit theorems for empirical processes. *The Annals of Probability* **12**, 929–89.
Gross, S. and Steiger, W. L. (1979) Least absolute deviation estimates in autoregression with infinite variance. *Journal of Applied Probability* **16**, 104–16.
Hahn, M. G., Weiner, D. C. and Mason, D. M. (1991) *Sums, Trimmed Sums and Extremes*. Berlin: Birkhäuser.
Hall, P. and Yao, Q. (2003) Inference in ARCH and GARCH models with Heavy-Tailed Errors. *Econometrica* **71**, 285–317.
Hall, P., Peng, L. and Yao, Q. (2002) Prediction and nonparametric estimation for time series with heavy tails. *Journal of Time Series Analysis* **23**, 251–75.
Hannan, E. J. and Kanter, M. (1977) Autoregressive processes with infinite variance. *Journal of Applied Probability* **14**, 411–15.
Hill, B. M. (1975), A simple general approach to inference about the tail of a distribution. *The Annals of Statistics* **3**, 1163–74.
Hill, J. B. (2009) On functional central limit theorems for dependent, heterogeneous arrays with applications to tail index and tail dependence estimation. *Journal of Statistical Planning and Inference* **139**, 2091–110.
Hill, J. B. (2010) On tail index estimation for dependent, heterogeneous data. *Econometric Theory* **26**, 1398–436.
Hill, J. B. (2011a) Extremal memory of stochastic volatility with an application to tail shape inference. *Journal of Statistical Planning and Inference* **141**, 663–76.
Hill, J. B. (2011b) Tail and non-tail memory with applications to extreme value and robust statistics. *Econometric Theory* **27**, 844–84.
Hill, J. B. (2012) Heavy-tail and plug-in robust consistent conditional moment tests of functional form. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr*. (eds. X. Chen and N. Swanson), New York: Springer, pp. 241–74.
Hill, J. B. and Aguilar, M. (2012) Moment condition tests for heavy tailed time series. *Journal of Econometrics* (in press).
Huber, P. J. (1977) *Robust Statistical Procedures*. Society for Industrial and Applied Mathematics: Philadelphia.
Kesten, H. (1973) Random difference equations and renewal theory for products of random matrices. *Acta Mathematica* **131**, 207–48.
Knight, K. (1987) Rate of convergence of centered estimates of autoregressive parameters for infinite variance regressions.. *Journal of Time Series Analysis* **8**, 51–60.
Lanne, M. and Lütkepohl, H. (2010) Structural vector autoregressions with nonnormal residuals. *Journal of Business & Economic Statistics* **28**, 159–68.
Lanne, M., Luoto, J. and Saikkonen, P. (2012) Optimal forecasting of noncausal autoregressive time series. *International Journal of Forecasting* **28**, 623–31.
Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983) *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer-Verlag.
Lighthill, M. J. (1958) *Introduction to Fourier Analysis and Generalised Functions*. Cambridge: Cambridge University Press.
Ling, S. (2005) Self-weighted LAD estimation for infinite variance autoregressive models. *Journal of the Royal Statistical Society Series B* **67**, 381–93.
Ling, S. (2007) Self-weighted and local quasi-maximum likelihood estimators for ARMA-GARCH/IGARCH models. *Journal of Econometrics* **150**, 849–73.
Liu, J-C. (2006) On the tail behaviors of a family of GARCH processes. *Econometric Theory* **22**, 852–62.
Nelson, D. (1990) Stationarity and Persistence in the GARCH(1,1) Model. *Econometric Theory* **6**, 318–34.
Neykov, N. M. and Neytchev, P. N. (1990) A Robust Alternative of the Maximum Likelihood Estimator. *Short. Comm. Compstat, Dubrovnik 1990*, 99–100.
Owen, A. (2001) *Empirical Likelihood*. Chapman & Hall: New York.
Pakes, A. and Pollard, D. (1989) Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**, 1027–57.
Pan, J., Wang, H. and Yao, Q. (2007) Weighted least absolute deviations estimation for ARMA models with infinite variance. *Econometric Theory* **23**, 852–79.
Peligrad, M. (1996) On the asymptotic normality of sequences of weak dependent random variables. *Journal of Theoretical Probability* **9**, 703–15.
Peters, G. W., Kannan, B., Lasscock, B., Mellon, C. and Godsill, S. (2011) Bayesian cointegrated vector autoregression models incorporating α-stable noise for inter-day price movements via approximate bayesian computation. *Bayesian Analysis* **6**, 755–92.
Pham, T. D. and Tran, L. T. (1985) Some mixing properties of time series models. *Stochastic Processes and Their Applications* **19**, 297–303.
Pollard, D. (1984) *Convergence of Stochastic Processes*. New York: Springer.
Powell, J.L. (1986) Symmetrically trimmed least squares estimation for tobit models. *Econometrica* **54**, 1435–60.
Resnick, S. I. (1987) *Extreme Values, Regular Variation and Point Processes*. New York.: Springer-Verlag.
Resnick, S. I. (1997) Heavy tail modeling and teletraffice data (with discussion). *Annals of Statistics* **25**, 1805–69.

Resnick, S. I. and Stărică, C. (1997) Asymptotic behavior of hill's estimator for autoregressive data. *Comm. Stat. Stoch. Models.* **13**, 703–21.
Robinson, P. M. (1991) Consistent nonparametric entropy-based testing. *Review of Economic Studies* **58**, 437–453.
Roitershtein, A. (2007) One-dimensional linear recursions with markov dependent coefficients. *Journal of American Statistical Association* **17**, 572–608.
Rousseeuw, P.J. (1984) Least median of squares regression. *Journal of American Statistical Association* **79**, 871–80.
Ruppert, D. and Carroll, J. (1980) Trimmed least squares estimation in the linear model. *67. Journal of American Statistical Association* **75**, 828–838.
Sims, C. A. (1980) Macroeconomics and reality. *Econometrica* **48**, 1–48.
Tableman, M. (1994) The asymptotics of the least trimmed absolute deviations (LTAD) estimator. *Statistics & Probability Letters* **19**, 329–37.
Tsay, R. (2002) *The Analysis of Financial Time Series.* New York: Wiley.
Zhu, K. and Ling, S. (2011) Global self-weighted and local quasi-maximum exponential likelihood estimators for ARMA-GARCH/IGARCH models. *Annals of Statistics* **39**, 2131–63.
Zhu, K. and Ling, S. (2012) The global LAD estimators for finite/infinite variance ARMA(p,q) models. *Econometric Theory* **28**, 1065–86.