# Robust Minimum Distance Estimation for Nonlinear Semi-Strong GARCH Models

This is a work in progress. Comments are welcome.

Jonathan B. Hill*
Dept. of Economics
University of North Carolina - Chapel Hill

October 27, 2008

**Abstract**

We develop a class of Minimum Distance Estimators for semi-strong Nonlinear ARMAX-Nonlinear GARCH processes. The estimators are asymptotically normal for possibly very heavy-tailed data due to underlying shocks and/or model parameter values. In particular we only impose trivial moment conditions on the GARCH errors, covering non-stationary GARCH. The MDE class is couched within a Method of Moments framework based on tail-trimming nonlinear functions of the data. The theory applies at least to tail-trimmed versions of Generalized Method of Moments, Quasi-Maximum Likelihood, and Least Absolute Deviations. As opposed to trimming the data or the criterion function directly as is universally done, we trim the estimating equations that govern asymptotics. Finally, we propose a unique method for selecting the trimming proportion based on exploiting the untrimmed criterion.

**1. INTRODUCTION** In this paper we develop a class of Minimum Distance Estimators [MDE's] that are robust to very heavy-tails by tail-trimming functions of the data. Applicable MDE's include tail-trimmed versions of Generalized Method of Moments [GMM], Quasi-Maximum Likelihood [QML], and Least Absolute Deviations [LAD]. Unlike extant conventions, we do not directly trim the data or the criterion, but the estimating equations that govern asymptotics.

We focus on a semi-strong Nonlinear ARMAX-Nonlinear GARCH models:

$$(1)\qquad \begin{aligned} y_t &= f\left(x_{t-1}, \alpha\right) + u_t,\ \alpha \in \mathbb{R}^q,\ \beta \in \mathbb{R}^r \\ u_t &= h_t\left(\alpha, \beta\right) \epsilon_t\left(\alpha, \beta\right), \end{aligned}$$

*Dept. of Economics, University of North Carolina-Chapel Hill, www.unc.edu/~jbhill, jbhill@email.unc.edu.

*Key words and phrases*: Minimum Distance Estimation, Method of Moments, tail trimming, robust estimation, extreme value theory, heavy-tails.

*AMS 2000 subject classifications.* Primary ###; secondary ###.

where $h_t(\alpha, \beta)$ is a Borel function measurable with respect to $\Im_{t-1} = \sigma(x_\tau : \tau \leq t - 1)$. The regressors $x_t \in \mathbb{R}^p$ may contain $y_t$ and other random variables, and their lags, and we assume $x_t$ satisfies a weak mixing condition. See Section 2.2 for all assumptions.

We require $f(x_{t-1}, \alpha)$ and $h_t(\alpha, \beta)$ to represent versions of the conditional mean and variance when evaluated at a unique point

$$\theta_0 = [\alpha_0, \beta_0]' \in \mathbb{R}^k, \quad k = q + r.$$

Thus, $\epsilon_t = \epsilon_t(\theta_0)$ satisfies $E[\epsilon_t | x_{t-1}] = 0$ and $E[\epsilon_t^2 | x_{t-1}] = 1$. We now drop redundant parameter arguments to ease the flow.

Class (1) is broad enough to have general appeal to financial and macroeconomic analysts since it nests Smooth Transition ARX, GARCH(1,1), ARMA-GARCH, ARX-Quadratic GARCH, and so on. Nonlinear models of conditional first and second moments capture the asymmetric arrival and impact of news in financial markets. Further, financial time series in particular are known to be heavy-tailed due to random noise $\epsilon_t$, or the structure of volatility clusters $h_t$, or both, and a non-iid framework for $\epsilon_t$ allows additional heterogeneity to be exhibited by higher moments. See Mandelbrot (1963), Engle and Ng (1993), Engle et al (1994), Mittnik and Rachev (2000), Embrechts et al (2003), and Finkenstadt and Rootzén (2003).

### 1.1 Method of Moments

Our estimators are couched within a method of moments framework. Let $m_t(\theta) \in \mathbb{R}^s$, $s \geq k$, be some vector of $\Im_t$-measurable equations, and define a class of MDE's

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \hat{Q}(\theta) \right\} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \left( \frac{1}{n} \sum_{t=1}^{n} m_t(\theta) \right)' \times \hat{\Upsilon}_n \times \left( \frac{1}{n} \sum_{t=1}^{n} m_t(\theta) \right) \right\}$$

for compact $\Theta \subset \mathbb{R}^k$, where $\hat{\Upsilon}_n$ is an $\mathbb{R}^{s \times s}$-valued matrix. Under mild assumptions (e.g. Newey and McFadden 1994)

$$\text{(2)} \qquad \sqrt{n}\left(\hat{\theta} - \theta_0\right) = A_n \times \frac{1}{\sqrt{n}} \sum_{t=1}^{n} m_t(\theta_0) + o_p(1) \text{ for some } A_n \in \mathbb{R}^{k \times s}.$$

An obvious example is GMM where $m_t(\theta) \in \mathbb{R}^s$, $s \geq k$, are estimating equations that identify $\theta_0$ by the moment condition $E[m_t(\theta)] = 0$ *if and only if* $\theta = \theta_0$.

A second example is QML where $m_t(\theta) \in \mathbb{R}^k$ are induced from the first order condition:

$$\text{(3)} \qquad m_t(\theta) = \frac{\partial}{\partial \theta} \ln \phi_t(\theta)$$

where $\phi_t(\theta)$ denotes the standard normal probability density function:

$$\text{(4)} \qquad \phi_t(\theta) := \frac{1}{\sqrt{2\pi} h_t(\alpha, \beta)} \exp\left\{ -\frac{1}{2} \left( y_t - f(x_{t-1}, \alpha) \right)^2 / h_t^2(\alpha, \beta) \right\}.$$

A third example is Least Absolute Deviations [LAD] with criterion

$$(5) \qquad \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\left(y_t - f\left(x_{t-1},\alpha\right)\right)^2}{h_t^2\left(\alpha,\beta\right)} - 1\right|$$

and equations

$$(6) \qquad m_t\left(\theta\right) = \frac{\partial}{\partial\theta}\left\{\frac{\left(y_t - f\left(x_{t-1},\alpha\right)\right)^2}{h_t^2\left(\alpha,\beta\right)}\right\} \times sgn\left(\epsilon_t^2 - 1\right).$$

A fourth example is Log-Transformed LAD [LLAD] with criterion (e.g. Peng and Yao 2003)

$$(7) \qquad \frac{1}{n}\sum_{t=1}^{n}\left|\ln\left(y_t - f\left(x_{t-1},\alpha\right)\right)^2 - \ln h_t^2\left(\alpha,\beta\right)\right|$$

and equations

$$(8) \qquad m_t\left(\theta\right) = \frac{\partial}{\partial\theta}\left\{\ln\left(y_t - f\left(x_{t-1},\alpha\right)\right)^2 - \ln h_t^2\left(\alpha,\beta\right)\right\} \times sgn\left(\ln\epsilon_t^2\right).$$

Due to our martingale difference assumptions we only focus on pure Nonlinear GARCH for these criterion:

$$(1') \qquad y_t = h_t\left(\theta\right)\epsilon_t.$$

Thus, our strongest results are for a robust QMLE.

Even for comparatively non-heavy tailed $\{y_t, \epsilon_t\}$ the equations $m_t(\theta_0)$ can easily have an infinite variance, in which case the limit in (2) is non-Gaussian or does not exist (after re-scaling). Consider GMM estimation of a pure ARCH(1) $y_t = (\beta_0 + \beta_1 y_{t-1}^2)\epsilon_t$ with equations

$$m_t\left(\theta\right) = \left[\begin{array}{c} y_t^2 - \left(\beta_0 + \beta_1 y_{t-1}^2\right) \\ \left\{y_t^2 - \left(\beta_0 + \beta_1 y_{t-1}^2\right) y_{t-1}^2\right\} \\ \left\{y_t^2 - \left(\beta_0 + \beta_1 y_{t-1}^2\right) y_{t-1}^4\right\} \end{array}\right].$$

Then $E[m_{3,t}^2(\theta_0)] = \infty$ if $E[\epsilon_t^4] = \infty$ and/or $E[y_t^8] = \infty$. Since $E[\epsilon_t^4] < \infty$ and $E[y_t^8] = \infty$ are not typically associated with "heavy tails", GMM has only limited used for even a simple, non-heavy tailed GARCH process in finance (see Hill and Renault 2008a).

### 1.2 Tail-Trimmed MDE

In order to ensure asymptotic normality of $\hat{\theta}$ we trim $k_{1,n}$ left-tailed and $k_{2,n}$ right-tailed observations from $\{m_t\left(\theta\right)\}_{t=1}^n$, where each $k_{j,n} \to \infty$ and $k_{j,n}/n \to 0$. Define tail specific observations of $m_{i,t}(\theta)$ and their order statistics,

$$m_{i,t}^{(-)}(\theta) := \min\left\{m_{i,t}(\theta), 0\right\} \quad and \quad m_{i,(1)}^{(-)}(\theta) \leq \cdots \leq m_{i,(n)}^{(-)}(\theta) \leq 0$$
$$m_{i,t}^{(+)}(\theta) := \max\left\{m_{i,t}(\theta), 0\right\} \quad and \quad m_{i,(1)}^{(+)}(\theta) \geq \cdots \geq m_{i,(n)}^{(+)}(\theta) \geq 0,$$

and the stochastically trimmed equations $m_t(\theta)$

$$(9) \qquad \hat{m}_{n,t}(\theta) := \left[ m_{i,t}(\theta) \times I\left( m^{(-)}_{i,(k_{1,n}+1)}(\theta) \leq m_{i,t}(\theta) \leq m^{(+)}_{i,(k_{2,n}+1)}(\theta) \right) \right]_{i=1}^{s},$$

where $I(A) = 1$ is $A$ is true, and 0 otherwise. The Tail-Trimmed MDE [TTMDE] is

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{t=1}^{n} \hat{m}_{n,t}(\theta)' \times \hat{\Upsilon}_n \times \frac{1}{n} \sum_{t=1}^{n} \hat{m}_{n,t}(\theta) \right\}.$$

As long as a deterministically trimmed version $\{m_{n,t}(\theta_0), \Im_t\}$ forms a martingale difference array [*mda*], then

$$V_n^{1/2}\left(\hat{\theta}_n - \theta_0\right) = A_n^{-1/2} \sum_{t=1}^{n} m_{n,t}(\theta_0) + o_p(1) \xrightarrow{d} N(0, I_k).$$

for some sequences $\{V_n, A_n\}$ of positive definite matrices. Notice we propose trimming the process $\{m_t(\theta)\}$ that governs asymptotics of $\hat{\theta}$ a la (2). The martingale difference assumption is not too binding. For each MDE considered $\epsilon_t$ and $\epsilon_t^2 - 1$ must be martingale differences; for LAD and LLAD $\epsilon_t^2 - 1$ must have a zero conditional median, each a common requirement in the literature. Otherwise, we only require $E|\epsilon_t|^p < \infty$ for some $p > 0$.

### 1.2 Literature

Francq and Zakoïan (2004) prove asymptotic normality of the QMLE for strong GARCH and ARMA-GARCH under $E(\epsilon_t^4) < \infty$. Jensen and Rahbek (2004) prove asymptotic normality for non-stationary ARCH. See, also, Hansen and Lee (1994), Lumsdaine (1996) for QML asymptotic theory on GARCH(1,1). In this literature $y_t$ is symmetrically distributed and $E(\epsilon_t^4) < \infty$. Hall and Yao (2003) allow $E(\epsilon_t^4) = \infty$ as long as $E(y_t^2) < \infty$. Each severely restricts applications to financial time series where the arrival of unexpected information $\epsilon_t$ may be heavy tailed; and returns and their volatility clusters may be nonlinear, and heavy-tailed due to the volatility structure.

Linton et al (2008) prove asymptotic normality of the LLAD estimator for non-stationary linear GARCH, provided $E(\epsilon_t^2) < \infty$ for martingale difference $\epsilon_t$, and $E|\epsilon_t|^p < \infty$ for some $p > 0$ for iid $\epsilon_t$. See also Peng and Yao (2003). Asymmetric GARCH and non-pure GARCH including AR-GARCH are not treated. Further, while versions of LAD are touted for being less sensitive to outliers, tail-trimming *is* robust to outliers by construction.

Whittle estimation has far fewer competitive results. Linear, finite variance GARCH $y_t$ with finite kurtosis errors $\epsilon_t$ is the highlight. See, e.g., Giraitis and Robinson (2000) and Mikosch and Straumann (2002).

Trimmed and weighted versions of least squares, LAD and QML are available. Trimming and weighting are always imposed symmetrically on the criterion function $\hat{Q}(\theta)$ or data $y_t$, and always based on a fixed (i.e. non-tail) quantile of $y_t$, or of residuals $\hat{\epsilon}_t$ in a multi-step algorithm. Note symmetric trimming is invalid for asymmetric processes.

Ling (2005, 2007), for example, proves asymptotic normality for a class of self-weighted LADE's for linear autoregressions and QMLE's for symmetric GARCH models with $E(\epsilon_t^2) < \infty$, as long as $E|y_t|^p < \infty$ for some $p > 0$. The criteria are symmetrically weighed and the author only suggests the weight criterion $|y_t| \geq c$ for fixed $c$.

Čižek (2008) delivers a class of General Trimmed Estimators for nonlinear models of the conditional mean. Trimming is symmetrically imposed on the criterion and $\epsilon_t$ must have an everywhere continuous distribution. Both Ling (2007) and Čižek (2008) fail to suggest a means for selecting the weighting or trimming portion and neither performs a simulation experiment. Even Ling's (2005) simulation study of Least Absolute Weighted Deviations lacks theoretical support: the author provides theory for a deterministic weighting criterion $|y_t| \geq c$ but uses a stochastic plug-in $\hat{c}$ for $c$ in the simulation.

See, also, Rubert and Carroll (1980), Rousseeuw (1985), Welsh (1987), Bassett (1991), Chen, Welsh and Chan (2001), Chan and Peng (2005), Čižek (2006), and Agulló, Croux and Van Aelst (2008).

By comparison we asymmetrically tail-trim $m_t(\theta)$,the process that uniquely governs asymptotics; the errors need not have a smooth distribution for GMM and QML, they need only form a *mds* and may be arbitrarily heavy-tailed; the conditional mean and variance models may be nonlinear; we only require trivial moment conditions on $\{y_t, \epsilon_t\}$ allowing non-stationary and nonlinear GARCH; we deliver theory for the case of stochastic trimming thresholds and we offer a unique method for selecting the trimming portion.

Hill and Renault (2008a) develop the Tail-Trimmed GMM estimator for a wide range of models. The present work is an extension and generalization of that paper. The major contributions here are *i*. new robust QML and LAD estimators for asymmetric heavy-tailed semi-strong GARCH models, with a parametric conditional mean in the QML case. *ii*. We provide new proofs of when stochastic $\hat{m}_{n,t}(\theta)$ and deterministic $m_{n,t}(\theta)$ trimming are asymptotically equivalent, a cornerstone requirement that was essentially assumed in Hill and Renault (2008a). *iii*. We provide detailed proofs showing variations on (1) satisfy the major assumptions. *iv*. Finally, we develop for the first time a praxis for selecting the asymmetric tail threshold fractiles $\{k_{1,n}, k_{2,n}\}$ for symmetric or asymmetric models.

The TTMDE is fully developed in Section 2, Sections 3 and 4 discuss assumptions and provide detailed examples. In Section 5 we develop a method for selecting the fractiles $\{k_{1,n}, k_{2,n}\}$, a simulation study is presented in Section 6 and parting comments are left for Section 7. Appendix 1 contains proofs of the main results and Appendix 2 contains supporting results.

Throughout $||\cdot||_p$ denotes both the $L_p$-norm for stochastic matrices $||x||_p := (\sum_{i,j} E|x_{i,j}|^p)^{1/p}$, and the $l_p$-norm otherwise $||x||_p := (\sum_{i,j} |x_{i,j}|^p)^{1/p}$. $||\cdot|| = ||\cdot||_2$. $(z)_+ := \max\{0, z\}$. $K$ denotes a positive, finite constant whose value may change from line to line. Similarly, $\iota > 0$ is an arbitrarily tiny constant whose value may change. Denote by $\overset{p}{\to}$ and $\overset{d}{\to}$ convergence in probability and in distribution. $I_d$ is a $d$-dimensional identity matrix.

## 2. ROBUST MDE

Define the $\sigma$-field

$$\Im_t := \sigma(x_\tau : \tau \leq t).$$

Let $\{m_t(\theta_0)\} = \{m_t(\theta_0)\}_{t\in\mathbb{Z}}$ be a stationary, zero mean $\Im_t$-measurable stochastic process in $\mathbb{R}^s$, $s \geq k$. In this section we do not make any reference to a particular MDE criterion. In Sections 3 and 4, however, we treat standard criteria in order to exemplify the conditions imposed in the present section.

### 2.1 TAIL-TRIMMING

The theory for tail-trimming is molded from fundamental arguments developed in Hill and Renault (2008a,c). First central limit theory is delivered for non-stochastically trimmed $m_{n,t}(\theta)$, defined below, then $\{\hat{m}_{n,t}(\theta) - m_{n,t}(\theta)\} \xrightarrow{p} 0$ sufficiently fast is shown so that all asymptotic theory rests on $m_{n,t}(\theta)$. We proceed in the same manner here. Let positive, real-valued sequences $\{l_{i,n}, u_{i,n}\}_{i=1}^{s}$ and integer fractiles $\{k_{1,n}, k_{2,n}\}$ satisfy

$$k_{j,n} \to \infty,\ k_{j,n}/n \to 0 \text{ and } 1 \leq k_{1,n} + k_{2,n} < n$$

$$l_{i,n} \to \infty \text{ and } u_{i,n} \to \infty,$$

and

$$\frac{n}{k_{1,n}} P\left(m_{i,t}(\theta_0) < -l_{i,n}\right) \to 1 \quad \text{and} \quad \frac{n}{k_{2,n}} P\left(m_{i,t}(\theta_0) > u_{i,n}\right) \to 1. \tag{10}$$

Compactly write

$$c_{i,n} := \max\{l_{i,n}, u_{i,n}\},\ \ c_n := \max_i\{c_{i,n}\},\ \text{ and } k_n := \max\{k_{1,n}, k_{2,n}\}.$$

Now build the tail trimmed moment process

$$m_{n,t}(\theta) := \left[m_{i,t}(\theta) \times I\left(-l_{i,n} \leq m_{i,t}(\theta) \leq u_{i,n}\right)\right]_{i=1}^{s}. \tag{11}$$

Thus $m_{i,n,t}(\theta)$ covers $m_{i,t}(\theta)$ between its lower $k_{1,n}/n^{th} \to 0$ and upper $k_{2,n}/n^{th} \to 0$ quantiles asymptotically.

Symmetric trimming is appropriate for symmetrically distributed data (e.g. AR, GARCH and AR-GARCH). In this case $l_{i,n} = u_{i,n} = c_{i,n}$ $\forall i$ hence

$$m_{n,t}(\theta) := \left[m_{i,t}(\theta) \times I\left(|m_{i,t}(\theta)| \leq c_{i,n}\right)\right]_{i=1}^{s}$$

and

$$\hat{m}_{n,t}(\theta) := \left[m_{i,t}(\theta) \times I\left(|m_{i,t}(\theta)| \leq m^{(a)}_{i,(k_n+1)}(\theta)\right)\right]_{i=1}^{s}$$

where $m^{(a)}_{i,t}(\theta) := |m_{i,t}(\theta)|$ and $(n/k_n)P(|m_{i,t}(\theta_0)| > c_{i,n}) \to 1$. In the sequel we will do not distinguish between the two cases and treat asymmetric trimming (9) and (11) by default.

## 2.2 MAIN RESULTS

Define the trimmed sum covariance matrix and its norm

$$\Sigma_n := E\left[\left(\sum_{t=1}^{n} m_{n,t}(\theta_0)\right)\left(\sum_{t=1}^{n} m_{n,t}(\theta_0)\right)'\right] \text{ and } \sigma_n^2 := \|\Sigma_n\|.$$

The criterion function evaluated with and without trimming is

$$Q_0(\theta) = E\left[m_t(\theta)'\right] \times \Upsilon_0 \times E\left[m_t(\theta)\right]$$

$$Q_n(\theta) = E\left[m_{n,t}(\theta)'\right] \times \Upsilon_n \times E\left[m_{n,t}(\theta)\right]$$

for some positive semi-definite $\Upsilon_0 \in \mathbb{R}^{s\times s}$ and sequence $\{\Upsilon_n\}$, $\Upsilon_n \in \mathbb{R}^{s\times s}$. The sample version of $Q_n(\theta)$ is

$$\hat{Q}_n(\theta) := \left(\frac{1}{n}\sum_{t=1}^{n} \hat{m}_{n,t}(\theta)\right)' \times \hat{\Upsilon}_n \times \left(\frac{1}{n}\sum_{t=1}^{n} \hat{m}_{n,t}(\theta)\right). \quad (12)$$

The TTMDE is therefore $\hat{\theta}_n = \arg\min_{\theta\in\Theta}\{\hat{Q}_n(\theta)\}$.

We require two sets of assumptions that ensure $\hat{Q}_n(\theta)$ is a proper criterion near $\theta_0$, and ensure a properly scaled trimmed sum $\sum_{t=1}^{n} \hat{m}_{n,t}(\theta_0)$ is asymptotically normal. We first state the assumptions and main results, then discuss the assumptions in Sections 3 and 4.

Compactly write the trimmed mean, Jacobian and Hessian

$$\bar{m}_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} m_{n,t}(\theta)$$
$$G_n(\theta) := \frac{\partial}{\partial\theta}E\left[m_{n,t}(\theta)\right] \in \mathbb{R}^{q\times k} \text{ and } G_n = G_n(\theta_0)$$
$$H_n(\theta) := -G_n(\theta)'\Upsilon_n G_n(\theta) \text{ and } H_n = H_n(\theta_0)$$
$$V_n := n^2 H_n \left(G_n'\Upsilon_n\Sigma_n\Upsilon_n G_n\right)^{-1} H_n \text{ and } v_n^2 := \|V_n\|.$$

The first set of conditions concern the criterion $\hat{Q}_n(\theta)$.

**A1.** $||\hat{\Upsilon}_n - \Upsilon_n|| \xrightarrow{p} 0$ *and* $||\Upsilon_n - \Upsilon_0|| \to 0$ *for some sequence* $\{\Upsilon_n\}$ *of positive semi-definite matrices on* $\mathbb{R}^{q\times q}$ *with positive semi-definite limit* $\Upsilon_0$, $0 < ||\Upsilon_0|| < \infty$.

**A2.** $\hat{Q}_n(\hat{\theta}_n) \le \inf_{\theta\in\Theta}\hat{Q}_n(\theta) + o_p(v_n^{-2})$ *and* $\sup_{||\theta-\theta_0||>\delta}\{\hat{Q}_n(\theta)^{-1}\} = O_p(1)$ *for each* $\delta > 0$.

**A3.** $G_n(\theta)$ *exists for all* $\theta$ *in some neighborhood* $U_{\theta_0}$ *of* $\theta_0$; $\liminf_{n\ge1}\inf_{\theta\in U_{\theta_0}}||G_n(\theta)|| > 0$; *and* $||\hat{\theta}_n - \theta_0|| = o_p(1/||G_n||)$.

**A4.** $||H_n^{-1}|| = O(h_n^{-1})$; $\min_{1\le i\le k}\{V_{i,i,n}\} \to \infty$; $||V_n^{-1}|| = O(v_n^{-2})$, $||V_n^{-1/2}|| = O(v_n^{-1})$, *and* $||V_n^{1/2}|| = O(v_n)$.

**A5.** *For all* $\{\delta_n\}$, $\delta_n \to 0$, *that satisfy* $||\hat{\theta}_n - \theta_0|| = O_p(\delta_n)$,

$$\sup_{||\theta-\theta_0||\leq\delta_n} \left\{ \frac{v_n \, \|\bar{m}_n(\theta) - E\left[m_{n,t}(\theta)\right] - \bar{m}_n(\theta_0)\|}{1 + v_n \, \|\theta - \theta_0\|} \right\} \xrightarrow{p} 0.$$

*Remark*: Although A1-A2 help ensure consistency, $||\hat{\theta}_n - \theta_0|| = o_p(1/||G_n||)$ in A3 helps ensure asymptotic normality.

The second set of conditions concern $\{x_t, m_t(\theta_0), m_{n,t}(\theta_0)\}$.

**B1.** *For each* $i = 1...s$ *let* $\lim_{\varepsilon\to\infty} P(|m_{i,t}(\theta_0)| > \varepsilon)/[\varepsilon^{-\kappa_i} L_i(\varepsilon)] \leq 1$ *for some* $\kappa_i > 0$ *and slowly varying* $L_i$, $0 < L_i(\varepsilon) \leq K$ *uniformly in* $\varepsilon > 0$. *In particular,* $L_i$ *are slowly varying with remainder:* $L_i(\lambda x)/L_i(x) = 1 + O(g_i(x))$ *for some measurable function* $g_i : \mathbb{R}_+ \to \mathbb{R}_+$ *with bounded increase*[1], *and* $k_n^{1/2} g_i(c_{i,n}) \to 0$.

**B2.** *There exists* $N \in \mathbb{N}$ *such that* $E[m_{n,t}(\theta_0) m_{n,t}'(\theta_0)]$ *is positive definite,* $||\Sigma_n^{-1}|| = O(\sigma_n^{-2})$, *and* $||\Sigma_n^{-1/2}|| = O(\sigma_n^{-1})$ $\forall n \geq N$. *Further* $c_n^2 = o(n)$.

**B3.** $E[m_t(\theta)] = 0$ if and only if $\theta = \theta_0$, *a unique interior point of compact* $\Theta \subset \mathbb{R}^k$, *and* $\{m_{n,t}(\theta), \Im_t\}$ *forms an adapted mda only at* $\theta = \theta_0$.

**B4.** $\{x_t\}$ *is strictly stationary and strong mixing with coefficients* $\alpha_l = O(l^{-\lambda})$ *with size* $\lambda > r/(r-2)$ *for some* $r > 2$ *(e.g. Ibragimov and Linnik 1971). Further,* $\sigma(\cup \cap_{t\in\mathbb{Z}} \Im_t)$ *is trivial.*

We require one preliminary lemma in order to prove $\hat{m}_{n,t}(\theta) - m_{n,t}(\theta) \xrightarrow{p} 0$ sufficiently fast. Let $m_t^{(\cdot)}(\theta)$ denote any scalar $m_{i,t}^{(-)}(\theta)$ or $m_{i,t}^{(+)}(\theta)$ with a corresponding trimmed version $m_{n,t}^{(\cdot)}(\theta)$ and threshold sequence $\{c_n\}$.

**LEMMA 2.1** *If* $m_t(\theta_0)$ *is adapted to* $\Im_t$ *then under B1 and B4* $1/k_n^{1/2} \sum_{t=1}^n \{I(|m_t^{(\cdot)}(\theta_0)| > c_n) - P(|m_t^{(\cdot)}(\theta_0)| > c_n)\} \xrightarrow{d} N(0, v_1^2)$ *and* $k_n^{1/2} \ln\{|m_{(k_n+1)}^{(\cdot)}(\theta_0)|/c_n\} \xrightarrow{d} N(0, v_2^2)$ *where each* $v_i^2 < \infty$.

**LEMMA 2.2** *If* $\Im_t$ *is adapted to* $m_t(\theta_0)$ *then under B1 and B4* $1/\sigma_n \sum_{t=1}^n \{\hat{m}_{n,t}(\theta_0) - m_{n,t}(\theta_0)\} = o_p(1)$.

*Remark* 1: If $x_t$ is strong mixing and $\sigma(m_\tau(\theta_0) : \tau \leq t) \subseteq \sigma(x_\tau : \tau \leq t)$ then stochastic and deterministic trimming are asymptotically equivalent. Notice adaptability says little about dependence per se[2].

*Remark* 2: By comparison, Hill and Renault (2008a) *assume* $1/k_n \sum_{t=1}^n \{I(|m_t^{(\cdot)}(\theta_0)| > c_n) - P(|m_t^{(\cdot)}(\theta_0)| > c_n)\} = O_p(1/k_n^{1/2})$ and $m_{(k_n+1)}^{(\cdot)}(\theta_0)/c_n = 1 + O_p(1/k_n^{1/2})$ to

[1] There exists $0 < D, z_0 < \infty$ and $\tau \leq 0$ such that $g_i(\lambda z)/g_i(z) \leq D\lambda^\tau$ some for $\lambda \geq 1$, $z \geq z_0$ (Goldie and Smith 1987).

[2] Suppose, for example, $m_t(\theta_0) = \sum_{i=0}^\infty a_i x_{t-i}$, $a_i = O(i^{-\mu})$, $\mu > 1$, an unlikely possibility. Then $\Im_t$ is adapted to $m_t(\theta_0)$ but clearly $m_t(\theta_0)$ need not be strong mixing (e.g. Gorodetskii 1977, Andrews 1984, Davidson 1994).

ensure $1/\sigma_n \sum_{t=1}^{n}\{\hat{m}_{n,t}(\theta_0) - m_{n,t}(\theta_0)\} = o_p(1)$, an irreducibly important condition for asymptotic normality Theorem 2.4, below.

The main results follow.

**THEOREM 2.3** *Under A1-A2 and B1-B4* $\hat{\theta}_n \xrightarrow{p} \theta_0$.

**THEOREM 2.4** *Under A1-A5 and B1-B4* $V_n^{1/2}(\hat{\theta}_n - \theta_0) = A_n^{1/2}\sum_{t=1}^{n} m_{n,t}(\theta_0) + o_p(1) \xrightarrow{d} N(0, I_k)$, *where* $A_n = G_n'\Upsilon_n \{G_n'\Upsilon_n\Sigma_n\Upsilon_n G_n\}^{-1}\Upsilon_n G_n$.

## 2.3 COVARIANCE MATRIX ESTIMATION

Giver the method of moments criterion (12) the optimal weighting sequence $\{\Upsilon_n\}$ in the sense of asymptotic efficiency is $\{\sigma_n^2\Sigma_n^{-1}\}$. See Hill and Renault (2008b: Lemma 2.6), cf. Hansen (1982) and Newey and McFadden (1994). In this case it is easy to show

$$V_n = n^2 \left(G_n'\Sigma_n^{-1}G_n\right),$$

The martingale difference condition B3 and strict stationarity B4 imply $\Sigma_n = nE[m_{n,t}(\theta_0)m_{n,t}'(\theta_0)]$ which is estimated by $\hat{\Sigma}_n = \sum_{t=1}^{n}\hat{m}_{n,t}(\hat{\theta}_n)\hat{m}_{n,t}(\hat{\theta}_n)'$.

**THEOREM 2.5** *Under A1-A5 and B1-B4* $||\hat{\Sigma}_n - \Sigma_n|| = o_p(\sigma_n^2)$.

The asymptotically optimal matrix $V_n$ is now estimated by

$$\hat{V}_n = n^2(\hat{G}_n'\hat{\Sigma}_n^{-1}\hat{G}_{n_n})$$

where a gradient estimator $\hat{G}_n$ can be constructed as follows. See Hill and Renault (2008a); and see McFadden (1989), Pakes and Pollard (1989) and Newey and McFadden (1994) for the finite variance case.

Let $e_j$ be the unit vector, $\{\varepsilon_n\}$ is a $k$-vector sequence of positive numbers, $||\varepsilon_n|| \rightarrow 0$, and construct the $j^{th}$-column

$$\hat{G}_{j,n} = \frac{1}{n}\sum_{t=1}^{n}\left(\hat{m}_{n,t}(\hat{\theta}_n + e_j\varepsilon_{j,n}) - \hat{m}_{n,t}(\hat{\theta}_n - e_j\varepsilon_{j,n})\right) / \left(2\varepsilon_{j,n}\right),\ j = 1...k.$$

**THEOREM 2.6** *Let A1-A5, B1-B4,* $||(G_n'\Sigma_n^{-1}G_n)^{-1}|| = O(||G_n'\Sigma_n^{-1}G_n||^{-1})$, $||\Sigma_n^{-1}|| = O(||\Sigma_n||^{-1})$, $||\varepsilon_n|| \rightarrow 0$ *and* $v_n||\varepsilon_n|| \rightarrow \infty$ *hold. Then* $||V_n\hat{V}_n^{-1} - I_k||| = o_p(1)$.

**3. MDE CRITERIA : A1-A5** Much of A1-A5 represents standard regulatory conditions that ensure a smooth criterion function in some sufficient sense, cf. Pakes and Pollard (1989), Newey and McFadden (1994) and Hill and Renault (2008a,b,c).

**A1** This is standard within the method of moments literature. See Hansen (1982) and Newey and McFadden (1994).

**A2** These are standard smoothness conditions that ensure a unique estimator $\hat{\theta}_n$. The first $\hat{Q}_n(\hat{\theta}_n) \leq \inf_{\theta \in \Theta} \hat{Q}_n(\theta) + o_p(v_n^{-2})$ implies whichever iterative method is used to compute $\hat{\theta}_n$ in fact optimizes $\hat{Q}_n(\theta)$ on $\Theta$. The second $\sup_{||\theta - \theta_0|| > \delta}\{\hat{Q}_n(\theta)^{-1}\}$ $= O_p(1)$ implies the criterion does not get arbitrarily close to its minimum 0 at discrete distances from $\theta_0$.

In the cases of LAD and LLAD condition A2 is partially ensured if $\epsilon_t$ is stationary with a positive continuous density in a neighborhood of 0 (Pollard 1991, Davis et al 1992).

**A3** This is standard for non-differentiable criteria: while $\hat{m}_{n,t}(\theta)$ is not differentiable in $\theta$ we require $E[m_{n,t}(\theta)]$ to be differentiable near $\theta_0$ so that classic optimization arguments dating to Huber (1967) apply, cf. Pakes and Pollard (1989) and Newey and McFadden (1994). The gradient $G_n(\theta)$ will exist and be non-generate near $\theta_0$ evidently except in extraordinary cases. The rate $||\hat{\theta}_n - \theta_0|| = o_p(1/||G_n||)$ seems to force consistency, but consistency is a primitive property that requires only A1-A2 and consistency of $1/n \sum_{t=1}^{n} \hat{m}_{n,t}(\theta)$ (see the proof of Theorem 2.3). Thus, $||\hat{\theta}_n - \theta_0|| = o_p(1/||G_n||)$ simply sharpens the rate of convergence, it automatically holds under A1-A2 and B1-B3 when $\lim_{n\to\infty} ||G_n|| < \infty$ as in thin-tailed cases, and holds in general when $\Upsilon_n$ is selected to optimize asymptotic efficiency (Hill and Renault 2008a: Section 2.4).

We require $\inf_{\theta \in U_{\theta_0}} ||G_n(\theta)|| > 0$ uniformly in $n$ to ensure the Hessian does not break down near $\theta$. Since $U_{\theta_0}$ is otherwise arbitrary this requirement is mild.

**A4** Nonsingularity of the Hessian is standard. The remaining properties imply $H_n$ and $V_n$ are proportional in norm to their normed inverses, properties that simplify asymptotic arguments.

**A5** This, along with A3, substitutes for differentiability of $m_{n,t}(\theta)$ while imposing minimal smoothness, cf. Huber (1967). See Pakes and Pollard (1989), New and McFadden (1994) and Hill and Renault (2008b).

## 4 TAILS, THRESHOLDS AND MDA : B1 - B4

Conditions B1-B3 concern the data generating processes of the untrimmed and trimmed equations $m_t(\theta_0)$ and $m_{n,t}(\theta_0)$. We therefore focus attention on $m_t(\theta_0)$ generated by GMM and QML from (1), and LAD and LLAD for (1′). The equations $m_t(\theta)$ are constructed by case. Since $m_t(\theta)$ in all cases represents a first order condition process we require differentiability.

**C1.** *$f(x_{t-1}, \alpha)$ and $h_t(\alpha, \beta)$ are $\Im_{t-1}$-measurable with $\Im_{t-1}$-measurable gradients in $\theta$.*

**GMM** The estimating equations $m_t(\theta_0)$ are constructed as

$$m_t(\theta) = \begin{bmatrix} u_t(\alpha) \times g_t^{(1)}(\theta) \\ \\ \{u_t^2(\alpha) - h_t(\alpha, \beta)\} \times g_t^{(2)}(\theta) \end{bmatrix}. \tag{13}$$

We write $g_t^{(1)}(\theta)$ to denote an $\Im_{t-1}$-measurable vector of stochastic weights that contains at least $(\partial/\partial\alpha)f(x_t,\alpha)$. Similarly, $g_t^{(2)}(\theta)$ is $\Im_{t-1}$-measurable and contains at least $(\partial/\partial\theta)h_t(\alpha,\beta)$. Certainly other equations are possible depending on the economic and statistical information available. See, e.g., Hall (2005).

**QML** Use the QML framework (3) and (4) to deduce, sans multiplicative constants,

$$(14) \qquad m_t(\theta) = \begin{bmatrix} \left\{ \frac{1}{2}\left(\epsilon_t^2 - 1\right) \frac{\partial}{\partial\alpha} h_t^2(\alpha,\beta) + u_t(\alpha)\frac{\partial}{\partial\alpha} f\left(x_{t-1},\alpha\right) \right\} \frac{1}{h_t^2(\alpha,\beta)} \\ \\ \left\{ \epsilon_t^2 - 1 \right\} \frac{\partial}{\partial\beta} \ln h_t^2(\alpha,\beta) \end{bmatrix}.$$

**LAD** The LAD criterion (5) and equations (6) imply for pure Nonlinear GARCH ($1'$)

$$(15) \qquad m_t(\theta) = sgn\left(\epsilon_t^2 - 1\right) \times \epsilon_t^2 \times \frac{\partial}{\partial\theta} \ln h_t^2(\alpha,\beta).$$

**LLAD** Finally, the LLAD framework (7) and (8) for ($1'$) implies

$$(16) \qquad m_t(\theta) = sgn\left(\ln \epsilon_t^2\right) \times \frac{\partial}{\partial\theta} \ln h_t^2(\theta).$$

### 4.1 TAIL BOUND : B1

The tail bound both ensures a moment $E|m_{i,t}(\theta_0)|^p < \infty$ for some $p > 0$, and expedites extreme value theoretic arguments used in tail trimming theory (Hill and Renault 2008a,c).

The tail bound holds for thin (exponential) and heavy (regularly varying) tailed data generating processes for $\{m_{i,t}(\theta_0)\}$. This will be widely satisfied, and it is difficult to imagine a distribution class encountered in finance and macroeconomics where the tails of $\{m_{i,t}(\theta_0)\}$ decay slower than a regularly varying function. See, e.g., Leadbetter et al (1983), Resnick (1987) and Finkenstadt and Rootzén (2003).

What is clear is the more "nonlinear" a model becomes the more tedious a proof of B1. We therefore restrict attention to three popular versions of (1) and consider simple GMM weight specifications $g_t^{(i)} := g_t^{(i)}(\theta_0)$ for the sake of brevity.

**Example 1 (STAR):** The first model is Smooth Transition AR(1):

$$(17) \qquad y_t = \alpha_1 y_{t-1} \exp\left\{-\left(y_{t-1} - \alpha_2\right)^2\right\} + \epsilon_t, \;\; |\alpha_1| < 1$$

$$g_t^{(1)} = \begin{bmatrix} y_{t-1} \exp\left\{-\left(y_{t-1} - \alpha_2\right)^2\right\} \\ \alpha_1 y_{t-1} \exp\left\{-\left(y_{t-1} - \alpha_2\right)^2\right\} \times \left(y_{t-1} - \alpha_2\right) \\ y_{t-2} \end{bmatrix};$$

**Example 2 (AR-ARCH):** Suppose $y_t$ is AR(1) with ARCH(1) errors:

(18) $$y_t = \alpha y_{t-1} + u_t, u_t = h_t\epsilon_t,\ h_t^2 = \beta_0 + \beta_1 u_{t-1}^2,\ |\alpha| < 1, \beta_0 > 0,$$

$$\sup_{t\in\mathbb{Z}} E\,|u_t|^\iota < \infty \text{ and } \rho := \sup_{t\in\mathbb{Z}} E\,|\epsilon_t|^\iota \in (0,1) \text{ for infinitessimal } \iota > 0,$$

$$0 \leq \beta_1 \|\epsilon_t\|_\iota^{1/\kappa} \leq \tilde{\rho} \in (0,1),$$

$$g_t^{(1)} = [y_{t-1},\ y_{t-2}]' \text{ and } g_t^{(2)} = \frac{1}{h_t^2} \times \left[1,\ u_{t-1}^2,\ u_{t-2}^2\right]';$$

**Example 3 (Threshold-GARCH(1,1)):** Assume $y_t$ is GARCH(1,1) if $y_t < 0$:

(19) $$y_t = h_t\epsilon_t, \text{ and } h_t^2 = \beta_0 + \left(\beta_1 y_{t-1}^2 + \beta_2 h_{t-1}^2\right) \times I\left(y_{t-1} < 0\right), \beta_0 > 0,$$

$$\sup_{t\in\mathbb{Z}} E\,|y_t|^\iota < \infty \text{ and } \sup_{t\in\mathbb{Z}} E\,|\epsilon_t|^\iota \in (0,1) \text{ for infinitessimal } \iota > 0,$$

$$0 \leq \beta_i \|\epsilon_t\|_\iota^{1/\kappa} < 1, i = 1, 2,$$

$$g_t^{(2)} = \left[1,\ y_{t-1}^2 I\left(y_{t-1} < 0\right),\ y_{t-2}^2,\ h_{t-1}^2 I\left(y_{t-1} < 0\right),\ h_{t-2}^2\right]'.$$

**C2.** $\limsup_{\epsilon>0} P(|\epsilon_t| > \epsilon)/\varepsilon^{-\kappa} \leq K < \infty$ *for some* $\kappa > 0$ *such that* $E|\epsilon_t|^{\kappa-\iota} < \infty$.

**C3.** $\limsup_{\epsilon>0} P(|\epsilon_t| > \epsilon|\Im_{t-1})/\varepsilon^{-\kappa} \leq K < \infty$.

*Remark*: C2 and independence suffice for C3: $P(|\epsilon_t| > \epsilon|\Im_{t-1}) = P(|\epsilon_t| > \epsilon) = O(\varepsilon^{-\kappa})$ with probability one. Independence, however, is far more severe than actually required, and C3 acts as an extremal version of martingale difference B3.

The logic of the proofs of the following claims can be straightforwardly extended to cases like STAR-GARCH, AR-STGARCH, Quadratic GARCH, etc.

**LEMMA 4.1 [STAR]** *Under C1-C3 tail bound B1 holds for GMM and QML* $m_t(\theta_0)$ *in (13)-(14) generated from STAR (17).*

**LEMMA 4.2 [AR-ARCH]** *Under C1-C3 tail bound B1 holds for GMM and QML* $m_t(\theta_0)$ *in (13)-(14) generated from AR-ARCH (18).*

**LEMMA 4.3 [TGARCH]** *Under C1-C3 tail bound B1 holds for GMM, QML, LAD and LLAD* $m_t(\theta_0)$ *in (13)-(16) generated from TGARCH (19).*

*Remark*: The AR-ARCH and TGARCH condition $\beta_i||\epsilon_t||_\iota^{1/\kappa} \leq \tilde{\rho} \in (0,1)$ is similar in spirit to constraints imposed in Davidson's (2004) analysis of ARCH($\infty$) processes. Since $\iota > 0$ is arbitrarily small it is mild to suppose $E|\epsilon_t|^\iota \leq \rho \in (0,1)$. Then $\beta_i||\epsilon_t||_\iota^{1/\kappa} \leq \beta_i\rho^{1/(\iota\kappa)} < 1$ requires $\beta_i \leq \rho^{-1/(\iota\kappa)}$ which can be arbitrarily large therefore covering slopes within the non-stationary range.

### 4.2 Covariance Matrix and Thresholds : B2

Positive definiteness of $E[m_{n,t}(\theta_0)\, m'_{n,t}(\theta_0)]$ is standard. We state it for sufficiently large $n \geq N$ to ensure non-degeneracy in the presence of trimming. Under martingale difference B3 this implies $\Sigma_n = nE[m_{n,t}(\theta_0)\, m'_{n,t}(\theta_0)]$ is positive definite and $\liminf_{nN} \sigma_n^2/n > 0$.

The condition $c_n^2/n \to 0$ helps regulate the rate of convergence of the trimmed mean $1/n \sum_{t=1}^{n} m_{n,t}(\theta_0)$ and therefore of the TTMDE $\hat{\theta}_n$. Notice under martingale difference B3 and the construction of each $c_{i,n}$, $\sigma_n^2/n^2 = ||E[m_{n,t}(\theta_0)\, m'_{n,t}(\theta_0)]||/n \leq Kc_n^2/n \to 0$. Thus, we are effectively ruling out super-consistency of the trimmed mean (cf. Hill and Renault 2008a,c).

If $m_t(\theta_0)$ is symmetrically distributed then $c_n^2/n \to 0$ holds simply if $k_n \to \infty$ sufficiently fast. In order to see this, under tail bound B1 $m_{i,t}(\theta_0)$, $c_{i,n}$, and some $k_n$ satisfy $k_n/n \sim P(|m_{i,t}(\theta_0)| > c_{i,n}) = O(c_{i,n}^{-\kappa_i})$ hence $c_{i,n} = O((n/k_n)^{1/\kappa_i})$. Therefore $c_n^2/n = \max_i\{c_{i,n}^2\}/n = \max\{(n/k_n)^{2/\kappa_i}/n\} \to 0$ only if $n^{1-\min_i\{\kappa_i\}/2}/k_n \to 0$. This is trivial when $\min_i\{\kappa_i\} \geq 2$ since we assume $k_n \to \infty$. A similar argument applies in the asymmetric case.

**LEMMA 4.4** *If $m_t(\theta_0)$ is symmetrically distributed then B1 and $k_n \sim n^{\delta}$ for any $\delta \in (1 - \min_i\{\kappa_i\}/2, 1)$ suffice for B2.*

Finally, by standard matrix norm properties $||\Sigma_n||^{-1} \leq K||\Sigma_n^{-1}||$ and $||\Sigma_n||^{1/2} \leq ||\Sigma_n^{1/2}||$. Thus, $||\Sigma_n^{-1}|| = O(\sigma_n^{-2})$ and $||\Sigma_n^{-1/2}|| = O(\sigma_n^{-1})$ requires proportionality: $||\Sigma_n||^{-1} \sim K||\Sigma_n^{-1}||$ and $||\Sigma_n||^{1/2} \sim K||\Sigma_n^{1/2}||$. As in A4 this greatly expedites limit theory for trimmed sums of martingale differences.

### 4.3 Martingale Difference : B3

If $\{m_t(\theta_0), \Im_t\}$ forms a *mds* then all components of B3 hold: the untrimmed mean $E[m_t(\theta_0)] = 0$ by iterated expectations and the conditionally trimmed $E[m_{n,t}(\theta_0)|\Im_{t-1}] = 0$ for well chosen threshold sequences $\{l_{i,n}, u_{i,n}\}$. If $m_t(\theta_0)$ is symmetrically distributed then $E[m_{n,t}(\theta_0)|\Im_{t-1}] = 0$ under symmetric trimming for any threshold sequence $\{c_n\}$: see Hill and Renault (2008a,b).

Consider verifying the *mds* property for $\{m_t(\theta_0), \Im_t\}$.

**C4.** $E[\epsilon_t|\Im_{t-1}] = 0$.

**C5.** $E[\epsilon_t^2|\Im_{t-1}] = 1$.

**C6.** $P(\epsilon_t^2 > 1|\Im_{t-1}) = P(\epsilon_t^2 \leq 1|\Im_{t-1})$.

**C7.** $\int_0^1 P(\epsilon_t^2 > u|\Im_{t-1}) = 1/2$.

**LEMMA 4.5 (GMM)** *Under GMM (13), $\{m_t(\theta_0), \Im_t\}$ forms a* mds *sufficiently if C1, C4 and C5 hold, and $g_t^{(i)}(\theta)$ is $\Im_{t-1}$ measurable.*

**LEMMA 4.6 (QML)** *Under QML (14), $\{m_t(\theta_0), \Im_t\}$ forms a* mds *sufficiently if C1, C4 and C5 hold, and $(\partial/\partial\alpha) f(x_t, \alpha)$ and $(\partial/\partial\theta) h_t^2(\alpha, \beta)$ are $\Im_{t-1}$ measurable.*

Finally, the LAD and LLAD equations can only be shown to form martingale difference sequences in the pure GARCH case.

**LEMMA 4.7 (LAD)** *Under LAD (14) for GARCH (1′), $\{m_t(\theta_0), \Im_t\}$ forms a* mds *sufficiently if C1, C5 and C7 and $(\partial/\partial\theta)h_t^2(\alpha,\beta)$ is $\Im_{t-1}$-measurable.*

**LEMMA 4.8 (LLAD)** *Under LAD (14) for GARCH (1′), $\{m_t(\theta_0), \Im_t\}$ forms a* mds *sufficiently if C1 and C6 hold and $(\partial/\partial\theta)h_t^2(\alpha,\beta)$ is $\Im_{t-1}$-measurable.*

### 4.4 Mixing Regressors $\mathbf{x_t}$ : B4

We require $x_t$ to be strictly stationary and mixing in order to ensure $\hat{m}_{n,t}(\theta_0)$ approximates $m_{n,t}(\theta_0)$ sufficiently fast. By assuming $x_t$ is stationary and strong mixing we ensure $\Im_t$-measurable tail arrays of $m_{n,t}(\theta_0)$ have the central limit property a la Hannan (1973), which then links $\hat{m}_{n,t}(\theta_0)$ to $m_{n,t}(\theta_0)$ (Hill and Renault 2008a,c). Sufficient conditions for nonlinear autoregressions (e.g. STAR in Example 1) to be geometrically ergodic, and therefore geometrically strong mixing, are treated in An and Huang (1996) and Leibscher (2005), to name a very few. Sufficient conditions for linear and nonlinear GARCH and AR-GARCH processes (e.g. AR-ARCH and TGARCH Examples 2 and 3) to be geometrically strong mixing are delivered in Carrasco and Chen (2002) and Meitz and Saikkonen (2007).

**5. OPTIMAL TAIL FRACTILE SELECTION** Implicit in condition B1 is the assumption the trimming fractiles $\{k_{1,n}, k_{2,n}\}$, and therefore the trimming thresholds $\{l_{i,n}, u_{i,n}\}$, satisfy $E[m_{n,t}(\theta)] = 0$ *if and only if* $\theta = \theta_0$. Clearly this arrangement is not unique: for any $\theta \in \Theta$ there exist possibly infinitely many sequences $\{l_{i,n}, c_{i,n}\}$ that satisfy $E[m_{n,t}(\theta)] = 0$.

Consider the class of tail fractiles $k_{1,n} = [n^{\delta_1}]$ and $k_{2,n} = [n^{\delta_2}]$, $\delta = [\delta_1, \delta_2] \in (0,1)^2$. We need a criterion for selecting any $\delta$ that uniquely ensures $E[m_{n,t}(\theta)] = 0$ *if and only if* $\theta = \theta_0$. Write explicitly

$$\hat{m}_{n,t}(\theta,\delta) := \left[m_{i,t}(\theta) \times I\left(m^{(-)}_{i,([n^{\delta_1}]+1)}(\theta) \leq m_{i,t}(\theta) \leq m^{(+)}_{i,([n^{\delta_2}]+1)}(\theta)\right)\right]_{i=1}^{s},$$

and re-write the TTMDE criterion

$$\hat{Q}_n(\theta,\delta) = \frac{1}{n}\sum_{t=1}^{n}\hat{m}_{n,t}(\theta,\delta)' \times \hat{\Upsilon}_n^{(\delta)} \times \frac{1}{n}\sum_{t=1}^{n}\hat{m}_{n,t}(\theta,\delta)$$

and the estimator

$$\hat{\theta}_n(\delta) := \underset{\theta\in\Theta}{\operatorname{argmin}}\left\{\hat{Q}_n(\theta,\delta)\right\}.$$

We write $\hat{\Upsilon}_n(\delta)$ to cover cases like the asymptotically efficient weight

$$\hat{\Upsilon}_n(\delta) = \left(E\left[m_{n,t}(\theta_0,\delta)m_{n,t}(\theta_0,\delta)'\right] / \left\|E\left[m_{n,t}(\theta_0,\delta)m_{n,t}(\theta_0,\delta)'\right]\right\|\right)^{-1}.$$

Let $D$ be the largest compact subset of $(0,1)^2$, and let $D^* \subseteq D$ contain all $\delta^*$ that satisfy B3: $E[m_{n,t}(\delta^*, \theta)] = 0$ *if and only if* $\theta = \theta_0$. Theorem 2.4 implies $\forall \delta^* \in D^*$ and some positive definite matrix $V_n(\delta^*)$, under A1-A5 and B1-B4

$$(20) \qquad V_n^{1/2}(\delta^*)\left(\hat{\theta}_n(\delta^*) - \theta_0\right) = A_n^{1/2}(\delta^*)\sum_{t=1}^{n} m_{n,t}(\delta^*, \theta_0) + o_p(1) \xrightarrow{d} N\left(0, I_k\right).$$

Define the sample criterion with untrimmed equations

$$\hat{Q}\left(\theta\right) = \frac{1}{n}\sum_{t=1}^{n} m_t\left(\theta\right)' \times \hat{\Upsilon} \times \frac{1}{n}\sum_{t=1}^{n} m_t\left(\theta\right).$$

The estimator $\hat{\theta}_n(\delta)$ satisfies $1/n\sum_{t=1}^{n} m_{n,t}(\hat{\theta}_n(\delta)) \xrightarrow{p} 0$, while *any* method of moments estimator $\hat{\theta}$ *must* satisfy $1/n\sum_{t=1}^{n} m_t(\hat{\theta}) \xrightarrow{p} 0$. Since only the latter identifies $\theta_0$ we suggest choosing $\delta$ by minimizing the *untrimmed* $\hat{Q}(\theta)$ evaluated at $\hat{\theta}_n(\delta)$. Define

$$(21) \qquad \hat{\delta}_n^* := \text{med}\left\{\hat{\delta}_{1,n} + \hat{\delta}_{2,n} : \hat{\delta}_n = \underset{\delta \in D}{\text{argmin}}\left\{\hat{Q}(\hat{\theta}_n(\delta))\right\}\right\}$$

and use $\hat{\theta}_n(\hat{\delta}_n^*)$ as the final estimate of $\theta_0$. Intuitively, since only $\delta^* \in D^*$ identify $\theta_0$ evidently $\text{argmin}_{\delta \in D}\{\hat{Q}(\hat{\theta}_n(\delta))\} \in D^*$ as $n \to \infty$ with probability one. The construction of $\hat{\delta}_n^*$ controls for the possibility that $D^*$ contains more than one element. If, for example, $m_t(\theta_0)$ is symmetrically distributed then $\hat{\theta}_n(\delta)$ is consistent and asymptotically normal for *all* $\delta_1 = \delta_2$ (i.e. $D^* = D$). Nevertheless values near 0 will render $\hat{\theta}_n(\hat{\delta}_n^*)$ very volatile and non-Gaussian for even large samples, and values near 1 will remove too many $m_t(\theta)$ from the criterion and therefore severely bias $\hat{\theta}_n(\hat{\delta}_n^*)$. See Hill and Renault's (2008a) TTGMM simulation study. Clearly there are other ways to select $\hat{\delta}_n^*$ from the set $\{\text{argmin}_{\delta \in D}\hat{Q}(\hat{\theta}_n(\delta))\}$ as $n \to \infty$. The vector $\hat{\delta}_n^*$ that minimizes the normed covariance $V_n^{-1}(\hat{\delta}_n^*)$ is one intuitive alternative.

**THEOREM 5.1** *Under A1-A5 and B1-B4* $V_n^{1/2}(\hat{\delta}_n^*)(\hat{\theta}_n(\hat{\delta}_n^*) - \theta_0) \xrightarrow{d} N(0,1)$.

*Remark* 1: In practice the analyst will perform a grid search over a countable version of $D$. For example, $\hat{\delta} = \text{argmin}_{\delta \in D_n}\hat{Q}_n^{(\delta)}(\hat{\theta}_n(\delta))$ for $D_n = \{1/[n^{1/4}], .2/[n^{1/4}], ..., ([n^{1/4}] - 1)/[n^{1/4}]\}^2$.

*Remark* 2: If the data generating process is symmetric then there is one fractile $k_n = [n^\delta]$ and the problem reduces to a line search.

**6. SIMULATION STUDY** In this section we analyze the small sample properties of Tail-Trimmed QML [TTQML] and Tail-Trimmed LAD [TTLAD] estimators $\hat{\theta}_{n,q}$ and $\hat{\theta}_{n,l}$. We compare our estimators to Ling's (2005, 2007) Quasi-Maximum Weighted Likelihood [QMWL] and Least Absolute Weighted Deviations [LAWD] estimators $\hat{\theta}_{n,q}^*$ and $\hat{\theta}_{n,l}^*$, and Čížek's (2008) Generalized Trimmed Estimators based on

QML and LAD criteria, respectively $\hat{\theta}^{\circ}_{n,g}$ and $\hat{\theta}^{\circ}_{n,l}$. See Hill and Renault (2008a) for a comparison of TTGMM with GMM and QML.

### 6.1 Models

The models considered are AR(1), GARCH(1,1), ARX(1)-GARCH(1,1), STARX(1)-Threshold Integrated ARCH(1), and Smooth Transition GARCH(1,1). Let $N_{0,1}$ be a standard normal law and $P_\gamma$ a symmetric Pareto law with index $\gamma > 0$ : if $\epsilon_t$ is governed by $P_\gamma$ then $P(\epsilon_t > \epsilon) = P(\epsilon_t < -\epsilon) = .5 \times (1+\epsilon)^{-\gamma}$. Each data generating process, auxiliary variables and the error distribution, are described in Table 1. We generate 100 samples of size $n = 1000$ for each model.

**TABLE 1 - Data Generating Processes**

| Type | Model | iid $\epsilon_t$ | $x_t \in \mathbb{R}$ | $\theta_{0,k}$ |
|---|---|---|---|---|
| AR | $y_t = .9y_{t-1} + \epsilon_t$ | $P_{1.5}$ | | .9 |
| GARCH | $y_t = h_t\epsilon_t$ | | | |
| | $h_t^2 = 1 + .4u_{t-1}^2 + .5h_{t-1}^2$ | $P_{2.5}$ | | .5 |
| | $h_t^2 = 1 + .4u_{t-1}^2 + .6h_{t-1}^2$ | $N_{0,1}$ | | .6 |
| STGARCH | $y_t = h_t\epsilon_t$ | | | |
| | $h_t^2 = 1 + \exp\left\{-.1 \times y_{t-1}^2\right\} \times \left(.4y_{t-1}^2 + .5h_{t-1}^2\right)$ | $P_{2.5}$ | | .5 |
| | $h_t^2 = 1 + \exp\left\{-.1 \times y_{t-1}^2\right\} \times \left(.4y_{t-1}^2 + .6h_{t-1}^2\right)$ | $N_{0,1}$ | | .6 |
| ARX-IGARCH | $y_t = .9y_{t-1} + .5x_t + u_t,\ u_t = h_t\epsilon_t$ | | iid $P_{2.5}$ | |
| | $h_t^2 = 1 + .4u_{t-1}^2 + .6h_{t-1}^2$ | $N_{0,1}$ | | .6 |
| STARX-TARCH | $y_t = e^{-.5\times y_{t-1}^2} \times \{.9y_t + .5x_t\} + u_t,\ u_t = h_t\epsilon_t$ | | iid $P_{2.5}$ | |
| | $h_t^2 = 1 + .6y_{t-1}^2 \times I(y_t < 0)$ | $P_{2.5}$ | | .6 |
| | $h_t^2 = 1 + .6y_{t-1}^2 \times I(y_t < 0)$ | $N_{0,1}$ | | .6 |

### 6.2 Estimator Computation

The TTQML and TTLAD estimators $\hat{\theta}_n \in \{\hat{\theta}_{n,q}, \hat{\theta}_{n,l}\}$ are computed based on equations (14) and (16) with weight $\hat{\Upsilon}_n = I_s$. Hill and Renault (2008a) do not find any improvement in the TTGMM estimator when the asymptotically efficient matrix $\hat{\Upsilon}_n = \hat{\Sigma}_n^{-1}$ is used, and related simulations (not reported here) corroborate this finding for other TTMDE's.

We compute each $\hat{\theta}_n$ using the two-step algorithm of Section 5. Define the set $d_n = \{1/[n^{1/3}], 2/[n^{1/3}]..., ([n^{1/3}] - 1)/[n^{1/3}]\}$. First $\hat{\theta}_n(\delta)$ is computed for each $\delta \in D_n$, and then $\hat{\delta}_n^* = \arg\min_{\delta \in D_n}\{\hat{Q}(\hat{\theta}_n(\delta))\}$, where $D_n = d_n$ for symmetric DGP's and $D_n = d_n \times d_n$ for asymmetric DGP's.

Ling's (2005, 2007) QMWL and LAWD estimators $\hat{\theta}^*_{n,q}$ and $\hat{\theta}^*_{n,l}$ are computed by minimizing

$$\text{QMWL} : \hat{Q}(\theta) = \frac{1}{n}\sum_{t=1}^{n} w_t(c) \times \ln \phi_t(\theta)$$

$$\text{LAWD} : \hat{Q}(\theta) = \frac{1}{n}\sum_{t=1}^{n} w_t(c) \times \left|(y_t - f(x_{t-1}, \alpha))^2 - h_t^2(\alpha, \beta)\right|$$

where $\phi_t(\theta)$ is defined in (4), the weight $w_t(c)$ is measurable with respect to $\sigma(y_t, y_{t-1}, ...)$, and $c > 0$ is a pre-chosen threshold. In both cases we use Ling's (2005, 2007) recommended symmetric weight, cf. Huber (1977):

$$w_t = \left(\max\left\{1, c^{-1}\sum_{l=1}^{4}\frac{1}{l^9}|y_{t-l}|\, I(|y_{t-l}| > c)\right\}\right)^{-4}.$$

Ling (2005, 2007) does not provide a theoretical basis for selecting $c$, does not study the QMWL estimator in a simulation experiment, and only studies the LAWT estimator by using the 90th and 95th percentiles of $y_t$ as plug-ins for $c$[3]. We simply replicate this choice for both $\hat{\theta}^*_{n,q}$ and $\hat{\theta}^*_{n,l}$.

Čižek's (2008) criterion is for some pre-chosen integer $1 \leq k < n$

$$\text{GTE} : \hat{Q}(\theta) = \frac{1}{n}\sum_{i=n-k+1}^{n} s_{(n-i+1)}(\theta)$$

where $s_t(\theta) \geq 0$ is an estimating equation, and $s_{(1)}(\theta) \geq s_{(2)}(\theta) \geq \cdots$ the order statistics[4]. Thus only the $k$ smallest equations are used. The QML and LAD versions are $s_t(\theta) = -\ln\phi_t(\theta)$ and $s_t(\theta) = |(y_t - f(x_{t-1}, \alpha))^2 - h_t^2(\alpha, \beta)|$ resulting in $\hat{\theta}^{\circ}_{n,g}$ and $\hat{\theta}^{\circ}_{n,l}$. Similar to Ling (2007), since Čižek (2008) does not suggest what $k$ should be or how to choose it we simply use $k = [\xi n]$ for $\xi \in \{.90, .95\}$, thus trimming 5% or 10% of the largest $s_t(\theta)$.

Note only our TTQMLE (and Hill and Renault's 2008 TTGMME) is shown to be asymptotically normal for general models of the conditional mean *and* variance. We only focus on GARCH models for the TTLADE; Ling's (2005) LAWDE and all of Čižek's (2008) estimators are analyzed solely for models of the conditional mean; and Ling's (2007) QMWLE is shown to be asymptotically normal only for linear GARCH models. Nevertheless we estimate all models with all estimators and focus attention on those estimator-model matches supported by the literature.

### 6.3 Summary of Results

[3] Ling's (2005) use of a stochastic threshold in his simulation experiment for $\hat{\theta}^*_{n,l}$ is not supported by the given theory since that requires a deterministic $c$. It remains to be proven that either $\hat{\theta}^*_{n,q}$ or $\hat{\theta}^*_{n,l}$ is asymptotically when the threshold is stochastic.

[4] Čižek's (2008) generalizes the trimming criteria beyond simply ordering $s_t(\theta)$ based on its own values by introducing a trimming criterion $r_t(\theta)$. This generality apparently serves data generating processes associated with discrete choice models, which we do not consider.

We analyze estimator performance by focusing on the $k^{th}$-parameter estimate $\hat{\theta}_k$ in each case. See Table for the true value of $\theta_{0,k}$ for each model. Let $\hat{\theta}$ denote any estimator used in this study, and let $\{\hat{\theta}_{j,k}\}_{j=1}^{100}$ be the sample of iid $k^{th}$-parameter estimates. We use the simulation standard deviations $s_\theta = \{(1/100)\sum_{j=1}^{100}(\hat{\theta}_{j,k} - \theta_{0,k})^2\}^{1/2}$ to generate a sequence of ratios $\{Z_j\}_{j=1}^{100}$, $Z_j = (\hat{\theta}_{j,k} - \theta_{0,k})/s_\theta$. Estimator performance is restricted to simulation sample means, standard deviations, rejection frequencies of $Z$-tests of the hypothesis $\theta_k = \theta_{0,k}$ and Kolmogorov-Smirnov tests of standard normality performed on $\{Z_j\}_{j=1}^{100}$. See Tables 2 and 3 for all QML estimates and Tables 4 and 5 for all LAD estimates.

####

## 7. CONCLUSION

####

# APPENDIX 1: Proofs of Main Results

**Proof of Lemma 2.1.** In the following we exploit $L_2$-Weak Dependence [$L_2$-WD], $L_p$-Mixingale Sequence [$L_p$-MS] and $L_p$-Near Epoch Dependence [$L_p$-NED] properties. Consult respectively Hannan (1973), McLeish (1975) and Davidson (1994: Chapter 17) for details. We define the properties in context, below.

Throughout write

$$z_t := |m_t^{(\cdot)}(\theta_0)|, \quad I_{n,t}(u) := I\left(z_t > c_n e^u\right) \quad \text{and} \quad S_n(u) = \frac{1}{k_n}\sum_{t=1}^{n} I_{n,t}\left(u/k_n^{1/2}\right).$$

**Step 1:** Suppose $\{I_{n,t}(u)\}$ is $L_2$-WD on $\{\Im_t\}$: $\sum_{t=1}^{\infty} ||E[I_{n,t}(u)|\Im_1] - E[I_{n,t}(u)|\Im_0]|| < \infty$ (Hannan 1973; cf. Wu and Min 2005). Note $I_{n,t}(u)$ is regular under B4 since it is $\Im_t$-measurable. Therefore $k_n^{1/2}(S_n(u) - E[S_n(u)]) \xrightarrow{d} N(0, v_1^2)$ with $v_1^2 < \infty$ by Theorem 1.i of Hannan (1973). We will prove adaptation suffices for $\{I_{n,t}(u), \Im_t\}$ to be an $L_2$-MS, which suffices for $\{I_{n,t}(u)\}$ to be $L_2$-WD on $\{\Im_t\}$.

Since $\Im_t$ is adapted to $m_t(\theta_0)$, any $\{z_t\}$ is trivially $L_p$-NED on $\{\Im_t\}$ for any $p > 0$ with constants $d_t = 0$ an coefficients $\psi_l = o(l^{-\lambda})$ for any size $\lambda > 0$:

$$\left\| z_t - E\left[z_t|\Im_{t-l}^{t+l}\right]\right\|_p = \|z_t - z_t\|_p = 0 \leq d_t \times \psi_l = 0 \times o(l^{-\lambda}).$$

See Davidson (1994: Chapter 17). By Lemma 2.1 of Hill (2008) this implies $\{I_{n,t}(u)\}$ is $L_2$-NED on $\{\Im_t\}$ with constants $d_{n,t}(u)$, $\sup_{t\in\mathbb{Z}}\sup_{u\geq 0} d_{n,t}(u) = O((k_n/n)^{1/r})$ and coefficients $\varphi_{l_n} = o((k_n/n)^{1/2-1/r} l_n^{-\lambda})$ for any $r > 2$ and size $\lambda > 0$, where $l_n \to \infty$ as $n \to \infty$:

$$\left\| I_{n,t}(u) - E\left[I_{n,t}(u)|\Im_{t-l_n}^{t+l_n}\right]\right\|_2 \leq d_{n,t}(u) \times \varphi_{l_n} \leq o(l_n^{-\lambda}).$$

Therefore $\{I_{n,t}(u), \Im_t\}$ forms an $L_2$-MS with uniformly bounded constants $e_{n,t}(u)$ and coefficients $\varphi_{l_n}$ since $\Im_t$ is induced by a strong mixing $x_t$ with size $r/(r-2)$: $||E[I_{n,t}(u)|\Im_{-\infty}^{t-l_n}]||_2 \leq e_{n,t}(u) \times \varphi_{l_n}$ and $||I_{n,t}(u) - E[I_{n,t}(u)|\Im_{-\infty}^{t+l_n}]||_2 \leq e_{n,t}(u) \times \varphi_{l_n+1}$. See Theorem 17.5 of Davidson (1994), cf. McLeish (1975).

Now use Minkowski's inequality, the mixingale property with uniformly bounded constants, and the definition of size to deduce for any $\lambda > 1$

$$\begin{aligned}
&\sum_{t=1}^{\infty} \| E[I_{n,t}(u)|\Im_1] - E[I_{n,t}(u)|\Im_0] \| \\
&\qquad \leq \sum_{l=1}^{\infty} \left\| E\left[I_{n,t}(u)|\Im_{-\infty}^{t-l}\right]\right\| + \sum_{l=1}^{\infty} \left\| E\left[I_{n,t}(u)|\Im_{-\infty}^{t-l+1}\right]\right\| \\
&\qquad \leq K \sum_{l=1}^{\infty} \varphi_l \leq K \sum_{l=1}^{\infty} l^{-\lambda} < \infty \; \forall \lambda > 1.
\end{aligned}$$

Therefore $\{I_{n,t}(u)\}$ is $L_2$-WD on $\{\Im_t\}$.

**Step 2:** Consider the second claim. Using arguments similar to Hsing (1991: p. 1553) we will prove $k_n^{1/2} \ln\{z_{(m_n+1)}/c_n\} \xrightarrow{d} N(0, v_2^2)$ is a consequence of Step 1. By construction $k_n^{1/2} \ln\{z_{(m_n+1)}/c_n\} \leq u$ sufficiently if $S_n(u) \leq \rho$ for any $\rho \in (0,1]$ to be chosen below, in turn if

$$\begin{aligned}
k_n^{1/2}(S_n(u) - E[S_n(u)]) &\leq k_n^{1/2}\left(\rho - \frac{n}{k_n} P\left(z_t > c_n e^{u/k_n^{1/2}}\right)\right) \\
&= k_n^{1/2}\left(\rho - \frac{n}{k_n} P(z_t > c_n) \frac{P\left(z_t > c_n e^{u/k_n^{1/2}}\right)}{P(z_t > c_n)}\right).
\end{aligned}$$

Since the tail bound implies

$$P(z_t > z) = \mathbf{O}(z^{-\kappa}) \times L(z),$$

where $\mathbf{O}(z) \leq z$, an argument identical to Hsing's (1991: p. 1553) reveals

$$\frac{n}{k_n} P(z_t > c_n) = \mathbf{O}(1) \times [1 + O(g(c_n))] = \mathbf{O}(1) + o\left(1/k_n^{1/2}\right)$$

and

$$\frac{P(z_t > c_n e^u)}{P(z_t > c_n)} = \mathbf{O}(e^{-u\kappa}) \times (1 + O(g(c_n))) = \mathbf{O}(e^{-u\kappa}) \times \left(1 + o\left(1/k_n^{1/2}\right)\right).$$

Therefore $k_n^{1/2} \ln\{z_{(m_n+1)}/c_n\} \leq u$ sufficiently if

$$\begin{aligned}
&\kappa^{-1} k_n^{1/2}(S_n(u) - E[S_n(u)]) \\
&\quad \leq \kappa^{-1} k_n^{1/2}\left(\rho - \left(\mathbf{O}(1) + o\left(1/k_n^{1/2}\right)\right) \times \mathbf{O}(e^{-u\kappa}) \times \left(1 + o\left(1/k_n^{1/2}\right)\right)\right) \\
&\quad \leq \kappa^{-1} k_n^{1/2}\left\{\rho - \mathbf{O}\left(e^{-u\kappa/k_n^{1/2}}\right) \times \left(1 + o\left(1/k_n^{1/2}\right)\right)\right\} \\
&\quad \leq \kappa^{-1} k_n^{1/2}\left\{u\kappa/k_n^{1/2} + o\left(1/k_n^{1/2}\right)\right\} \\
&\quad = u + o(1),
\end{aligned}$$

where the third inequality exploits $\rho = \mathbf{O}(1) \in (0,1]$ since $\rho$ is otherwise arbitrary. Since $\kappa^{-1}k_n^{1/2}\{S_n(u) - E[S_n(u)]\} \xrightarrow{d} Z$ a normal law by Step 1, it follows

$$\begin{aligned}\lim_{n\to\infty} P &\left(k_n^{1/2}\ln\{z_{(m_n+1)}/c_n\} \leq u\right) \\ &= \lim_{n\to\infty} P\left(p^{-1}k_n^{1/2}\left(S_n(u) - E\left[S_n(u)\right]\right) \leq u + o(1)\right) \\ &= P\left(Z \leq u\right).\end{aligned}$$

Therefore $k_n^{1/2}\ln\{z_{(m_n+1)}/c_n\} \xrightarrow{d} N(0, v_2^2)$ where $v_2^2 < \infty$. ■

**Proof of Lemma 2.2.** In order to conserve space assume $m_t \geq 0$ $a.s.$ is univariate ($q = 1$), the general case being similar but laborious. Then $\hat{m}_{n,t} = m_t \times I(m_t \leq m_{(k_n+1)})$ and $m_{n,t} = m_t \times I(m_t \leq c_n)$ where for some $k_n = o(n)$ and $k_n \to \infty$

$$1_n := \frac{n}{k_n}P\left(m_t > c_n\right) \to 1.$$

Define the number of sample threshold exceedances

$$k_n^* := \sum_{t=1}^{n} I\left(m_t > c_n\right),$$

and write

$$\begin{aligned}(22)\qquad \sum_{t=1}^{n}\{\hat{m}_{n,t} - m_{n,t}\} &= \sum_{i=1}^{n} m_{(i)}\left\{I\left(m_{(i)} \leq m_{(k_n+1)}\right) - I\left(m_{(i)} \leq c_n\right)\right\} \\ &= \sum_{i=k_n+1}^{k_n^*} m_{(i)} - \sum_{i=k_n^*+1}^{k_n} m_{(i)}.\end{aligned}$$

**Step 1:** We first bound $\sum_{i=k_n+1}^{k_n^*} m_{(i)}$ and $\sum_{i=k_n^*+1}^{k_n} m_{(i)}$. For any sequence $\{\varepsilon_n\}$ of positive real numbers that satisfies $\varepsilon_n \to 0$ and $\varepsilon_n k_n^{1/2-\iota} \sim K$ for infinitessimal $\iota > 0$

$$\begin{aligned}(23)\qquad \sum_{i=k_n+1}^{k_n^*} m_{(i)} &\leq \sum_{i=k_n+1}^{[k_n(1_n+\varepsilon_n)]} m_{(i)} + \sum_{i=[k_n(1_n+\varepsilon_n)]+1}^{k_n^*} m_{(i)} \\ &\leq \sum_{i=k_n+1}^{[k_n(1_n+\varepsilon_n)]} m_{(i)} + \left(k_n^* - \left[k_n\left(1_n + \varepsilon_n\right)\right]\right)_+ \times m_{(k_n+1)} \\ &= \sum_{i=k_n+1}^{[k_n(1_n+\varepsilon_n)]} m_{(i)} + o_p(c_n)\end{aligned}$$

and

(24)

$$\begin{aligned}
\sum_{i=k_n^*+1}^{k_n} m_{(i)} &\leq \sum_{i=k_n^*+1}^{[k_n(1_n-\varepsilon_n)]} m_{(i)} + \sum_{i=[k_n(1_n-\varepsilon_n)]+1}^{k_n} m_{(i)} \\
&\leq \sum_{i=[k_n(1-\varepsilon_n)]+1}^{k_n} m_{(i)} + \left(k_n^* - [k_n(1_n-\varepsilon_n)]\right)_+ \times m_{([k_n(1_n-\varepsilon_n)]+1)} \\
&\quad + \left(k_n^* - [k_n(1_n-\varepsilon_n)]\right)_+ \times \left(m_{(k_n^*+1)} - m_{([k_n(1_n-\varepsilon_n)]+1)}\right)_+ \\
&= \sum_{i=[k_n(1-\varepsilon_n)]+1}^{k_n} m_{(i)} + o_p(c_n)
\end{aligned}$$

The last line in (23) follows from Chebychev's inequality, $\varepsilon_n k_n^{1/2-\iota} \sim K$, $1_n = (n/k_n)P(m_t > c_n)$ and Lemma 2.1 since $m_{(k_n+1)}/c_n = 1 + O_p(1)$ and

$$\begin{aligned}
&P\left(k_n^* \geq k_n\left(1_n + \varepsilon_n\right)\right) \\
&\quad \leq P\left(k_n^{1/2-\iota}\left|\frac{1}{k_n}\sum_{t=1}^{n} I\left(m_t > c_n\right) - 1_n\right| \geq k_n^{1/2-\iota}\varepsilon_n\right) \\
&\quad \leq K \times E\left(k_n^{1/2-\iota}\left\{\frac{1}{k_n}\sum_{t=1}^{n}\left\{I\left(m_t > c_n\right) - P\left(m_t > c_n\right)\right\}\right\}\right)^2 = o(1).
\end{aligned}$$

Similarly, the last line in (24) is a consequence of

$$\begin{aligned}
&P\left(k_n^* \leq k_n\left(1 - \varepsilon_n\right)\right) \\
&\quad = P\left(\frac{1}{k_n}\sum_{t=1}^{n} I\left(m_t > c_n\right) - 1_n \leq -\varepsilon_n\right) \\
&\quad \leq P\left(k_n^{1/2-\iota}\left|\frac{1}{k_n}\sum_{t=1}^{n}\left\{I\left(m_t > c_n\right) - P\left(m_t > c_n\right)\right\}\right| \geq k_n^{1/2-\iota}\varepsilon_n\right) = o(1)
\end{aligned}$$

hence

$$P\left(\left(m_{(k_n^*+1)} - m_{([k_n(1_n-\varepsilon_n)]+1)}\right)_+ \geq 0\right) = P\left(k_n^* \leq k_n\left(1 - \varepsilon_n\right)\right) = o(1).$$

**Step 2:** We now prove the claim. Use (22)-(24), the triangular inequality and

$m_{([k_n(1_n+\varepsilon_n)]+1)} \leq m_{(i)}$ $\forall i \leq [k_n(1_n + \varepsilon_n)] + 1$ to deduce

$$\begin{aligned}
&\left|\sum_{t=1}^{n} \{\hat{m}_{n,t} - m_{n,t}\}\right| \\
&\quad \leq \left|\sum_{i=k_n+1}^{k_n^*} m_{(i)} - k_n\varepsilon_n m_{([k_n(1_n+\varepsilon_n)]+1)}\right| + \left|\sum_{i=k_n^*+1}^{k_n} m_{(i)} - k_n\varepsilon_n m_{([k_n(1_n-\varepsilon_n)]+1)}\right| \\
&\quad \leq \left|\sum_{i=k_n+1}^{[k_n(1_n+\varepsilon_n)]} \left(m_{(i)} - m_{([k_n(1_n+\varepsilon_n)]+1)}\right)\right| + \left|\sum_{i=[k_n(1_n-\varepsilon_n)]+1}^{k_n} \left(m_{(i)} - m_{([k_n(1_n-\varepsilon_n)]+1)}\right)\right| \\
&\qquad + \left|k_n\left(1_n + \varepsilon_n\right) - \left[k_n\left(1_n + \varepsilon_n\right)\right]\right| \times m_{([k_n(1_n+\varepsilon_n)]+1)} \\
&\qquad + \left|k_n\left(1_n - \varepsilon_n\right) - \left[k_n\left(1_n - \varepsilon_n\right)\right]\right| \times m_{([k_n(1_n-\varepsilon_n)]+1)} + o_p(1) \\
&\quad = A_n + B_n + C_n + D_n + o_p(1),
\end{aligned}$$

say. We need only show $\{A_n, B_n, C_n, D_n\} = O_p\left(c_n k_n^{\iota}\right)$ for infinitessimal $\iota > 0$ since $c_n k_n^{\iota}/\sigma_n \leq c_n k_n^{\iota}/n^{1/2} = o(1)$ under B2 and a continuity argument.

The proofs for $A_n$ and $B_n$ are identical so consider $A_n$. By two applications of the triangular inequality

$$\begin{aligned}
&\frac{1}{c_n}\left|\sum_{i=k_n+1}^{[k_n(1_n+\varepsilon_n)]} \left(m_{(i)} - m_{([k_n(1_n+\varepsilon_n)]+1)}\right)\right| \\
&\qquad \leq O\left(k_n\varepsilon_n\right) \times \left(\left|\frac{m_{(k_n+1)}}{c_n} - 1\right| + \left|\frac{m_{([k_n(1_n+\varepsilon_n)]+1)}}{c_n} - 1\right|\right).
\end{aligned}$$

Since $(n/k_n)P(m_t > c_n) \to 1$ and $(n/k_n(1_n + \varepsilon_n))P(m_t > c_n) \sim (1_n + \varepsilon_n)^{-1} \to 1$, Lemma 2.1 applies to equally to $m_{(k_n+1)}$ and $m_{([k_n(1_n+\varepsilon_n)])}$. Therefore

$$\begin{aligned}
&\frac{1}{k_n^{\iota} c_n}\left|\sum_{i=k_n+1}^{[k_n(1+\varepsilon_n)]} \left(m_{(i)} - m_{([k_n(1_n+\varepsilon_n)]+1)}\right)\right| \\
&\qquad \leq O\left(k_n^{1-\iota}\varepsilon_n\right) \times \left(\left|\frac{m_{(k_n+1)}}{c_n} - 1\right| + \left|\frac{m_{([k_n(1_n+\varepsilon_n)]+1)}}{c_n} - 1\right|\right) \\
&\qquad = O_p\left(\varepsilon_n k_n^{1/2-\iota}\right) = O_p\left(1\right),
\end{aligned}$$

hence $A_n = O_p\left(c_n k_n^{\iota}\right)$.

Finally, consider $C_n$, the proof for $D_n$ being identical. Since truncation implies $z - [z] \in [0,1]$ for $z > 0$, use Lemma 2.1 to deduce

$$\begin{aligned}
\frac{C_n}{c_n k_n^{\iota}} &= \left|\frac{k_n\left(1_n + \varepsilon_n\right) - \left[k_n\left(1_n + \varepsilon_n\right)\right]}{k_n^{\iota}}\right| \times \frac{m_{([k_n(1_n+\varepsilon_n)]+1)}}{c_n} \\
&\leq k_n^{-\iota} \times \left(1 + O_p\left(1/k_n^{1/2}\right)\right) = o_p(1).
\end{aligned}$$

■

**Proof of Theorem 2.3.** Consistency $\hat{\theta}_n \xrightarrow{p} \theta_0$ follows from Theorem 2.1 of Hill and Renault (2008b) under A1-A2, provided $Q_0(\theta_0) = 0$, $Q_n(\theta_0) = 0$, $Q_n(\theta_0) = o_p(1)$ and $\hat{Q}_n(\theta_0) = o_p(1)$. Identification $E[m_t(\theta_0)] = 0$ under B3 implies $Q_0(\theta_0) = 0$, and martingale difference B3 and iterated expectations imply $E[m_{n,t}(\theta_0)] = 0$ hence $Q_n(\theta_0) = 0$.

Finally, $\hat{Q}_n(\theta_0) = o_p(1)$ follows from $||\Upsilon_n - \Upsilon_0|| < \infty$ and $||\Upsilon_0|| < \infty$ under A1 provided $1/n\sum_{t=1}^n \hat{m}_{n,t}(\theta_0) \xrightarrow{p} 0$. The latter limit holds under Theorem 4.1 of Hill and Renault (2008a) if their Assumptions A-D hold. Assumptions A-C hold respectively under B1-B3, and Lemma 2.1 ensures Assumption D under B1, B3 and B4. ■

**Proof of Theorem 2.4.** The claim $V_n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_k)$ follows from Theorem 2.4 of Hill and Renault (2008b) provided the following conditions hold[5]:

**H1.** *$\Sigma_n$ is positive definite for each $n$, where $\sigma_n^2/n^2 = o(1)$ and $||\Sigma_n^{-1}|| = O(\sigma_n^{-2})$.*

**H2.** *$G_n(\theta_0)$ exists on $U_{n,0}$, $\liminf_{n\geq 1} ||G_n|| > 0$, and $||\hat{\theta}_n - \theta_{n,0}|| = o_p(1/||G_n||)$.*

**H3.** *$||\hat{\Upsilon}_n - \Upsilon_n|| \xrightarrow{p} 0$ and $||\Upsilon_n - \Upsilon_0|| \to 0$ for some sequence $\{\Upsilon_n\}$ of deterministic, positive semi-definite matrices, with positive semi-definite limit $\Upsilon_0$, $0 < ||\Upsilon_0|| < \infty$.*

**H4.** *$E[m_t(\theta_0)] = 0$ and $E[m_{n,t}(\theta_0)] = 0$ for a unique interior point $\theta_0 \in \Theta$.*

**H5.** *$H_n$ is non-singular for each $n$, $h_n := \inf_{\theta\in U_{n,0}} ||H_n(\theta)|| > 0$ uniformly in $n$, and $||H_n^{-1}|| = O(h_n^{-1})$. Further $\min_{1\leq i\leq k} V_{n,i,i} \to \infty$, $||V_n^{-1}|| = O(v_n^{-2})$, $||V_n^{-1/2}|| = O(v_n^{-1})$, and $||V_n^{1/2}|| = O(v_n)$.*

**H6.** *Let $A_n \sum_{t=1}^n \hat{m}_{n,t}(\theta_0) \xrightarrow{d} N(0, I_q)$ for all sequences of $\mathbb{R}^{q\times q}$-valued matrices $\{A_n\}$ satisfying $A_n\Sigma_n A_n' \to I_q$.*

**H7.** *For all $\{\delta_n\}$, $\delta_n \to 0$, that satisfy $||\hat{\theta}_n - \theta_0|| = O_p(\delta_n)$,*

$$\sup_{||\theta-\theta_{n,0}||\leq\delta_n} \left\{ \frac{v_n \left\| m_n(\theta) - m_n(\theta_0) - E\left[m_{n,t}(\theta)\right]\right\|}{1 + v_n \left\|\theta - \theta_0\right\|} \right\} \xrightarrow{p} 0.$$

We verify the conditions in two groups:

**H1-H5, H7:** Condition H1 holds under B2 and the *mds* condition B3 since by properties of the Euclidean norm and the Cauchy-Schwartz inequality $\sigma_n^2/n^2 = ||E\left[m_{n,t}(\theta_0)m_{n,t}(\theta_0)'\right]||/n \leq Kc_n^2/n \to 0$; conditions H2 and H3 are A1 and A3 respectively. Condition H4 holds under B3 while H5 holds under A1 and A4. Finally, H7 is A5.

**H6:** The limit law $A_n \sum_{t=1}^n \hat{m}_{n,t}(\theta_0) \xrightarrow{d} N(0, I_q)$ is valid under Theorem 4.2 of Hill

[5] These are labeled differently than in Hill and Renault (2008b) to avoid confusion. Also, condition H2 has been reduced from the original condition to simple sufficient conditions.

and Renault (2008a) if their Assumptions A-D hold, which hold under B1-B4 and Lemma 2.1. ■

**Proof of Theorems 2.5 and 2.6.** Using arguments similar to the proofs of Theorem 2.2 and 2.3, Theorems 2.4 and 2.5 follow respectively from Theorem 2.4 of Hill and Renault (2008a) and Theorem 2.8 of Hill and Renault (2008b). ■

**Proof of Lemma 4.1.** We first establish tail bounds for $y_t$ and $\epsilon_t y_t$.

**Step 1 ($y_t$):** Since $y_t = \sum_{i=0}^{\infty} \psi_{t,i}\epsilon_{t-i}$ with probability one, where $|\psi_{t,i}| \leq |\alpha_1|^i$ *a.s.*, write $x_t = \sum_{i=0}^{\infty} \psi_{t,i}\epsilon_{t-i}$ and $x_t^{(N)} = \sum_{i=0}^{N} \psi_{t,i}\epsilon_{t-i}$ for arbitrary $N \in \mathbb{N}$. Subadditivity and Markov's and Minkowski's inequalities imply for some small $\iota > 0$

$$\begin{aligned} P(x_t > x) &\leq P\left(\left|x_t - x_t^{(N)}\right| > x/2\right) + P\left(\left|x_t^{(N)}\right| > x/2\right) \\ &\leq Kx^{-\kappa+\iota}\left\|x_t - x_t^{(N)}\right\|_{\kappa-\iota}^{\kappa-\iota} + P\left(\left|x_t^{(N)}\right| > x/2\right) \\ &\leq Kx^{-\kappa+\iota}\left|\alpha_1\right|^{(N+1)(\kappa-\iota)} + P\left(\left|x_t^{(N)}\right| > x/2\right) \\ &= o\left(x^{-\kappa+\iota}\right) + P\left(\left|x_t^{(N)}\right| > x/2\right), \end{aligned}$$

as $N \to \infty$. It therefore suffices to bound $P(|x_t^{(N)}| > x)$. Now use subadditivity, $|\psi_{t,i}| \leq |\alpha_1|^i$ and C2 to deduce

$$P\left(\left|x_t^{(N)}\right| > x\right) \leq \sum_{i=0}^{N} P\left(\left|\alpha_1\right|^i \left|\epsilon_{t-i}\right| > x/N\right) = O\left(N^{\kappa}\sum_{i=0}^{N}\left|\alpha_1\right|^i\right) \times x^{-\kappa}.$$

Since $N$ is arbitrary we can always fix $N \sim x^{\iota/\kappa}$ such that $KN^{\kappa}\sum_{i=0}^{N}\left|\alpha_1\right|^i x^{-\kappa} \leq Kx^{\kappa-\iota}$. This proves $P(|y_t| > y) = O(y^{\kappa-\iota})$.

**Step 2 ($\epsilon_t y_{t-i}$):** Under C2 and C3 $P(|\epsilon_t| > \epsilon) = O(\epsilon^{-\kappa})$ and $P(|\epsilon_t| > \epsilon|y_{t-i}) = O(\epsilon^{-\kappa})$ for any $i \geq 1$, and $P(|y_t| > y) = O(y^{\kappa-\iota})$ by Step 1 for tiny $\iota > 0$. Now use Lemma A.1 to conclude

$$P\left(\left|\epsilon_t y_{t-i}\right| > x\right) = O\left(y^{\kappa-2\iota}\right).$$

**Step 3 ($m_t(\theta)$ : GMM, QML):** We prove the claim for GMM since QML is similar. Since $|m_{1,t}(\theta)| \leq |\epsilon_t y_{t-1}|$ and $|m_{1,t}(\theta)| = |\epsilon_t y_{t-2}|$, tail bounds follows from Step 2. Similarly, since $|m_{2,t}(\theta)| \leq |\epsilon_t|y_{t-1}^2| + \alpha_2|\epsilon_t| \times |y_{t-1}|$ the tail bound follows from subadditivity and Steps 1 and 2. ■

**Proof of Lemma 4.2.** In Steps 1-3 preliminary probability bounds for $\epsilon_t u_{t-1}$, $y_t$, $u_t$, $u_t u_{t-1}$, and $u_t y_t$ are established. Steps 4-5 then deliver the bound for all elements $m_{i,t}(\theta_0)$.

**Step 1 ($\epsilon_t u_{t-1}$):** Define $w_t := \epsilon_t^2 u_{t-1}^2$. By repeated substitution for any $N \in \mathbb{N}$

$$w_t = \beta_0 \sum_{i=0}^{N-1} \beta_1^i \prod_{j=0}^{i+1} \epsilon_{t-j}^2 + \beta_1^N \prod_{j=0}^{N+1} \epsilon_{t-j}^2 h_{t-N-1}^2 = w_t^{(N)} + R_N$$

say. Subadditivity, Markov's inequality, $E|u_t|^\iota < \infty$ and $\beta_1 ||\epsilon_t||_\iota^{1/\kappa} \leq \tilde{\rho} \in (0,1)$ imply

$$\begin{aligned} P(w_t > x) &\leq Kx^{-\iota} \|R_N\|_\iota^\iota + P\left(w_t^{(N)} > x/2\right) \\ &= Kx^{-\iota}\beta_1^N \left(E|\epsilon_t|^{2\iota}\right)^{N+2} + P\left(w_t^{(N)} > x/2\right) \\ &\leq K\tilde{\rho}^N x^{-\iota} + P\left(w_t^{(N)} > x/2\right) = o\left(x^{-\iota}\right) + P\left(w_t^{(N)} > x/2\right) \end{aligned}$$

hence it suffices to bound $P(w_t^{(N)} > x/2)$ as $x \to \infty$. By subadditivity

$$P\left(w_t^{(N)} > x^2/2\right) \leq \sum_{i=0}^{N} P\left(\beta_0\beta_1^i \prod_{j=0}^{i+1} \epsilon_{t-j}^2 > x^2/[2N]\right).$$

Since $N \to \infty$ is otherwise arbitrary fix $N \sim x^{1/2-3\iota/(2\kappa)}$. The bounds $P(|\epsilon_t| > \epsilon) = O(\epsilon^{-\kappa}) = o(\epsilon^{-\kappa+\iota})$ and $P(|\epsilon_t| > \epsilon|\Im_{t-1}) = o(\epsilon^{-\kappa+\iota})$ under C2 and C3, $E|\epsilon_t|^{\kappa-\iota} < \infty$, Lemma A.1, and $\beta_1||\epsilon_t||_\iota^{1/\kappa} \leq \tilde{\rho}$ imply

$$P\left(\beta_0\beta_1^i \prod_{j=0}^{i+1} \epsilon_{t-j}^2 > x^2/[2N]\right) = O\left(N^\kappa \tilde{\rho}^{2i}\right) x^{-(\kappa-\iota)/2} = O\left(\tilde{\rho}^{2i} x^{-\iota}\right)$$

hence $P(w_t^{(N)} > x^2/2) \leq O(\sum_{i=0}^N \tilde{\rho}^{2i} x^{-2\iota}) = O\left(x^{-\iota}\right)$.

**Step 2 ($y_t$):** Define $x_t = \sum_{i=0}^\infty \alpha^i u_{t-i}$ and $x_t^{(N)} = \sum_{i=0}^N \alpha^i u_{t-i}$. Repeating the proof of Lemma 4.1 it suffices to bound $P(|x_t^{(N)}| > x/2)$. Use subadditivity again to get

(25)

$$\begin{aligned} P\left(\left|x_t^{(N)}\right| > x\right) &\leq \sum_{i=0}^{N} P\left(\alpha^{2i}\epsilon_{t-i}^2 h_{t-i}^2 > \frac{x^2}{N^2}\right) \\ &\leq \sum_{i=0}^{N} \left\{P\left(\beta_0\alpha^{2i}\epsilon_{t-i}^2 > \frac{x^2}{2N^2}\right) + P\left(\beta_1\alpha^{2i}\epsilon_{t-i}^2 u_{t-i-1}^2 > \frac{x^2}{2N^2}\right)\right\} \end{aligned}$$

where

$$P\left(\beta_0\alpha^{2i}\epsilon_{t-i}^2 > x^2/\left[2N^2\right]\right) = O(N^\kappa |\alpha|^{\kappa i} x^{-\kappa}).$$

Fix $N \sim x^{\iota/\max\{2,\kappa\}}$ such that

$$\sum_{i=0}^{N} P\left(\beta_0\alpha^{2i}\epsilon_{t-i}^2 > \frac{x^2}{2N^2}\right) = O\left(\sum_{i=0}^{N} |\alpha|^{\kappa i} \times N^\kappa x^{-\kappa}\right) = O\left(x^{-(\kappa-\iota)}\right).$$

For the remaining probability use Step 1 and $N = x^{\iota/\max\{2,\kappa\}}$ to deduce

$$\begin{aligned} \sum_{i=0}^{N} P\left(\beta_1\alpha^{2i}\epsilon_{t-i}^2 u_{t-i-1}^2 > \frac{x^2}{2N^2}\right) &\leq K\sum_{i=0}^{N} \beta_1^{2i\iota}\alpha^{4i\iota} x^{-\iota} N^\iota \leq Kx^{-\iota}N^{1+\iota} \\ &= O\left(x^{-\iota+\left(\iota+\iota^2\right)/\max\{2,\kappa\}}\right) = O\left(x^{-\iota^2/2}\right). \end{aligned}$$

**Step 3** $(u_t,\ u_t u_{t-1}$ *and* $u_t y_t)$**:** The same logic used in Steps 1 and 2 suffices to prove $P(|u_t| > u) = O(u^{-\iota})$, $P(|u_t u_{t-1}| > u) = O(u^{-\iota})$ and $P(|u_t y_t| > x) = O(x^{-\iota})$.

**Step 4** $(m_t(\theta)$ : **GMM, QML):** Consider GMM with a reduced weight set:

$$m_t(\theta) = \left[\begin{array}{l} (y_t - \alpha y_{t-1}) \times y_{t-1} \\ \left\{(y_t - \alpha y_{t-1})^2 - \beta_0 - \beta_1 u_{t-1}^2\right\} \times \left[1, u_{t-1}^2\right]' \end{array}\right].$$

The following easily extends to more general weight structures, and QML is similar.

Consider $m_{1,t}^2(\theta_0) = u_t^2 y_{t-1}^2 = \beta_{0,0}\epsilon_t^2 y_{t-1}^2 + \beta_{0,1}\epsilon_t^2 u_{t-1}^2 y_{t-1}^2$. If $\beta_{0,1} = 0$ the result is a special case of Lemma 4.1, so consider $\beta_{0,1} > 0$. Subadditivity, the tail bounds C2 and C3, $P(|y_t| > y) = O(y^{-\iota})$ and $P(|u_t y_t| > x) = O(x^{-\iota})$ for arbitrarily small $0 < \iota < \kappa$ by Steps 2 and 3, and Lemma A.1 imply

$$\begin{aligned} P\left(|m_{1,t}(\theta_0)| > x\right) &\leq P\left(\beta_{0,0}\epsilon_t^2 y_{t-1}^2 > x^2/2\right) + P\left(\beta_{0,1}\epsilon_t^2 u_{t-1}^2 y_{t-1}^2 > x^2/2\right) \\ &= O\left(\beta_{0,0}^{\iota} E\left|\epsilon_t\right|^{\iota} x^{-\iota}\right) + O\left(\beta_{0,1}^{\iota} E\left|\epsilon_t\right|^{\iota} x^{-\iota}\right) = O\left(x^{-\iota}\right). \end{aligned}$$

Since $m_{2,t}(\theta_0) = u_t^2 - \beta_{0,0} - \beta_{0,1}u_{t-1}^2$, subadditivity and Step 3 imply $P(|m_{2,t}(\theta_0)| > x) = O(x^{-\iota})$.

By the same reasoning since $m_{3,t}(\theta_0) = u_t^2 u_{t-1}^2 - \beta_{0,0}u_{t-1}^2 - \beta_{0,1}u_{t-1}^4$ use subadditivity and Step 3 to deduce $P(|m_{3,t}(\theta_0)| > x) = O(x^{-\iota})$. ■

**Proof of Lemma 4.3.** All arguments follow the same logic used above: repetitions of subadditivity, the tail bound on $\epsilon_t$ and Lemma A.1 suffice to establish tail bounds for each $m_{i,t}(\theta_0)$. ■

**Proof of Lemma 4.5.** Under the stated conditions $E[u_t(\alpha_0)g_t^{(1)}(\theta_0)|\Im_{t-1}] = g_t^{(1)}(\theta_0)(E[y_t|\Im_{t-1}] - f(x_{t-1},\alpha_0)) = 0$ with probability one. Similarly, $E[\{u_t^2(\alpha_0) - h_t(\alpha_0,\beta_0)\} \times g_t^{(2)}(\theta_0)|\Im_{t-1}] = g_t^{(2)}(\theta_0) \times h_t^2(\alpha_0,\beta_0) \times \{E[\epsilon_t^2|\Im_{t-1}] - 1\} = 0$ with probability one given $\Im_{t-1}$-measurability of $g_t^{(i)}(\theta_0)$ and $h_t(\theta_0)$. ■

**Proof of Lemma 4.6.** The proof in the QML case is essentially identical to the proof of Lemma 4.5. ■

**Proof of Lemma 4.7.** In this case $m_t(\theta) = sgn(\epsilon_t^2 - 1) \times \epsilon_t^2 \times (\partial/\partial\theta)\ln h_t^2(\alpha,\beta)$, hence $E[m_t(\theta_0)|\Im_{t-1}] = 0$ *if and only if* $E[\epsilon_t^2 sgn(\epsilon_t^2 - 1)|\Im_{t-1}] = 0$. Use $E[\epsilon_t^2|\Im_{t-1}] = \int_0^\infty P(\epsilon_t^2 > u|\Im_{t-1})du = 1$ under C5 to deduce under C7

$$\begin{aligned} E\left[\epsilon_t^2 sgn\left(\epsilon_t^2 - 1\right)|\Im_{t-1}\right] &= E\left[\epsilon_t^2\left(I\left(\epsilon_t^2 \geq 1\right) - I\left(\epsilon_t^2 < 1\right)\right)|\Im_{t-1}\right] \\ &= \int_1^\infty P\left(\epsilon_t^2 > u|\Im_{t-1}\right)du - \int_0^1 P\left(\epsilon_t^2 > u|\Im_{t-1}\right)du \\ &= \int_0^\infty P\left(\epsilon_t^2 > u|\Im_{t-1}\right)du - 2\int_0^1 P\left(\epsilon_t^2 > u|\Im_{t-1}\right)du \\ &= 1 - 2\int_0^1 P\left(\epsilon_t^2 > u|\Im_{t-1}\right)du = 0. \end{aligned}$$

Therefore $E[m_t(\theta_0)|\Im_{t-1}] = 0$. ■

**Proof of Lemma 4.8.** Write $h_t'(\theta_0) := (\partial/\partial\theta) \ln h_t^2(\alpha, \beta)|_{\theta=\theta_0}$ so that $m_t(\theta_0) = h_t'(\theta_0) \times sgn(\ln \epsilon_t^2)$, and notice $sgn(\ln \epsilon_t^2) = I(\epsilon_t^2 > 1) - I(\epsilon_t^2 \leq 1)$. Then $E[m_t(\theta_0)|\Im_{t-1}] = h_t'(\theta_0) \times \{P(\epsilon_t^2 > 1|\Im_{t-1}) - P(\epsilon_t^2 \leq 1|\Im_{t-1})\} = 0$ with probability one under C6 and $\Im_{t-1}$-measurability of $h_t'(\theta_0)$. ■

**Proof of Theorem 5.1.** We claim $\text{argmin}_{\delta \in D}\{\hat{Q}(\hat{\theta}_n(\delta))\} \in D^*$ asymptotically with probability one such that $\hat{\delta}_n^*$ satisfies (20) as $n \rightarrow \infty$. We prove the claim by contradiction. Assume

$$\lim_{n\rightarrow\infty} P\left(\text{argmin}_{\delta\in D}\left\{\hat{Q}(\hat{\theta}_n(\delta))\right\} \in D/D^*\right) > 0, \tag{26}$$

In the symmetric case trimming is irrelevant: $D/D^* = \varnothing$, a contradiction of (22).

Consider the asymmetric case.

####

■

# Appendix 2: Supporting Lemmata

**LEMMA A.1** *Consider two random variables $y_t$ and $z_t$ with tail bounds $P(|y_t| > y) = O(y^{-\kappa_y})$ and $P(|z_t| > z) = O(z^{-\kappa_z})$. In particular $P(|y_t| > y|z_t) = O(y^{-\kappa_y})$ and $E|z_t|^{\kappa_z - \iota} < \infty$. Then $P(|y_t z_t| > x) = o(x^{-\delta})$ for some $\delta < \min\{\kappa_y, \kappa_z\}$.*

**Proof.** A variation on Breiman's (1965) argument suffices. For any $\delta < \min\{\kappa_y, \kappa_z\}$ trivially $P(|y_t| > y|z_t) = O(y^{-\kappa_y}) = o(y^{-\delta})$. By bounded convergence and $E|z_t|^\delta < \infty$

$$P\left(|y_t z_t| > x\right) = E\left[P\left(|y_t| > x/z_t|z_t\right)\right] = o\left(x^{-\delta} E\left|z_t\right|^\delta\right) = o\left(x^{-\delta}\right).$$

■

**REFERENCES**


[1] An H.Z., and F.C. Huang (1996). The Geometrical Ergodicity of Nonlinear Autoregressive Models, Statistica Sinica, 6, 943-956.

[2] Agulló, J., C. Croux, S. Van Aelst (2008). The Multivariate Least-Trimmed Squares Estimator, Journal of Multivariate Analysis 99, 311-338.

[3] Andrews, D.W.K. (1984). Non-Strong Mixing Autoregressive Processes, Journal of Applied Probability 21, 930-934.

[4] Basrak, B., R.A. Davis, and T, Mikosch (2002). Regular Variation of GARCH Processes, Stochastic Processes and Their Applications 12, 908-920.

[5] Bassett, G.W. (1991). Equivariant, Monotonic, 50% Breakdown Estimators, American Statistician 45, 135–137.

[6] Breiman, L. (1965). On Some Limit Theorems Similar to the Arc-Sine Law, Theory of Probability and its Applications 10, 351–360.

[7] Carrasco and Chen (2002). Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models, Econometric Theory 18, 17-39.
[8] Chan, N.H., Peng, L., 2005. Weighted Least Absolute Deviations Estimation for an AR(1) Process with ARCH(1) Errors, Biometrika 92, 477–484.
[9] Chen, L.-A., A. H. Welsh and W. Chan (2001). Estimators for the Linear Regression Model Based on Winsorized Observations, Statistica Sinica, 11, 147-172.
[10] Čižek, P. (2006). Least Trimmed Squares in Nonlinear Regression Under Dependence, Journal of Statistical Planning and Inference 136, 3967–3988.
[11] Čižek, P. (2008). General Trimmed Estimation: Robust Approach to Nonlinear and Limited Dependent Variable Models, Econometric Theory: *forthcoming*.
[12] Davidson, J. (1994). Stochastic Limit Theory. Oxford Univ. Press: Oxford.
[13] Davis, R.A., K. Knight and J. Liu (1992). M-estimation for Autoregressions with Infinite Variance, Stochastic Processes and Their Applications 40, 145-180.
[14] Embrechts, P., Klüppelberg, C. and Mikosch, T. (2003). Modelling Extremal Events for Insurance and Finance. Springer-Verlag: Frankfurt.
[15] Engle, R. amd V. Ng (1993). Measuring and Testing the Impact of News On Volatility. Journal of Finance 48, 1749-1778.
[16] Engle, R., D. Nelson and T. Bollerslev (1994). ARCH Models, in R. Engle and D. McFadden (eds.), Handbook of Econometrics, Volume IV, pp. 2959-3038. North Holland: Amsterdam.
[17] Finkenstadt, B., and H. Rootzén (2003). Extreme Values in Finance, Telecommunications and the Environment. Chapman and Hall: New York.
[18] Francq, C. and J.M. Zakoïan (2004). Maximum Likelihood Estimation of Pure GARCH and ARMA-GARCH Processes. Bernoulli 10, 605–637.
[19] Giraitis, L. and P.M. Robinson (2000). Whittle Estimation of ARCH Models, Econometric Theory 17, 608–631.
[20] Goldie, C.M., and R.L. Smith (1987). Slow Variation with Remainder: Theory and Applications, Quarterly Journal of Mathematics 38, 45-71.
[21] Gorodetskii, V. V. (1977). On the Strong Mixing Property for Linear Sequences, Theory of Probability and its Applications 22, 411-413.
[22] Hall, A. R. (2005). Generalized Method of Moments. Oxford University Press: Oxford.
[23] Hall, P. and Q. Yao (2003). Inference in ARCH and GARCH Models with Heavy-Tailed Errors, Econometrica 71, 285–317.
[24] Hansen, L.P. (1982). Large Sample Properties of Generalized Method of Moments Estimators , Econometrica 50, 1029-1054.
[25] Hansen, B. and S.W. Lee (1994). Asymptotic Theory for the Garch (1,1) Quasi-Maximum Likelihood Estimator, Econometric Theory 10, 29-52.
[26] Hill, J.B. (2005). On Tail Index Estimation for Dependent, Heterogeneous Data, Dept. of Economics, University of North Carolina - Chapel Hill. Revised for *Econometric Theory*. Available at www.unc.edu/~jbhill/fclt_tail_array.pdf.
[27] Hill, J.B. (2007). Functional Central Limit Theorems for Dependent, Heterogeneous Arrays with Applications, Dept. of Economics, University of North Carolina - Chapel Hill. Revised for *Journal of Statistical Planning and Inference*. Available at www.unc.edu/~jbhill/fclt_tail_array.pdf.

[28] Hill, J.B. (2008). Tail and Non-Tail Memory with Applications to Extreme Value and Robust Statistics, Dept. of Economics, University of North Carolina - Chapel Hill. Submitted. Available at www.unc.edu/∼jbhill/tail_nontail_garch.pdf.
[29] Hill, J.B. and E. Renault (2008a). Tail-Trimmed Generalized Method of Moments, Dept. of Economics, University of North Carolina - Chapel Hill.
[30] Hill, J.B. and E. Renault (2008b). On Minimum Distance Estimation for Non-Differentiable Criteria under Moment Condition Failure, Dept. of Economics, University of North Carolina - Chapel Hill.
[31] Hill, J.B. and E. Renault (2008c). Central Limit Theory for Tail-Trimmed Sums of Dependent, Heterogeneous Data, with an Application to Robust Least Squares, Dept. of Economics, University of North Carolina - Chapel Hill.
[32] Hsing, T. (1991). On Tail Index Estimation Using Dependent Data, Annals of Statistics 19, 1547-1569.
[33] Huber, P. (1967). The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions, in L.M. LeCam and J. Neyman, eds., Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press.
[34] Huber, P. J. (1977) Robust Statistical Procedures. Philadelphia: Society for Industrial and Applied Mathematics.
[35] Ibragimov, I. A. and Y. V. Linnik (1971). Independent and Stationary Sequences of Random Variables. Wolters-Noordhof: Berlin.
[36] Jensen, S.T. and A. Rahbek (2004). Asymptotic Normality of the QML Estimator of ARCH in the Nonstationary Case, Econometrica 72, 641-646.
[37] Leadbetter, M.R., G. Lindgren and H. Rootzén (1983). Extremes and Related Properties of Random Sequences and Processes. Springer-Verlag: New York.
[38] Lee, S.-W. and B.E. Hansen (1994). Asymptotic Theory for the GARCH (1,1) Quasi-Maximum Likelihood Estimator. Econometric Theory, 10, 29-52.
[39] Leibscher, E. (2005). Towards a Unified Approach for Proving Geometric Ergodicity and Mixing Properties of Nonlinear Autoregressive Processes, Journal of Time Series Analysis 26, 669-689.
[40] Ling, S. (2005). Self-Weighted Least Absolute Deviation Estimation for Infinite Variance Autoregressive Models, Journal of the Royal Statistical Society Series B 67, 381–393.
[41] Ling, S. (2007). Self-Weighted and Local Quasi-Maximum Likelihood Estimators for ARMA-GARCH/IGARCH Models, Journal of Econometrics 150, 849-873.
[42] Linton, O., J. Pan and H. Wang (2008). Estimation for a Non-stationary Semi-Strong GARCH(1,1) Model with Heavy-Tailed Errors, mimeo, London School of Economics.
[43] Lumsdaine, R. (1996). Consistency and Asymptotic Normality of the Quasi-Maximum Likelihood Estimator in IGARCH(1,1) and Covariance Stationary GARCH(1,1) Models, Econometrica 64, 575-596.
[44] Mandelbrot, B. (1963). The variation of certain speculative prices. Journal of Business 36, 394-419.
[45] McFadden D.L. (1989). A Method of Simulated Moments for Estimation of Multinomial Discrete Response Models Without Numerical Integration, Econometrica 57, 995-1026.
[46] McLeish, D.L. (1975). A Maximal Inequality and Dependent Strong Laws, Annals of Probability 3, 829-839.

[47] Meitz M. and P. Saikkonen (2008). Stability of Nonlinear AR-GARCH Models, Journal of Time Series Analysis 29, 453-475.
[48] Mikosch, T. and D. Straumann (2002). Whittle Estimation in a Heavy-Tailed GARCH(1,1) Model, Stochastic Processes and Their Applications 100, 187-222.
[49] Mittnik, S. and S.T. Rachev (2000). Stable Paretian Models in Finance. New York: Wiley.
[50] Newey, W.K., and D.L. McFadden (1994). Large Sample Estimation and Hypothesis Testing, in: R. Engle and D. McFadden, eds., Handbook of Econometrics, Vol. 4. Amsterdam: North-Holland.
[51] Pakes, A. and D. Pollard (1989). Simulation and the Asymptotics of Optimization Estimators, Econometrica, 57, 1027-1057.
[52] Peng, L. and Q. Yao (2003). Least Absolute Deviations Estimation for ARCH and GARCH Models, Biometrika 90, 967-975.
[53] Pollard, D. (1991). Asymptotics for Least Absolute Deviation Regression Estimators, Econometric Theory 7, 186-199.
[54] Resnick, S. (1987). Extreme Values, Regular Variation and Point Processes. Springer-Verlag: New York.
[55] Rubert, D. and J. Carroll (1980). Trimmed least squares estimation in the linear model, Journal of the American Statistical Association 75, 828-838..
[56] Tong, H. (1990). Non-linear Time Series: A Dynamic System Approach. Oxford University Press: Oxford.
[57] Welsh, A.H. (1987). The Trimmed Mean in the Linear Model, Annals of Statistics 15, 20-36.
[58] Whittle, P., 1953. Estimation and Information in Stationary Time Series, Arkiv für Matematik 2, 423–434.
[59] Wu, W. B. and W. Min (2005). On Linear Processes with Dependent Innovations, Stochastic Processess and Their Applications 115 939–958.