Least Tail-Trimmed Squares for Infinite Variance Autoregressions

Jonathan B. Hill*
Dept. of Economics, University of North Carolina

May 15, 2012

Abstract

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 regressor. 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 and 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.

1. INTRODUCTION We develop the Least Tail-Trimmed Squares estimator for a stationary autoregression that may be very heavy tailed. The model is

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

where $\theta = [\xi, \phi']' \in \mathbb{R}^{p+1}$, with regressors $x_t = [1, y_{t-1}, ..., y_{t-p}]'$, and the sample is $\{y_t\}_{t=1}^n$ with sample size $n \geq 1$. We assume there exists a point θ^0 such that the error $\epsilon_t = \epsilon_t(\theta^0)$ is independent and identically distributed (iid), 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:

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

*Dept. of Economics, University of North Carolina-Chapel Hill, www.unc.edu/∼jbhill, jbhill@email.unc.edu.
Key words and phrases: least squares; tail trimming; heavy tails; robust inference.
AMS subject classifications. Primary 62F35; secondary 62F07.

We are especially interested in the infinite variance case where the tail index $\kappa \leq 2$. We assume the errors are independent with a symmetric distribution for simplicity in order to focus on the pure idea of tail trimming. 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) and Davis 2010) for compendium discussions.

If the AR errors have an infinite second moment $E[\epsilon_t^2] = \infty$ then M-estimators like Least Squares and Least Absolute Deviations 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), Knight (1987), Davis and Resnick (1986), Cline (1989), Davis et al (1992), 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 Least Squares [LS], as well as Least Absolute Deviations [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) and Davis and Resnick (1986).

Robust M-estimators like Least Trimmed Squares (Rousseeuw 1984, Čižek 2008), Least Absolute Trimmed Deviations (Basset 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 Rousseeuw (1984), Powell (1986), Ling (2007), Agulló et al (2008), and Zhu and Ling (2011) to name a few.

In the case of Least Trimmed Squares on (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 $\sum_{t=1}^n \epsilon_t^2(\theta) I(\epsilon_t^2(\theta) \leq \epsilon_{([\lambda n])}^2(\theta))$ 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 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, Basset 1991, Chen, Welsh and Chan 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 $i = 1, ..., p$. Ling (2005), for example, presents the Least Weighted Absolute Deviations estimator for (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 (i.e. those with a very larger error ϵ_t), hence it is sensitive to error extremes in small samples. See the simulation study in Section 5.

In this paper 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 Least Tail-Trimmed Squares [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 in lieu of trimming. See Section 2.

Tail-trimming to date is used primarily for location estimation (e.g. Csörgo et al 1986, Hahn et al 1991) including average treatment effects (Crump et al 2009) and moment condition tests (Hill 2012, Hill and Aguilar 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 squared errors $\epsilon_t^2(\theta)$ 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 positive sequence $\{g_n\}$, $g_n \to \infty$, that depends on the number of trimmed least squares equations. 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 percent of the sample λn, and trimming by y_{t-i} should be minimized as long as the amount increases with n. By comparison LS is asymptotically non-Gaussian and $(n/\ln(n))^{1/\kappa}$-convergent when $\kappa < 2$ (An and Chen 1982, Davis and Resnick 1986, Davis et al 1992). Hence when $\kappa \in [1, 2)$ LTTS obtains a rate of convergence $n^{1/\kappa}/g_n$ that is *faster* than LWAD and LS since $g_n \to \infty$ can be made arbitrarily slow by controlling the amount of trimming as n increases. If $\kappa = 2$ then $\hat{\theta}_n$ is $(n/\ln(n))^{1/\kappa}$-convergent, 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. By contrast, for technical reasons when $\kappa < 1$ 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 (2012), and Hill and Aguilar (2012).

We study autoregressions with iid symmetric errors and least squares for simplicity since linearity and symmetry allow us to focus ideas on the benefits of trimming. The methods and

theory developed here easily generalize to ARIMAX with nonlinear GARCH errors, to GARCH models with ARIMA-in-squares representation, and to other M-estimators.

We complete the paper with a simulation study in Section 5 and an empirical study of financial asset returns in Section 6. 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 estimator is constructed as follows. Represent two-tailed observations and order statistics for any random variable z_t by

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

The determination of the number of trimmed large errors and regressors 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

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

where $\Theta \subset \mathbb{R}^{p+1}$, 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. Thus $k_n^{(z)}/n \to 0$ implies we trim asymptotically a vanishing sample portion of observations.

Asymptotic theory for $\hat{\theta}_n$ requires the non-random sequences $\{c_n^{(\epsilon)}(\theta), c_n^{(y)}\}$ which the order

statistics $\{\epsilon^{(a)}_{(k_n^{(\epsilon)})}(\theta), y^{(a)}_{(k_n^{(y)})}\}$ estimate. Let $\{c_n^{(\epsilon)}(\theta), c_n^{(y)}\}$ be the exact quantiles defined by

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

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 $\{c_n^{(\epsilon)}(\theta), c_n^{(y)}\}$ for any $\{k_n^{(\epsilon)}, k_n^{(y)}\}$. See below for all assumptions. Clearly $\{\epsilon^{(a)}_{(k_n^{(\epsilon)})}(\theta), y^{(a)}_{(k_n^{(y)})}\}$ estimate $\{c_n^{(\epsilon)}(\theta), c_n^{(y)}\}$, and under regularity conditions presented below intermediate order statistics are uniformly consistent, e.g. $\sup_{\theta\in\Theta}|\epsilon^{(a)}_{(k_n^{(\epsilon)})}(\theta)/c_n^{(\epsilon)}(\theta) - 1| = O_p(1/(k_n^{(\epsilon)})^{1/2})$. 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):

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

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, $\psi_0 = 1$ 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 $\kappa = 1$ under Assumption 3.a, and otherwise holds under Assumption 3.b. Notice 3.a implies trimming cannot be too light, for example we cannot have both $k_n^{(z)} \sim \ln(n)$. Further, 3.b 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 $g_n^{(\epsilon)}/g_n^{(y)} \to 0$ ensures 3.a holds. See Section 3, below.

2.2 MAIN RESULTS

We state the main results here and relegate proofs to the appendix. 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 (see the appendices).

THEOREM 2.1 (*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 2.2 (*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 2.3 (*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 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 rates of convergence $\mathcal{V}_{i,i,n}^{1/2}$ 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 $j = 1, ..., p$ by dominated convergence and stationarity

$$E\left[\prod_{i=1}^{p} I\left(|y_{t-i}| \leq c_n^{(y)}\right)\right] = 1 + o(1) \tag{5}$$

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

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

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

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

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

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]

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

$$\text{if } \kappa = 2 \text{ then } E\left[z_t^2 I\left(|z_t| \leq c_n^{(z)}\right)\right] \sim \frac{d_z}{4}\ln(n) \text{ for any intermediate order } \{k_n^{(z)}\}.$$

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

[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 $a > 0$ and some $K > 0$. If $\kappa = 2$ then $c_n^{(\epsilon)} = d^{1/2}(n/k_n^{(\epsilon)})^{1/2}$ and therefore $E[z_t^2 I(|z_t| \leq c_n^{(z)})] \sim K + d_z \int_a^{(c_n^{(z)})^2} u^{-1} du \sim (d_z/2)\ln(c_n^{(z)}) \sim (d_z/4)\ln(n)$.

THEOREM 3.1 (*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 it allows us to optimize the rate to the highest possible amongst M-estimators for model (1) when $\kappa \in [1, 2)$. Keeping in mind that Assumption 3.a requires $k_n^{(\epsilon)} k_n^{(y)}/n \to \infty$, in general for any positive sequences $\{g_n^{(\epsilon)}, g_n^{(y)}\}$ where $g_n^{(1)} \geq 1$, $g_n^{(\cdot)} \to \infty$ as slow as we choose and $g_n^{(\epsilon)}/g_n^{(y)} \to 0$, simply put

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

to satisfy Assumption 3.a, 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 3.a, but not Assumption 3.b 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^{12}-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 3.a holds and if the error variance is infinite $\kappa \in [1, 2)$ then

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

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 3.b. However, Assumption 3.a and 3.b 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 Assumption 3.b 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

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

Notice $1/n \sum_{t=1}^{n} \epsilon_t^2(\hat{\theta}_n) \hat{I}_{n,t}(\hat{\theta}_n)$ uses the composite trimming indicator $\hat{I}_{n,t}(\hat{\theta}_n)$ rather than the error specific one $\hat{I}_{n,t}^{(\epsilon)}(\hat{\theta}_n)$. 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 $f : \mathbb{R} \to \mathbb{R}$ 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 4.1 (*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 2.2 and 4.1 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 $J \geq 1$, 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. SIMULATION STUDY We now compare the small sample properties of Least Squares [LS], Least Trimmed Squares [LTS], Least Tail-Trimmed Squares [LTTS], and Ling's

(2005) Least Weighted Absolute Deviations [LWAD].

We draw $n \in \{100, 400, 800\}$ random variables y_t from an AR(2) model $y_t = .2 + .8y_{t-1} - .3y_{t-2} + \epsilon_t$. The errors ϵ_t are iid symmetric Pareto distributed $P(\epsilon_t > \epsilon) = P(\epsilon_t < -\epsilon) = .5 \times (1+\epsilon)^{-\kappa}$ with index $\kappa \in \{.75, 1.5, 2.5\}$ 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 $k_n^{(\epsilon)} = [.05n/\ln(n)]$ and $k_n^{(y)} = \max\{1, [.01n/(\ln(n))^2]\}$ 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 3.b when $\kappa = .75$. Nevertheless, a pair like $k_n^{(\epsilon)} = [.05n/\ln(n)]$ and $k_n^{(y)} \sim \max\{1, [.1\ln(n)]\}$ 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 $\operatorname{argmin}_{\theta\in\Theta}\{\sum_{t=1}^n \epsilon_t^2(\theta) I(\epsilon_t^2(\theta) \leq \epsilon^2_{([\lambda n])}(\theta))\}$ with $\lambda = .05$ as in Čižek (2008). The LWAD estimator is $\operatorname{argmin}_{\theta\in\Theta}\{\sum_{t=1}^n w_t|\epsilon_t(\theta)|\}$ with Ling's (2005: eq. (2.3) chosen weight w_t based on Huber's (1977) influence function. Define $a_t := \sum_{i=1}^p |y_{t-i}| I(|y_{t-i}| \geq y^{(a)}_{([\lambda n])})$: 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 $\lambda = .05$. The parameter space for all estimators is $\Theta = [-1, 1]^3$.

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 $\theta_3^0 = -.3$ (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, as expected (Theorem 4.2).

[2] The simulation results for LTTS based on $k_n^{(\epsilon)} = [.05n/\ln(n)]$ and $k_n^{(y)} \sim \max\{1, [.1\ln(n)]\}$ are qualitatively identical to the results reported here, and are available upon request.

6. 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 June 3, 1996 to May 31, 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 2 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 $\hat{\kappa}_{r_n} = (1/r_n \sum_{i=1}^{r_n} \ln(y_{(i)}^a/y_{(r_n+1)}^a))^{-1}$ over fractiles $r_n \in \{5, 2, ..., 200\}$. Although we assume an AR model with iid 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 AR with linear or nonlinear GARCH shocks with geometric or hyperbolic memory decay. See Hill (2010, 2011) 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 $\gamma_n = n^{.225}$. 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: we never reject the one-sided hypothesis $\kappa \leq 2$ against $\kappa > 2$.

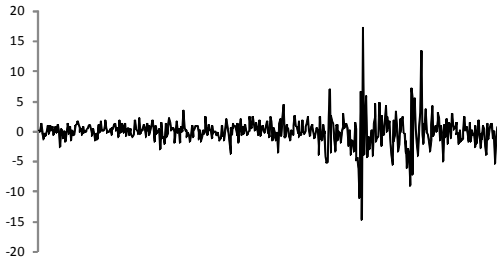

Figure 2: HSI Daily Log-Returns

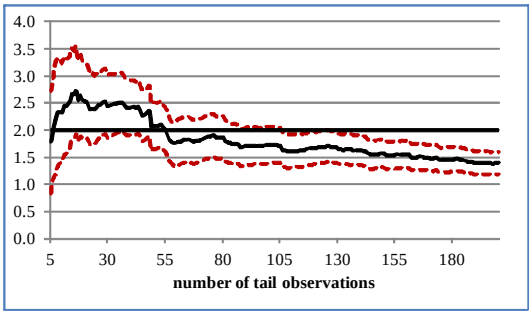

Figure 3: Hill-Plot and Robust 95% Bands

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 the LTTS fractiles are $k_n^{(\epsilon)} = [.05n/\ln(n)]$ and $k_n^{(y)} = \max\{1, [.01n/(\ln(n))^2]\}$ as in Section 5. The statistics are $\hat{W}_6 = 45.02$ (.000), $\hat{W}_8 = 43.99$ (.000), and $\hat{W}_{12} = 52.58$ (.000) with p-values in parentheses, hence white noise is rejected. We then estimate AR(p) models over $p = 1, ..., 12$ and test the

[3] Our data were taken from finance.yahoo.com, which may be slightly different from Ling's data. Ling reports 497 observations.

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, $p \geq 3$, 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 $\hat{W}_{12} = 9.88$ (.628), while AR(2) $\hat{W}_{12} = 27.62$ (.006) and AR(1) $\hat{W}_{12} = 18.45$ (.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)

$$\begin{array}{rcl} \hat{y}_t & = & -.10 + .25y_{t-1} - .02y_{t-2} + .11y_{t-3}. \\ & & \quad (.08)\ (.05) \qquad (.05) \qquad (.06) \end{array}$$

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

$$\begin{array}{rcl} \hat{y}_t & = & -.07 + .28y_{t-1} - .01y_{t-2} + .13y_{t-3} + .01y_{t-4} - .04y_{t-5} + .04y_{t-6} + .02y_{t-7}. \\ & & \quad (.08)\ (.05) \qquad (.05) \qquad (.05) \qquad (.05) \qquad (.07) \qquad (.06) \qquad (.09) \end{array}$$

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 (.000), .165 (.685) and 28.9 (.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

$$\begin{array}{l} \hat{y}_t = .07 + .07y_{t-1} + .00y_{t-2} + .11y_{t-3} + .03y_{t-4} - .08y_{t-5} + .02y_{t-6} - .09y_{t-7} \\ \quad (.06)\ (.04) \qquad (.04) \qquad (.04) \qquad (.04) \qquad (.04) \qquad (.04) \qquad (.04) \end{array}$$

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

6. CONCLUSION We present the Least Tail-Trimmed Squares 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. In fact, we can 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. The LTTS estimator is easy to compute, highly robust and efficient for very heavy tailed data. Although we do not present the details here, 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

LAD. These matters 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 $G_n(\theta)$:

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

As usual we drop θ^0. The proofs of consistency and asymptotic normality Theorems 2.1 and 2.2 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 A.1 (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 A.2 (variance bound). $\Sigma_n = o(n)$.

LEMMA A.3 (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 A.4 (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 A.5 (CLT). $n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^{n} m_{n,t} \xrightarrow{d} N(0, I_{p+1})$.

LEMMA A.6 (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.* $1/n\sum_{t=1}^n x_t x_t' \hat{I}_{n,t-1}^{(y)} = -G_n \times (1 + o_p(1))$.

See Appendix B for proofs of Lemmas A.1-A.6. We are now ready to prove Theorems 2.1 and 2.2.

PROOF OF THEOREM 2.1. The proof of consistency follows an argument in Pakes and Pollard (1989: Theorem 3.1, Corollary 3.2). Define $\mathcal{M}_n(\theta) := E[m_{n,t}(\theta)]$ 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 A.6.a $E[m_{n,t}(\theta)] = G_n \times (\theta - \theta^0) \times (1 + o(1))$, and by Lemma A.6.b 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 $n \geq N$ 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 $||G_n|| > 0\ \forall n \geq N$. 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 (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 $A_n(\hat{\theta}_n)$. 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 $A_n(\hat{\theta}_n) = 0$ *a.s.*

Next, $B_n(\hat{\theta}_n) \leq \sup_{\theta\in\Theta}\{\mathcal{B}_n(\theta)\}$. Combine $\sup_{\theta\in\Theta}\{||\hat{m}_n(\theta) - m_n(\theta)||/\mathfrak{e}_n\} = o_p(1)$ by Lemma A.1.b and $\sup_{\theta\in\Theta}\{||m_n(\theta) - \mathcal{M}_n(\theta)||/\mathfrak{e}_n\} = o_p(1)$ by ULLN Lemma A.3.b to deduce

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

PROOF OF THEOREM 2.2. By the proof of Theorem 2.1 $\hat{\theta}_n$ satisfies $1/n\sum_{t=1}^n \hat{m}_{n,t}(\hat{\theta}_n) = 0$ *a.s.* Apply expansion Lemma A.4.b to deduce

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

By Lemma A.6.c $\hat{G}_n(\hat{\theta}_n) = G_n(1 + o_p(1)) = -E[x_t x_t' I_{n,t}] \times (1 + o_p(1))$, 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

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

The claim now follows from approximation Lemma A.1.a and CLT Lemma A.5: $\mathcal{V}_n^{1/2}(\hat{\theta}_n - \theta^0) = -n^{-1/2}\Sigma_n^{-1/2}\sum_{t=1}^n m_{n,t} \times (1 + o_p(1)) \xrightarrow{d} N(0, I_{p+1})$. $\mathcal{QED}$.

PROOF OF THEOREM 4.1. In view of Jacobian consistency Lemma A.6.d 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 A.1.c and A.4.c $1/n\sum_{t=1}^n \epsilon_t^2(\hat{\theta}_n)\hat{I}_{n,t}(\hat{\theta}_n) = 1/n\sum_{t=1}^n \epsilon_t^2 I_{n,t} + o_p(1)$, 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 $1/n\sum_{t=1}^n \epsilon_t^2 I_{n,t}/E[\epsilon_t^2 I_{n,t}] \xrightarrow{p} 1$. $\mathcal{QED}$.

APPENDIX B: Proofs of Lemmas A.1-A.6

In order 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)

$$\lim_{a\to\infty}\sup_{\theta\in\Theta}\{|c^\kappa P(|w_t(\theta)| > a) - d_w(\theta)|\} = 0 \tag{12}$$

$$\inf_{\theta\in\Theta}\{d_w(\theta)\} > 0 \quad \text{and} \quad \sup_{\theta\in\Theta}\{d_w(\theta)\} < \infty,$$

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 (12) and Karamata's Theorem

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

The proofs of Lemmas A.1-A.6 require two supporting results. First, trimming indicators satisfy a uniform law.

LEMMA B.1 (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* $\Longrightarrow^*$ *denotes weak convergence on a Polish space.*

PROOF. By construction $\mathcal{I}_{n,t}(\theta)$ is L_2-bounded uniformly on $1 \le t \le n$, $n \ge 1$, and Θ, and geometrically β-mixing. Further, $\{\mathcal{I}_{n,t}(\theta) : \theta \in \Theta\}$ satisfies the metric entropy with L_2-bracketing bound $\int_0^1 \ln(N_{[\,]}(\varepsilon, \Theta, ||\cdot||_2))d\varepsilon < \infty$ 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 $\int_0^1 \ln(N_{[\,]}(\varepsilon, \Theta, ||\cdot||_2))d\varepsilon < \infty$ is then a classic exercise (Giné and Zinn 1984, Pollard 1984). We may therefore apply Doukhan et al's (1995: Theorem 1; eq. (2.17), Application 4) uniform central limit theorem to deduce $\{1/n^{1/2}\sum_{t=1}^n \mathcal{I}_{n,t}(\theta) : \theta \in \Theta\} \Longrightarrow^* \{\mathcal{I}(\theta) : \theta \in \Theta\}$. $\mathcal{QED}$.

Second, intermediate order statistics are uniformly bounded in probability.

LEMMA B.2 (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 $\inf_{\theta\in\Theta} w_t(\theta) \ge 0$ 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. There-

fore 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 $\sup_{u\geq 0} v_1(u) < \infty$. Since $u/k_n^{1/2} \to 0$ it therefore follows for any u

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

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 (15). By construction $k_n^{1/2}\ln(w_{(k_n)}/c_n) \leq u$ for $u \in \mathbb{R}$ sufficiently if $\mathcal{I}_n(u/k_n^{1/2}) \leq 1$, while $\mathcal{I}_n(u/k_n^{1/2}) \leq 1$ if

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

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

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

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 $k_n^{1/2}\ln(w_{(k_n)}/c_n) \leq u$ if $\mathcal{I}_n(u/k_n^{1/2}) \leq 1$ 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 $v_2^2 := \xi^{-2}v_1^2(0)$,

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

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 (16) to obtain for any $u \in \mathbb{R}$

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

The claim now follows from uniform indicator law Lemma B.1 and the mapping theorem. $\mathcal{QED}$.

The proofs of Lemmas A.1-A.6 now follow.

PROOF OF LEMMA A.1.

Claim (a): By Minkowski's inequality

$$\begin{aligned}
&\left|\sum_{t=1}^{n} \epsilon_t y_{t-1} \left\{\hat{I}_{n,t}^{(\epsilon)} \hat{I}_{n,t-1}^{(y)} - I_{n,t}^{(\epsilon)} I_{n,t-1}^{(y)}\right\}\right| \\
&\quad \leq \left|\sum_{t=1}^{n} \epsilon_t \left(\hat{I}_{n,t}^{(\epsilon)} - I_{n,t}^{(\epsilon)}\right) \times y_{t-1} I_{n,t-1}^{(y)}\right| + \left|\sum_{t=1}^{n} \epsilon_t I_{n,t}^{(\epsilon)} \times y_{t-1} \left(\hat{I}_{n,t-1}^{(y)} - I_{n,t-1}^{(y)}\right)\right| \\
&\qquad + \left|\sum_{t=1}^{n} \epsilon_t \left(\hat{I}_{n,t}^{(\epsilon)} - I_{n,t}^{(\epsilon)}\right) \times y_{t-1} \left(\hat{I}_{n,t-1}^{(y)} - I_{n,t-1}^{(y)}\right)\right|.
\end{aligned}$$

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

The indicator $I(u) := I(u \leq 0)$ can be approximated by a *regular* sequence $\{\mathfrak{I}_{\mathcal{N}_n}(u)\}_{n\geq 1}$, cf. Lighthill (1958), where $\{\mathcal{N}_n\}$ is a sequence of finite positive numbers, $\mathcal{N}_n \to \infty$, the rate to be chosen below. Define $\mathfrak{I}_{\mathcal{N}_n}(u) := \int_{-\infty}^{\infty} I(\varpi)\mathrm{S}(\mathcal{N}_n(\varpi - u))\mathcal{N}_n e^{-\varpi^2/\mathcal{N}_n^2} d\varpi$ where $\mathrm{S}(\xi) = e^{-1/(1-\xi^2)} / \int_{-1}^{1} e^{-1/(1-w^2)} dw$ if $|\xi| < 1$ and $\mathrm{S}(\xi) = 0$ if $|\xi| \geq 1$. 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 $\mathfrak{I}_{\mathcal{N}_n}(u)$ is uniformly bounded in u, and continuous and differentiable. Also, $I(u)$ is differentiable except at 0 with derivative $\delta(u) = (\partial/\partial u)I(u) = 0\ \forall u \neq 0$, the Dirac delta function. Therefore $\delta(u)$ has a regular sequence $\mathcal{D}_{\mathcal{N}_n}(u) := (\mathcal{N}_n/\pi)^{1/2} \exp\{-\mathcal{N}_n u^2\}$. 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 $\hat{I}_{n,t}^{(\epsilon)} = I(\mathcal{E}_t(\epsilon_{(k_n^{(\epsilon)})}))$ 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^*, $|c_n^* - c_n| \leq |\epsilon_{(k_n^{(\epsilon)})} - c_n|$,

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

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

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

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

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

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

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

In view of $\sup_{\theta\in\Theta} I_{n,t}^{(\epsilon)}(\theta) I_{n,t-1}^{(y)} \xrightarrow{a.s.} 0$ 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 $K > 0$

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

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

Therefore, by ULLN Lemma A.3.b and dominated convergence each term is $o_p(\sup_{\theta\in\Theta} E|m_{n,t}(\theta)|)$.

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

PROOF OF LEMMA A.2. If $E[\epsilon_t^2] < \infty$ then $\Sigma_n = O(1) = o(n)$. If $E[\epsilon_t^2] = \infty$ consider $\kappa = 2$ and note by independence and two applications of trimmed moment properties (14) $\Sigma_n = E[\epsilon_t^2 I_{n,t}^{(\epsilon)}] \times E[y_{t-1}^2 I_{n,t-1}^{(y)}] = K(\ln(n))^2 = o(n)$. 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 $\Sigma_n = (n^2/(k_n^{(\epsilon)}k_n^{(y)}))^{2/\kappa-1} = o(n)$. $\mathcal{QED}$.

PROOF OF LEMMA A.3.

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 A.2.a variance bound $\Sigma_n = o(n)$ 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). $\mathcal{QED}$.

PROOF OF LEMMA A.4.

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

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

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

We exploit the regular sequences $\{\mathfrak{I}_{\mathcal{N}_n}(u), \mathcal{D}_{\mathcal{N}_n}(u)\}_{n\geq 1}$ defined in the proof of Lemma A.1 (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^*, $|\theta^* - \theta| \leq |\theta - \tilde{\theta}|$

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

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 $\mathcal{D}_{\mathcal{N}_n}(\mathcal{E}_{n,t}(\theta^*))$

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

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

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

Claim (b). In view of order statistic consistency Lemma B.2, the proof is identical to (a) above, and to Lemma A.1.a.

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

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^n \epsilon_t^2(\hat{\theta}_n)\hat{I}_{n,t}(\hat{\theta}_n) &= \frac{1}{n}\sum_{t=1}^n \epsilon_t^2\hat{I}_{n,t} - 2\left(\hat{\theta}_n-\theta^0\right)'\frac{1}{n}\sum_{t=1}^n \hat{m}_{n,t} \\ &\quad + \left(\hat{\theta}_n-\theta^0\right)'\frac{1}{n}\sum_{t=1}^n y_{t-1}^2\hat{I}_{n,t}(\hat{\theta}_n)\left(\hat{\theta}_n-\theta^0\right) + o_p(1).\end{aligned}$$

By Lemma A.1.c $1/n\sum_{t=1}^n \epsilon_t^2\hat{I}_{n,t} = 1/n\sum_{t=1}^n \epsilon_t^2 I_{n,t} + o_p(1)$ 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 A.1.a $1/n\sum_{t=1}^n \hat{m}_{n,t} = 1/n\sum_{t=1}^n m_{n,t} + o_p(|\Sigma_n/n|^{1/2})$ where $(|\Sigma_n/n|^{1/2} = o(1)$ by Lemma A.2 and $1/n\sum_{t=1}^n m_{n,t} = o_p(1)$ by Lemma A.3.a, hence the second term is $o_p(1)$. Next, invoke Lemma A.6.b 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 2.2, and by construction and Lemma A.2 $\mathcal{V}_n^{-1/2} G_n \mathcal{V}_n^{-1/2} = \mathcal{V}_n^{-1/2}\Sigma_n^{1/2}/n^{1/2} \to 0$. Hence the third term is $o_p(1)$.

Claim (d). The proof is the same as (c). $\mathcal{QED}$.

PROOF OF LEMMA A.5. 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 $\kappa = 2$ then $\Sigma_n \sim K(\ln(n))^2$ by (14), hence the integral bounds $(c_n^{(\epsilon)} c_n^{(y)})^2 < \varepsilon^2 \Sigma_n n$ 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 3.b. But this implies for some $N \in \mathbb{N}$ and all $n \geq N$ 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 (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 3.b. $\mathcal{QED}$.

PROOF OF LEMMA A.6.

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

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

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

$$\frac{E\left[m_{n,t}\left(\theta\right)\right] - E\left[m_{n,t}\right]}{\left|\theta - \theta^0\right|} = G_n \times \left(1 + o\left(1\right)\right). \tag{18}$$

Identically, by the definition of a derivative

$$\frac{E\left[m_{n,t}\left(\theta\right)\right] - E\left[m_{n,t}\right]}{\left|\theta - \theta^0\right|} = \frac{\partial}{\partial\theta} E\left[m_{n,t}\left(\theta\right)\right] \times \left(1 + o\left(\left|\theta - \theta^0\right|\right)\right) + o\left(\left\|G_n\right\|\right). \tag{19}$$

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

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

By construction $E[y_{t-1}^2 I_{n,t-1}^{(y)}] = |G_n|$, and by the proofs of Lemmas A.1.a and A.4.a 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 A.1.d and A.4.d $1/n\sum_{t=1}^n y_{t-1}^2 \hat{I}_{n,t}(\hat{\theta}_n) = 1/n\sum_{t=1}^n y_{t-1}^2 I_{n,t} + o_p(1)$, 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 A.1.d and A.4.d. $\mathcal{QED}$.

REFERENCES

Agulló, J., C. Croux, S. Van Aelst (2008). The Multivariate Least-Trimmed Squares Estimator, J. Mult. Anal. 99, 311-338.

An, H.Z. and Z.G. Chen (1982). On Convergence of LAD Estimates in Autoregression with Infinite Variance, J. Mult. Anal. 12, 335-345.

Bassett, G.W. (1991). Equivariant, Monotonic, 50% Breakdown Estimators, Amer. Stat. 45, 135–137.

Brockwell, P. and D.B.H. Cline (1985). Linear Prediction of ARMA Processes with Infinite Variance, Stoch. Proc. Appl. 19, 281-296.

Chen, L.-A., A. H. Welsh and W. Chan (2001). Estimators for the Linear Regression Model Based on Winsorized Observations, Stat. Sin. 11, 147-172.

Čižek, P. (2008). General Trimmed Estimation: Robust Approach to Nonlinear and Limited Dependent Variable Models, Econometric Theory 24 , 1500-1529.

Cline, D.B.H. (1989). Consistency for Least Squares Regression Estimators with Infinite Variance Data, J. Stat. Plan. Infer. 23, 163-179.

Crump, R.K., V.J. Hotz, G.W. Imbens, and O.A. Mitnik (2009). Dealing with Limited Overlap in Estimation of Average Treatment Effects, Biometrika 96, 187-199.

Csörgo, S., L. Horváth, and D. Mason (1986). What Portion of the Sample Makes a Partial Sum Asymptotically Stable or Normal? Prob. Theory Related Fields 72, 1-16.

Davis, R.A. (1996). Gauss-Newton and M-Estimation for ARMA Processes with Infinite Variance, Stoch. Proc. Appl. 63, 75-95.

Davis, R.A. (2010). Heavy Tails in Financial Time Series, in R. Cont (ed.), Encyclopedia Quantitative Finance, Wiley: New York.

Davis, R.A. and S.I. Resnick (1986). Limit Theory for Sample for the Sample Covariance and Correlation Functions of Moving Averages, Ann. Stat. 14, 533-558.

Davis, R.A., W. Wu (1997). M-Estimation for Linear Regression with Infinite Variance, Prob. Math. Stat. 17, 1–20.

Davis, R.A., K. Knight and J. Liu (1992). M-Estimation for Autoregressions with Infinite Variance, Stoch. Proc. Appl. 40, 145-180.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance. Springer-Verlag: Frankfurt.

Finkenstädt, B. and H. Roozén (2001). Extreme Values in Finance, Telecommunications, and the Environment, Chapman and Hall: New York.

Gross, S., and W. L. Steiger (1979). Least Absolute Deviation Estimates in Autoregression with Infinite Variance, J. Appl. Prob. 16, 104-116.

Hahn, M.G., D.C. Weiner and D.M. Mason (1991). Sums, Trimmed Sums and Extremes, Birkhäuser: Berlin.

Hannan, E.J. and M. Kanter (1977). Autoregressive Processes with Infinite Variance, J. Appl. Prob. 14, 411-415.

Hill, J.B. (2009). On Functional Central Limit Theorems for Dependent, Heterogeneous Arrays with Applications to Tail Index and Tail Dependence Estimatio, J. of Stat. Plan. Infer. 139, 2091-2110.

Hill, J.B. (2010). On Tail Index Estimation for Dependent, Heterogeneous Data. Econometric Theory 26, 1398-1436.

Hill, J.B. (2011). Tail and Non-Tail Memory with Applications to Extreme Value and Robust Statistics, Econometric Theory 27, 844-884.

Hill, J.B. (2012). Heavy-Tail and Plug-In Robust Consistent Conditional Moment Tests of Functional Form: to appear in X. Chen and N. Swanson (ed.'s), Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr, Springer: New York.

Hill, J.B. and M. Aguilar (2012). Moment Condition Tests for Heavy Tailed Time Series, J. Econometrics: forthcoming.

Huber, P. J. (1977). Robust Statistical Procedures. Philadelphia: Society for Industrial and Applied Mathematics.

Knight, K. (1987). Rate of Convergence of Centered Estimates of Autoregressive Parameters for Infinite Variance Regressions, J. Time Series Anal. 8, 51-60.

Leadbetter, M.R., G. Lindgren and H. Rootzén (1983). Extremes and Related Properties of Random Sequences and Processes. Springer-Verlag: New York.

Lighthill, M.J. (1958). Introduction to Fourier Analysis and Generalised Functions, Cambridge University Press.

Ling, S. (2005). Self-Weighted LAD Estimation for Infinite Variance Autoregressive Models, J. R. Stat. Soc. Ser. B 67, 381–393.

Ling, S. (2007). Self-Weighted and Local Quasi-Maximum Likelihood Estimators for ARMA-GARCH/IGARCH Models, J. Econometrics 150, 849-873.

Neykov, N.M. and P.N. Neytchev (1990). A Robust Alternative of the Maximum Likelihood Estimator, Short Comm. Compstat, Dubrovnik 1990, 99–100.

Pakes, A. and D. Pollard (1989). Simulation and the Asymptotics of Optimization Estimators, Econometrica, 57, 1027-1057.

Pan, J., H. Wang, and Q. Yao (2007). Weighted Least Absolute Deviations Estimation for ARMA Models with Infinite Variance, Econometric Theory 23, 852-879.

Pham, T.D. and L.T. Tran (1985). Some Mixing Properties of Time Series Models, Stoch. Proc. Appl. 19, 297-303.

Powell, J.L. (1986). Symmetrically Trimmed Least Squares Estimation for Tobit Models, Econometrica 54, 1435-1460.

Resnick, S. (1987). Extreme Values, Regular Variation and Point Processes. Springer-Verlag, New York.

Rousseeuw, P.J. (1984). Least Median of Squares Regression, J. Amer. Stat. Assoc. 79, 871-880.

Rubert, D. and J. Carroll (1980). Trimmed Least Squares Estimation in the Linear Model, J. Amer. Stat. Assoc. 75, 828-838.

Stigler, S.M. (1973). Simon Newcomb, Percy Daniell, and the History of Robust Estimation 1885-1920, J. Amer. Stat. Assoc. 68, 872-879.

Tableman, M. (1994). The Asymptotics of the Least Trimmed Absolute Deviations (LTAD) Estimator, Stat. Prob. Let. 19, 329-337.

Weiss, A.A. (1991). Estimating Nonlinear Dynamic Models Using Least Absolute Error Estimation, Econometric Theory 7, 46-68.

Zhu, K. and S. Ling (2011). The Global LAD Estimators for Finite/Infinite Variance ARMA(p,q) Models, Econometric Theory: forthcoming.

Zhu, K. and S. Ling (2012). Global Self-Weighted and Local Quasi-Maximum Exponential Likelihood Estimators for ARMA-GARCH/IGARCH Models, Ann. Stat.: forthcoming.

TABLE 1 : AR(2) Estimation Results ($\theta_2^0 = -.3$)

$n = 100$					$n = 400$				$n = 800$			
Tail Index $\kappa = .75$												
	Bias	MSE	KS[b]	%[c]	Bias	MSE	KS[b]	%[c]	Bias	MSE	KS	%
LS	-.003	.004	4.75	.000	-.001	.0009	6.50	.000	.000	.0006	6.79	.000
LTS	.005	.001	6.17	.050	.003	.0008	10.0	.050	.002	.0003	10.5	.050
LTTS	.000	.003	3.51	.030	.001	.0016	1.92	.030	.001	.0004	1.08	.020
LWAD	.000	.021	6.17	.050	.000	.0001	2.09	.050	.000	.0001	1.03	.050
Tail Index $\kappa = 1.5$												
	Bias	MSE	KS	%	Bias	MSE	KS	%	Bias	MSE	KS	%
LS	-.007	.007	2.52	.000	.001	.0015	3.18	.000	.000	.0007	2.68	.000
LTS	.009	.004	2.85	.050	-.002	.0003	2.50	.050	.001	.0001	2.42	.050
LTTS	-.000	.006	1.64	.030	.001	.0019	.842	.031	.000	.0008	.763	.021
LWAD	.001	.040	2.31	.050	.000	.0001	1.61	.050	.000	.0004	1.05	.050
Tail Index $\kappa = 2.5$												
	Bias	MSE	KS	%	Bias	MSE	KS	%	Bias	MSE	KS	%
LS	-.008	.008	1.94	.000	-.002	.0020	1.26	.000	-.001	.0003	1.13	.000
LTS	-.001	.004	1.32	.050	.001	.0008	.868	.050	.001	.0002	.816	.050
LTTS	-.001	.007	.758	.030	-.001	.0022	.737	.032	.000	.0008	.527	.021
LWAD	-.001	.050	1.03	.050	.001	.0001	.842	.050	.000	.0001	.632	.050

a. LTTS = Least Tail-Trimmed Squares; LTS = Least Trimmed Squares; LS = Least Squares; LWAD = Least Weighted Absolute Deviations.

b. 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.

c. The total sample proportion trimmed for LTS and LTTS (Tr% = .05 for LTS by construction). In the case of LWAD this represents the percentile used in the weight or .05, cf. Ling (2005).

TABLE 2 : LTTS Wald-Test of AR(1) vs. AR(2)[a]

		Tail Index $\kappa = .75$			Tail Index $\kappa = 1.5$			Tail Index $\kappa = 2.5$		
		$n = 100$			$n = 100$			$n = 100$		
θ_1^0	θ_2^0	10%	.5%	1%	10%	.5%	1%	10%	.5%	1%
.80	.00	.058	.031	.008	.110	.057	.012	.011	.053	.012
.80	-.20	.454	.333	.165	.577	.462	.225	.602	.495	.283
.80	-.30	.756	.654	.442	.849	.772	.571	.862	.781	.582
		$n = 400$			$n = 400$			$n = 400$		
θ_1^0	θ_2^0	10%	.5%	1%	10%	.5%	1%	10%	.5%	1%
.80	.00	.064	.041	.021	.093	.054	.010	.105	.051	.013
.80	-.20	.899	.812	.613	.949	.897	.760	.954	.912	.813
.80	-.30	.975	.944	.915	.993	.993	.982	.994	.994	.991
		$n = 800$			$n = 800$			$n = 800$		
θ_1^0	θ_2^0	10%	.5%	1%	10%	.5%	1%	10%	.5%	1%
.80	.00	.074	.049	.019	.093	.049	.009	.101	.049	.010
.80	-.20	.961	.938	.853	.997	.991	.983	.999	.994	.991
.80	-.30	.992	.985	.971	.999	.998	.997	.999	.998	.996

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