THE UNIVERSITY OF SYDNEY

**The University of Sydney Business School**
**The University of Sydney**

BUSINESS ANALYTICS WORKING PAPER SERIES

# GEL Estimation for Heavy-Tailed GARCH Models with Robust Empirical Likelihood Inference

Jonathan B. Hill
University of North Carolina Chapel Hill

Artem Prokhorov
University of Sydney

September 10, 2015

We construct a Generalized Empirical Likelihood estimator for a GARCH(1,1) model with a possibly heavy tailed error. The estimator imbeds tail-trimmed estimating equations allowing for over-identifying conditions, asymptotic normality, efficiency and empirical likelihood based confidence regions for very heavy-tailed random volatility data. We show the implied probabilities from the tail-trimmed Continuously Updated Estimator elevate weight for usable large values, assign large but not maximum weight to extreme observations, and give the lowest weight to non-leverage points. We derive a higher order expansion for GEL with imbedded tail-trimming (GELITT), which reveals higher order bias and efficiency properties, available when the GARCH error has a finite second moment. Higher order asymptotics for GEL without tail-trimming requires the error to have moments of substantially higher order. We use first order asymptotics and higher order bias to justify the choice of the number of trimmed observations in any given sample. We also present robust versions of Generalized Empirical Likelihood Ratio, Wald, and Lagrange Multiplier tests, and an efficient and heavy tail robust moment estimator with an application to expected shortfall estimation. Finally, we present a broad simulation study for GEL and GELITT, and demonstrate profile weighted expected shortfall for the Russian Ruble - US Dollar exchange rate. We show that tail-trimmed CUE-GMM dominates other estimators in terms of bias, mse and approximate normality.

Key words and phrases: GEL, GARCH, tail trimming, heavy tails, robust inference, efficient moment estimation, expected shortfall, Russian Ruble.
AMS classifications : 62M10 , 62F35.
JEL classifications : C13 , C49.

BA Working Paper No: BAWP-2015-03
http://sydney.edu.au/business/business_analytics/research/working_papers

# GEL Estimation for Heavy-Tailed GARCH Models with Robust Empirical Likelihood Inference

**Jonathan B. Hill**[*]
University of North Carolina – Chapel Hill

**Artem Prokhorov**[†]
University of Sydney

September 10, 2015

**Abstract**

We construct a Generalized Empirical Likelihood estimator for a GARCH(1,1) model with a possibly heavy tailed error. The estimator imbeds tail-trimmed estimating equations allowing for over-identifying conditions, asymptotic normality, efficiency and empirical likelihood based confidence regions for very heavy-tailed random volatility data. We show the implied probabilities from the tail-trimmed Continuously Updated Estimator elevate weight for usable large values, assign large but not maximum weight to extreme observations, and give the lowest weight to non-leverage points. We derive a higher order expansion for GEL with imbedded tail-trimming (GELITT), which reveals higher order bias and efficiency properties, available when the GARCH error has a finite second moment. Higher order asymptotics for GEL without tail-trimming requires the error to have moments of substantially higher order. We use first order asymptotics and higher order bias to justify the choice of the number of trimmed observations in any given sample. We also present robust versions of Generalized Empirical Likelihood Ratio, Wald, and Lagrange Multiplier tests, and an efficient and heavy tail robust moment estimator with an application to *expected shortfall* estimation. Finally, we present a broad simulation study for GEL and GELITT, and demonstrate profile weighted expected shortfall for the Russian Ruble - US Dollar exchange rate. We show that tail-trimmed CUE-GMM dominates other estimators in terms of bias, mse and approximate normality.

*Key words and phrases*: GEL, GARCH, tail trimming, heavy tails, robust inference, efficient moment estimation, expected shortfall, Russian Ruble.

*AMS classifications* : 62M10 , 62F35.

*JEL classifications* : C13 , C49.

[*]Corresponding author. Dept. of Economics, University of North Carolina, Chapel Hill, NC; http://www.unc.edu/∼jbhill; jbhill@email.unc.edu

[†]The University of Sydney Business School and St.Petersburg State University; http://sydney.edu.au/business/staff/artemp; artem.prokhorov@sydney.edu.au.

# 1 Introduction

We develop a Generalized Empirical Likelihood estimator for a potentially very heavy tailed GARCH(1,1) process by tail-trimming estimating equations. The setting is motivated by recent intense interest in information theoretic methods (Smith, 1997; Imbens, 1997; Kitamura, 1997; Antoine, Bonnal, and Renault, 2007), including the higher order properties of GEL estimators (Newey and Smith, 2004; Anatolyev, 2005), coupled with empirical evidence that the distributions of many financial returns have very heavy tails (e.g. Embrechts, Kluppleberg, and Mikosch, 1997; Wagner and Marsh, 2005; Ibragimov, 2009; Hill, 2015b) and exhibit volatility clustering (Bollerslev, 1986).

The time series of interest is a stationary ergodic scalar process $\{y_t\}$ with increasing $\sigma$-fields $\Im_t \equiv \sigma(\{y_\tau\} : \tau \leq t)$ and a strong-GARCH(1,1) representation

$$y_t = \sigma_t \epsilon_t \text{ where } \epsilon_t \text{ is iid, } E[\epsilon_t] = 0 \text{ and } E[\epsilon_t^2] = 1 \tag{1}$$

$$\sigma_t^2 = \omega^0 + \alpha^0 y_{t-1}^2 + \beta^0 \sigma_{t-1}^2, \text{ where } \omega^0 > 0,\ \alpha^0, \beta^0 \geq 0, \text{ and } \alpha^0 + \beta^0 > 0.$$

The assumption $\alpha^0 + \beta^0 > 0$ safeguards against well known estimation boundary problems, although allowing $\alpha^0 = 0$ and/or $\beta^0 = 0$ merely requires an additional functional limit theory (Andrews, 1999; Francq and Zakoïan, 2004). Assume $\Theta$ is a compact subset of points $\theta = [\omega, \alpha, \beta]'$ that contains $\theta^0$ as an interior point, and the stationarity and ergodicity condition $E[\ln(\alpha + \beta\epsilon_t^2)] < \infty$ holds (Nelson, 1990; Bougerol and Picard, 1992):

$$\Theta \subseteq \left\{\theta \in (0,\infty) \times (0,1) \times (0,1) : E\left[\ln\left(\alpha + \beta\epsilon_t^2\right)\right] < \infty\right\}. \tag{2}$$

We work with a linear strong-GARCH model solely to focus ideas and to motivate the use of tail-trimming to deliver a robust GEL estimator. An extension of our methods to higher order GARCH processes is trivial. In order to include a model of the conditional mean, however, a more nuanced trimming approach is required since the relevant QML estimating equations may have heavy tailed iterative terms which impact the resulting Jacobian in a more complicated way. See Appendix B for a brief discussion concerning an ARMA-GARCH model.[1] Our asymptotic

[1]We show how to construct trimmed estimating equations, and note that no additional moment conditions on $y_t$ are required. Francq and Zakoïan (2004, Theorem 3.2), however, show that the QML estimator requires $y_t$ itself to have a finite fourth moment, a tremendous requirement in practice since many finanncial time series show evidence of heavy tails (for evidence and further references, see Ibragimov, 2009; Aguilar and Hill, 2015; Hill, 2015b).

theory relies heavily on uniform asymptotics for stationary mixing data,[2] hence whether our required results extend to non-stationary cases is not yet known.[3]

The iid assumption for $\epsilon_t$ implies our trimmed QML-type estimating equations are martingale differences. This simplifies estimation since smoothing is not required (cf. Owen, 1990, 1991; Kitamura, 1997; Kitamura and Stutzer, 1997), and this leads to sharp details concerning how the implied probabilities relate information about usable sample extremes. Furthermore, the iid assumption allows us to explicitly show how higher order bias is reduced by reducing trimming. We can easily allow for weakly dependent errors by smoothing the estimating equations, but the cost is far fewer details about how the smoothed implied probabilities translate information about extremes, and essentially no information about how trimming impacts higher order bias.[4] Since the latter two are key contributions in this paper, we simply focus on iid errors.

Construct volatility and error functions

$$\sigma_t^2(\theta) = \omega + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2(\theta) \text{ and } \epsilon_t(\theta) = y_t/\sigma_t(\theta) \text{ where } \theta = [\omega, \alpha, \beta]' \in \mathbb{R}^3,$$

and let $m_t(\theta)$ denote estimating equations based on $\{y_t, \sigma_t(\theta)\}$, a stochastic mapping $m_t : \Theta \to \mathbb{R}^q$ with $q \geq 3$ that satisfies the global identification condition

$$E\left[m_t(\theta)\right] = 0 \textit{ if and only if } \theta = \theta^0 \text{ for unique } \theta^0 \text{ in compact } \Theta \subset \mathbb{R}^3.$$

In Section 2 we note that $\sigma_t^2(\theta)$ is not observed, and utilize an iterated approximation.

We consider equations $m_t(\theta) \in \mathbb{R}^q$, $q \geq 3$, based on QML score equations, with added over-identifying restrictions based on stochastic weights $w_t(\theta) \in \mathbb{R}^{q-3}$. Hence, we use:

$$m_t(\theta) = \left(\epsilon_t^2(\theta) - 1\right) \times x_t(\theta) \in \mathbb{R}^q,\ q \geq 3, \text{ where } x_t(\theta) \equiv [s_t'(\theta), w_t'(\theta)]' \text{ and } s_t(\theta) \equiv \frac{1}{\sigma_t^2(\theta)} \frac{\partial}{\partial\theta} \sigma_t^2(\theta).$$

[2]See the proof of Lemma A.5 in the technical appendix Hill and Prokhorov (2014). This result is crucial for showing the estimating equations $\{\hat{m}_{n,t}^*(\theta), m_{n,t}^*(\theta)\}$, defined below, satisfy $\sup_{\theta\in\Theta} ||n^{-1/2}\Sigma_n^{-1/2}(\theta) \sum_{t=1}^n \{\hat{m}_{n,t}^*(\theta) - m_{n,t}^*(\theta)\}|| = o_p(1)$, while a uniform limit is required since the tail-trimmed estimating equations are nonlinear functions of $\theta$. See especially the proof of Theorem 5.2 in Appendix A.4.

[3]Some uniform limit theory for QML score components in the nonstationary GARCH case is presented in Jensen and Rahbek (2004b, Lemma 5) and Linton, Pan, and Wang (2010, Lemma 5). These arguments, however, do not cover our required property $\sup_{\theta\in\Theta}\{1/n^{1/2}|\sum_{t=1}^n (s_{i,t}(\theta) - E\left[s_{i,t}(\theta)\right])|\} = O_p(1)$, where $s_t(\theta) \equiv (\partial/\partial\theta) \ln \sigma_t^2(\theta)$ and $\sigma_t^2(\theta) = \omega + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2(\theta)$. We use a uniform limit theory in Doukhan, Massart, and Rio (1995) for stationary mixing data to prove the required results.

[4]This follows since higher order bias is a function of higher moments of tail-trimmed partial sums. These moments are simple functions of trimming fractiles only in the case of iid errors, and otherwise we are limited to deducing bounds for these moments (see, e.g. Hill, 2012, 2015a,b) which do not illuminate how trimming impacts higher order bias.

Implicitly if $q = 3$ then $x_t(\theta) = s_t(\theta)$, while $q > 3$ aligns with over-identifying restrictions $E[(\epsilon_t^2 - 1)w_{i,t}] = 0$ for $i = 1, ..., q-3$. We assume $w_t(\theta)$ is $\Im_{t-1}$-measurable, continuous and differentiable. Identification $E[(\epsilon_t^2 - 1)x_t] = 0$ and $E[\epsilon_t^2] = 1$ imply $x_t$ must be integrable, while $s_t$ is square integrable when $\alpha^0 + \beta^0 > 0$ (Francq and Zakoïan, 2004), hence we assume $w_t$ is integrable. Instrument classes other than QML-equations are obviously possible (cf. Skoglund, 2010). The use of QML-equations is known to result in an efficient (exactly identified) GMM estimator in the sense of Godambe (1985), cf. Li and Turtle (2000). Further, since the instrument $s_t$ is square integrable, if $x_t$ contains only lags of $s_t$ then heavy tail challenges arise solely due to the error $\epsilon_t$.

Several recent papers consider properties of QML and LAD estimators of GARCH under heavy tailed errors. Hall and Yao (2003) derive the QML estimator limit distribution for linear GARCH when $\epsilon_t$ belongs to a domain of attraction of stable law with tail exponent $\kappa \in [2, 4]$. They show that the convergence rate is $n^{1-2/\kappa}/L(n)$ for slowly varying[5] $L(n) \to \infty$, where $n^{1-2/\kappa}/L(n) < n^{1/2}$ for any $\kappa \in [2, 4]$. See also Berkes and Horvath (2004) for consistency results. Although QML for GARCH is robust to heavy tails in possibly non-stationary $y_t$, as long as $\epsilon_t$ has a finite fourth moment, in small samples it is known to exhibit bias (e.g. Lumsdaine, 1995; Gonzalez-Rivera and Drost, 1999; Berkes and Horvath, 2004; Jensen and Rahbek, 2004a).

A finite variance $E[\epsilon_t^2] < \infty$ appears indispensable for obtaining an asymptotically normal estimator. Linton, Pan, and Wang (2010) prove $\sqrt{n}$-convergence and asymptotic normality of the log-LAD estimator $\arg\min_{\theta\in\Theta} \sum_{t=1}^n |\ln y_t^2 - \ln \sigma_t^2(\theta)|$ for non-stationary GARCH provided $\epsilon_t$ has a zero median. See also Peng and Yao (2003) for earlier work with iid errors. Zhu and Ling (2011) show the weighted Laplace QML estimator is $\sqrt{n}$-convergent and asymptotically normal if $\epsilon_t$ has a zero median and $E|\epsilon_t| = 1$. They only require $E[\epsilon_t^2] < \infty$, but in practice GARCH models are typically used under the assumption $E[\epsilon_t^2] = 1$ irrespective of the estimator chosen. The classic assumption $E[\epsilon_t^2] = 1$ coupled with $E|\epsilon_t| = 1$ seems to severely limit the available distributions for $\epsilon_t$. Berkes and Horvath (2004) tackle non-Gaussian QML which for identification requires moment conditions either beyond, or in place of, the traditional $E[\epsilon_t] = 0$ and $E[\epsilon_t^2] = 1$. Thus, in general these estimators are not technically for Bollerslev's (1986) seminal GARCH model (1) in which independence and $E[\epsilon_t^2] = 1$ imply identically $\sigma_t^2 = E[y_t^2|\Im_{t-1}]$, and they naturally do not allow for over-identifying restrictions.

Hill (2015a) uses a variety of trimming and weighting techniques for QML and method of moments estimators for heavy tailed GARCH. However, over-identifying restrictions are not allowed, profiles weights are not developed and therefore efficient moment estimators are not treated, and the empirical likelihood method for inference is not considered. See also Hill (2013)

[5]Recall slowly varying $L(n)$ satisfies $L(\xi n)/L(n) \to 1$ as $n \to \infty$ for all $\xi > 0$.

for a related least squares theory for autoregressions. Notice, though, that moment conditions not used for estimation can always be tested using heavy tail robust methods (Hill and Aguilar, 2013), while a large variety of model specification tests can be rendered heavy tail robust (Hill, 2012; Hill and Aguilar, 2013; Aguilar and Hill, 2015). Moreover, higher order asymptotics have evidently never been used for determining a reasonable negligible trimming strategy.

The present paper extends the line of heavy tail robust estimation and inference in Hill and Aguilar (2013), Aguilar and Hill (2015) and Hill (2012, 2013, 2015a,b) to a GEL framework and to the empirical likelihood method. As in those papers we apply a heavy tail robust, but negligible, data transform to the estimating equations. We allow over identifying restrictions with one-step estimation and inference that leads to Gaussian asymptotics by exploiting tail-trimming. GMM and GEL allow for over-identifying restrictions whereas the M-estimators developed in Hill (2013, 2015a) naturally do not. Over-identifying restrictions can reveal exploitable information about the data generating process, an idea dating at least to Owen (1990, 1991) and Qin and Lawless (1994), cf. Antoine, Bonnal, and Renault (2007). The classic example is IV estimation (see, e.g., Guggenberger and Smith, 2008). Indeed, in the GARCH model, moment conditions tie model parameters to the unconditional variance when it exists, an idea exploited in the variance targeting literature (cf. Engle and Mezrich, 1996; Hill and Renault, 2012) and for iid data stated in Qin and Lawless (1994, Example 1). As another example, model parameters identify the tail index by a moment condition (see Basrak, Davis, and Mikosch, 2002, e.g.).

The empirical likelihood method has the great advantage of allowing inference without covariance matrix estimation by inverting the likelihood function (Owen, 1990). See Section 2 for development of the infeasible and feasible estimators, and characterization of the rate of convergence. Standard and profile-weighted moment estimators are treated in Section 5, and are used for heavy tail robust (and efficient) score, Lagrange Multiplier, and Likelihood Ratio tests. Such tests can be used as heavy tail robust model specification tests, including GARCH order or the presence of GARCH effects, so they can be used as model selection tools.[6] However, testing when a parameter value is on the boundary of the maintained hypotheses leads to non-standard asymptotics (Andrews, 2001).

In Section 3 we show that the implied probabilities derived from the tail-trimmed Continuously Updated Estimator, which are especially tractable, differentiate between usable large values (i.e. values near the trimming threshold) and damaging extremes that are trimmed for estimation. Large values serve as leverage points and accelerate convergence rates, yet very large values impede normality and are therefore trimmed. Thus, extremes receive elevated weight, but

[6]We thank a referee for pointing out this possibility to us.

*near-extremes that are not trimmed receive the most weight*. We use the implied probabilities from tail-trimmed GEL to perform heavy tail robust and efficient tests of over-identification. Similar test statistics, without trimming, have been considered by Kitamura and Stutzer (1997), Newey and Smith (2004), and Smith (2011) amongst others.

In Section 4 we derive a higher order expansion for our estimator along the lines of Newey and Smith (2004, Sections 3 and 4). In the case of GARCH model estimation with QML-type estimating equations, GEL requires $E[\epsilon_t^6] < \infty$ for a second order expansion (necessary for bias) and $E[\epsilon_t^{10}] < \infty$ for a third order expansion, while GELITT always only needs $E[\epsilon_t^2] < \infty$ for any higher order expansion. GELITT bias decomposes into bias due to the GEL structure (when higher moments exist) and bias due to trimming. This is irrelevant for bias-correction since a composite bias estimator as in Newey and Smith (2004, Section 5) removes higher order GELITT bias whether due to the GEL form or trimming. Moreover, it does not require extreme value theory and therefore tail index estimation as in Hill (2015b).

We also show that under mild assumptions (higher order) bias is always small if few observations are trimmed, and monotonically smaller in the case of EL or exact identification. By first order asymptotics the rate of convergence is higher if the rate of trimming is nearly the sample size $n$, a feature common to M-estimators for GARCH models with negligible trimming, and to mean estimation, cf. Hill (2012, 2015b,a). Thus, trimming at a rate nearly equal to $\zeta n$, e.g. $\zeta n/\ln(n)$, is optimal as long as a small $\zeta$ is used. The usefulness of this combination is revealed by simulation in Section 8, and elsewhere (Aguilar and Hill, 2015; Hill, 2012, 2013, 2015b,a; Hill and Aguilar, 2013). Together, the use of higher order asymptotics to minimize and estimate bias marks a sharp improvement over existing tail-trimming methods for M-estimators (Hill, 2013, 2015b,a). In that literature, only first order asymptotics exist which, as in the present paper, invariably points toward elevating trimming by errors, but says little about the implications for trimming on bias.

We then use the probability profiles in Section 6 for tail-trimmed moment estimation which is shown to have the same efficiency property as without trimming. We generalized theory developed in Smith (2011) for GEL estimators to the heavy tail case, while Smith (2011) extends theory in Back and Brown (1993) and Brown and Newey (1998). As an example, in Section 7 we use the profiles for efficient and heavy tail robust estimation of a conditionally heteroscedastic asset's *expected shortfall*. We derive the limit distribution of a bias-corrected profile weighted tail-trimmed estimator, making a more efficient version of Hill's (2015b) robust estimator. Further, we improve on Hill's (2015b) proposed strategy for optimally estimating bias, and derive the appropriate limit theory.

A simulation study follows in Section 8. This is unique in the literature since the merit of

GEL estimators (untrimmed or trimmed) have not been thoroughly studied for GARCH model estimation.[7] We use EL, CUE and ET criteria, with and without trimming, and for trimming we use our higher order bias minimization theory for selecting the trimming fractile. Tail-trimmed CUE performs best overall in terms of bias, mse, and approximate normality, evidently due to the easily solved quadratic criterion and the fact that trimming a few errors per sample improves sampling properties. This is a useful result that may be of independent interest since EL with or without trimming has lower higher order bias in theory. That theory, however, does not account for substantial computational differences across GEL estimators, giving substantial credence to the argument for simplicity in Bonnal and Renault (2004) and Antoine, Bonnal, and Renault (2007). It also further demonstrates that trimming very few observations can have a strong positive impact on estimator performance, as shown also in Hill (2013, 2015a).

Finally, we perform a small scale empirical study based on financial returns in order to demonstrate our GEL estimator, and our robust, efficient and bias-improved estimator of the expected shortfall. We leave concluding remarks for Section 10.

The theory of GEL to date is designed for sufficiently thin tailed equations such that asymptotic normality is assured. See Qin and Lawless (1994), Hansen, Heaton, and Yaron (1996), Imbens (1997), Kitamura (1997), Kitamura and Stutzer (1997), Imbens, Spady, and Johnson (1998), Smith (1997, 2011), Newey and Smith (2004), and Antoine, Bonnal, and Renault (2007) for early contributions and broad theory developments. In a GARCH framework with QML-type equations and only lags of $s_t$ as instruments, we need $E[\epsilon_t^4] < \infty$ (cf. Francq and Zakoïan, 2004), but a far more restrictive moment condition is needed if least squares-type equations are used (see Francq and Zakoïan, 2000). Moreover, as discussed above, a higher order asymptotic expansion for GEL estimators of GARCH models with QML-type equations require prohibitive moment conditions, up to $E[\epsilon_t^{10}] < \infty$ for a third order expansion. Nevertheless, GEL estimators have beneficial properties: asymptotic bias of GEL does not grow with the number of estimating equations, contrary to GMM in well known cases, while bias-corrected EL is higher order asymptotically efficient (see Newey and Smith, 2004; Anatolyev, 2005). The higher order properties arise from different first order conditions for different GEL criteria, while first order asymptotics, including efficiency, are insensitive to the criteria, whether there is weak identification or not (cf. Newey and Smith, 2004; Guggenberger and Smith, 2008). We show that GELITT obtains the same type of higher order expansion as GEL, without the requirement of higher moments. Hence, the higher order bias and efficiency properties of GEL extend to GELITT under far less stringent conditions.

[7] Chan and Ling (2006) develope EL theory for AR-GARCH models, but only study a unit root test, and otherwise we are not familiar with other published simulation studies of GEL for GARCH.

Empirical likelihood for heavy tail robustness and for GARCH has limited use to date. Peng (2004) uses the empirical likelihood method for heavy tail robust confidence bands of the mean, and other than a similar use for tail parameter inference (Worms and Worms, 2011) there do not appear to be any other extensions to robust estimation. Chan and Ling (2006) develop empirical likelihood for GARCH and random walk-GARCH, where $E[\epsilon_t^4] < \infty$ and $\alpha^0 + \beta^0 < 1$, both unrealistic restrictions for many financial time series. Further, they only study a unit root test by simulation and therefore do not report GEL estimator properties for GARCH. Two-step GMM estimation for GARCH is treated in Skoglund (2010), amongst others.

We use the following notation. The $L_p$-norm for a matrix $A \equiv [A_{i,j}]$ is $||A||_p \equiv (\sum_{i,j} E|A_{i,j}|^p)^{1/p}$. The spectral norm is $||A|| = (\lambda_{\max}(A'A))^{1/2}$ where $\lambda_{\max}$ is the maximum eigenvalue. $K > 0$ is a finite constant whose value may change; $\iota, \delta > 0$ are tiny constants; and $N$ is a positive integer. $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and in distribution. $\to$ denotes convergence in $||\cdot||$. $a_n \sim b_n$ implies $a_n/b_n \to 1$ as $n \to \infty$. $I_d$ is a $d$-dimensional identity matrix. $L(n) \to \infty$ is a slowly varying function whose value or rate may change from line to line. An *intermediate order sequence* $\{k_n\}$ satisfies $k_n \in \{1, ..., n-1\}$, and $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$.

# 2 GEL with Tail-Trimming

We initially work with the unobserved process $\{\sigma_t^2(\theta)\}$ and derive an infeasible estimator of $\theta^0$. We then derive parallel results for the feasible estimator based on an iterated approximation to $\sigma_t^2(\theta)$. Drop $\theta^0$ throughout, e.g. $\sigma_t^2 = \sigma_t^2(\theta^0)$, $x_t = x_t(\theta^0)$.

## 2.1 Tail-Trimmed Equations

Our first task is to trim the equations $m_{i,t}(\theta)$ when they obtain an extreme value. Hill and Renault (2010) use $m_{i,t}(\theta)$ itself to gauge when an extreme value occurs. Since $m_t$ may be asymmetric this requires asymmetric trimming which in general induces small sample bias. In the present setting by a standard first order expansion we know asymptotics depend solely on $\epsilon_t(\theta)$ and $x_t(\theta)$. However, $s_t(\theta) = (\partial/\partial\theta) \ln \sigma_t^2(\theta)$ has an $L_2$-bounded envelope $\sup_{\theta \in \mathcal{N}_0} |s_{i,t}(\theta)|$ on some compact subset $\mathcal{N}_0 \subseteq \Theta$ containing $\theta^0$ (cf. Francq and Zakoïan, 2004), hence only $\epsilon_t(\theta)$ and the added weights $w_t(\theta)$ in $x_t(\theta)$ can be sources of extremes in $m_t(\theta)$. We therefore trim by these components separately.

Let $z_t(\theta)$ denote $\epsilon_t(\theta)$ or $w_{i,t}(\theta)$, and define the two-tailed process and its order statistics:

$$z_t^{(a)}(\theta) \equiv |z_t(\theta)| \quad \textit{and} \quad z_{(1)}^{(a)}(\theta) \geq \cdots \geq z_{(n)}^{(a)}(\theta) \geq 0.$$

Let $\{k_n^{(\epsilon)}, k_{i,n}^{(w)}\}$ for $i \in \{1, ..., q-3\}$ be intermediate order sequences. We use intermediate order statistics $\epsilon^{(a)}_{(k_n^{(\epsilon)})}(\theta)$ and $w^{(a)}_{i,(k_{i,n}^{(w)})}(\theta)$ to gauge when an extreme observation occurs, a common practice in the extreme value theory and robust estimation literatures. See Hill (2011) for references. Now define indicator functions for trimming

$$\hat{I}_{n,t}^{(\epsilon)}(\theta) \equiv I\left(|\epsilon_t(\theta)| \leq \epsilon^{(a)}_{(k_n^{(\epsilon)})}(\theta)\right)$$

$$\hat{I}_{i,n,t}^{(w)}(\theta) \equiv I\left(|w_{i,t}(\theta)| \leq w^{(a)}_{i,(k_{i,n}^{(w)})}(\theta)\right) \text{ and } \hat{I}_{n,t}^{(x)}(\theta) \equiv \begin{cases} \prod_{i=1}^{q-3} \hat{I}_{i,n,t}^{(w)}(\theta) \text{ if } q > 3 \\ 1 \text{ if } q = 3, \end{cases}$$

and tail-trimmed variables and equations

$$\hat{\epsilon}_{n,t}^*(\theta) \equiv \epsilon_t(\theta)\hat{I}_{n,t}^{(\epsilon)}(\theta) \text{ and } \hat{w}_{n,t}^*(\theta) \equiv w_t(\theta)\hat{I}_{n,t}^{(w)}(\theta) \text{ and } \hat{x}_{n,t}^*(\theta) \equiv \left[s_t(\theta)', \hat{w}_{n,t}^*(\theta)\right]'$$

$$\hat{m}_{n,t}^*(\theta) \equiv \left(\hat{\epsilon}_{n,t}^{*2}(\theta) - \frac{1}{n}\sum_{t=1}^{n} \hat{\epsilon}_{n,t}^{*2}(\theta)\right) \times \hat{x}_{n,t}^*(\theta). \tag{3}$$

As in Hill (2015a) and Aguilar and Hill (2015), we re-center $\epsilon_t(\theta)$ after trimming to eradicate small sample bias that arises from trimming. This allows for intrinsically simpler symmetric trimming even if $\epsilon_t$ has an asymmetric distribution.

If over-identifying restrictions are not used such that $x_t(\theta) = s_t(\theta)$, then we use

$$\hat{m}_{n,t}^*(\theta) \equiv \left(\hat{\epsilon}_{n,t}^{*2}(\theta) - \frac{1}{n}\sum_{t=1}^{n} \hat{\epsilon}_{n,t}^{*2}(\theta)\right) \times s_t(\theta) \text{ where } \hat{\epsilon}_{n,t}^*(\theta) \equiv \epsilon_t(\theta)\hat{I}_{n,t}^{(\epsilon)}(\theta).$$

If any added instrument $w_{i,t}$ has a finite variance then we do not need to trim by it. It is easy to show, however, that if we trim by all components in $w_t(\theta)$ then it is asymptotically equivalent to only trimming by those elements with an infinite variance (cf. Hill, 2015a, 2013). We therefore assume that each $w_{i,t}(\theta)$ is trimmed in order to reduce notation.

Although $s_t(\theta)$ has an $L_2$-bounded envelope, in small samples components of $s_t(\theta)$ may be influenced by large observations $y_{t-1}$. Consider that in the case of no GARCH effects $\alpha^0 + \beta^0 = 0$, it follows $s_t = (\omega^0)^{-1} \times [1, y_{t-1}^2, \omega^0]'$. Thus, in view of continuity, if $\alpha^0 + \beta^0$ is close to zero then $||s_t||$ may be large when $y_{t-1}$ is large. Although Gaussian asymptotics does not require trimming by $y_{t-1}$, we find that an improved robust GEL estimator uses extremal sample information from

$y_{t-1}$ for trimming, even when $\alpha^0 + \beta^0$ is far from zero. In this case the trimmed covariates are

$$\hat{x}^*_{n,t}(\theta) \equiv \left[\hat{s}^*_{n,t}(\theta)', \hat{w}^*_{n,t}(\theta)\right]' \text{ where } \hat{s}^*_{n,t}(\theta) \equiv s_t(\theta)\hat{I}^{(y)}_{n,t-1} \text{ and } \hat{I}^{(y)}_{n,t-1} \equiv I\left(|y_t| \leq y^{(a)}_{(k_n^{(y)})}\right). \quad (4)$$

Since the asymptotic theory for our GEL estimator with $\hat{x}^*_{n,t}(\theta)$ defined as $[s_t(\theta)', \hat{w}^*_{n,t}(\theta)']'$ or $[\hat{s}^*_{n,t}(\theta)', \hat{w}^*_{n,t}(\theta)']'$ is the same, we simply assume the former to reduce notation in the proofs.

## 2.2 Estimator

Let $\rho : \mathcal{D} \to \mathbb{R}_+$ be a twice continuously differentiable concave function, with domain $\mathcal{D}$ containing zero. Write $\rho^{(i)}(u) = (\partial/\partial u)^i \rho(u)$, $i = 0, 1, 2$, and $\rho^{(i)} = \rho^{(i)}(0)$, and assume the normalizations $\rho^{(0)} = \rho(0) = 0$ and $\rho^{(1)} = \rho^{(2)} = -1$. If $\rho(u) = -u^2/2 - u$ we have the Continuously Updated Estimator or Euclidean Empirical Likelihood (cf. Antoine, Bonnal, and Renault, 2007); $\rho(u) = \ln(1-u)$ for $u < 1$ leads to Empirical Likelihood; $\rho(u) = 1 - \exp\{u\}$ represents Exponential Tilting.

The GEL estimator with Imbedded Tail-Trimming (GELITT) solves a classic saddle-point optimization problem (Smith, 1997; Newey and Smith, 2004; Smith, 2011):

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \sup_{\lambda\in\hat{\Lambda}_n(\theta)} \left\{ \frac{1}{n}\sum_{t=1}^{n} \rho\left(\lambda'\hat{m}^*_{n,t}(\theta)\right) \right\} \text{ and } \hat{\lambda}_n = \arg\sup_{\lambda\in\hat{\Lambda}_n(\hat{\theta}_n)} \left\{ \frac{1}{n}\sum_{t=1}^{n} \rho\left(\lambda'\hat{m}^*_{n,t}(\hat{\theta}_n)\right) \right\}, \quad (5)$$

where $\hat{\Lambda}_n(\theta)$ contains those $\lambda$ such that sample $\lambda'\hat{m}^*_{n,t}(\theta) \in \mathcal{D}$ with probability one:

$$\hat{\Lambda}_n(\theta) = \left\{\lambda : \lambda'\hat{m}^*_{n,t}(\theta) \in \mathcal{D} \ a.s., \ t = 1, 2, ..., n\right\}.$$

The non-smoothness of $\hat{m}^*_{n,t}(\theta)$ is irrelevant as long as $w_{i,t}(\theta)$ are differentiable, and $\epsilon_t(\theta)$ and $w_{i,t}(\theta)$ have smooth distributions (Parente and Smith, 2011; Hill, 2015a, 2013).

Asymptotics for $[\hat{\theta}'_n, \hat{\lambda}'_n]'$ requires non-random threshold sequences associated with the sample order statistics. Let positive sequences of functions $\{c^{(\epsilon)}_n(\theta), c^{(w)}_{i,n}(\theta), \}$ satisfy for any $\theta \in \Theta$

$$P\left(|\epsilon_t(\theta)| \geq c^{(\epsilon)}_n(\theta)\right) = \frac{k^{(\epsilon)}_n}{n} \text{ and } P\left(|w_{i,t}(\theta)| \geq c^{(w)}_{i,n}(\theta)\right) = \frac{k^{(w)}_{i,n}}{n}. \quad (6)$$

Thus, for example, $\epsilon^{(a)}_{(k_n^{(\epsilon)})}(\theta)$ estimates $c^{(\epsilon)}_n(\theta)$ since $\epsilon^{(a)}_{(k_n^{(\epsilon)})}(\theta)$ is the sample $k^{(\epsilon)}_n/n$ upper two-tailed quantile. Since we assume below that $\epsilon_t(\theta)$ and $w_t(\theta)$ have continuous distributions, such sequences $\{c^{(\epsilon)}_n(\theta), c^{(w)}_{i,n}(\theta)\}$ exist for all $\theta$ and any choice of fractiles $\{k^{(\epsilon)}_n, k^{(w)}_{i,n}\}$. Now define

trimming indicator functions

$$I_{n,t}^{(\epsilon)}(\theta) \equiv I\left(|\epsilon_t(\theta)| \leq c_n^{(\epsilon)}(\theta)\right) \quad \text{and} \quad I_{i,n,t}^{(w)}(\theta) \equiv I\left(|w_{i,t}(\theta)| \leq c_{i,n}^{(w)}(\theta)\right),$$

write the composite covariate indicator $I_{n,t}^{(x)}(\theta) = \prod_{i=1}^{q-3} I_{i,n,t}^{(w)}(\theta)$, and define tail-trimmed variables and equations

$$\epsilon_{n,t}^*(\theta) \equiv \epsilon_t(\theta) I_{n,t}^{(\epsilon)}(\theta) \quad \text{and} \quad w_{n,t}^*(\theta) \equiv w_t(\theta) I_{n,t}^{(w)}(\theta)$$

$$m_{n,t}^*(\theta) \equiv \left(\epsilon_{n,t}^{*2}(\theta) - E\left[\epsilon_{n,t}^{*2}(\theta)\right]\right)\left(x_{n,t}^*(\theta) - E\left[x_{n,t}^*(\theta)\right]\right).$$

In view of the re-centering of $\epsilon_t(\theta)$ for $\hat{m}_{n,t}^*(\theta)$ in (3), it can be shown that asymptotics for $\hat{\theta}_n$ are grounded on $m_{n,t}^*(\theta)$. See the appendix.

Notice by error independence, re-centering, and $\Im_{t-1}$-measurability of $x_t$, it follows $m_{n,t}^*$ is a martingale difference with respect to $\Im_t$ since

$$E\left[m_{n,t}^* | \Im_{t-1}\right] = \left(x_{n,t}^* - E\left[x_{n,t}^*\right]\right) \times E\left[\left(\epsilon_{n,t}^{*2} - E\left[\epsilon_{n,t}^{*2}\right]\right) | \Im_{t-1}\right] = 0. \tag{7}$$

## 2.3 Main Results

Define moment suprema for $\epsilon_t(\theta)$, and $w_{i,t}(\theta)$ provided over-identifying weights are used:

$$\kappa_\epsilon(\theta) \equiv \sup\left\{\alpha > 0 : E|\epsilon_t(\theta)|^\alpha < \infty\right\} \quad \text{and} \quad \kappa_i(\theta) \equiv \sup\left\{\alpha > 0 : E|w_{i,t}(\theta)|^\alpha < \infty\right\}.$$

Note that $\kappa_\epsilon = \infty$ or $\kappa_i = \infty$ are possible, for example if $\epsilon_t$ is Gaussian, or $w_{i,t}$ is bounded.[8] Let $\Theta_{1,i} \subseteq \Theta$ be the set of all $\theta$ such that $\kappa_i(\theta) \leq 1$, where $\Theta_{1,i}$ may be empty. Drop $\theta^0$ such that $\kappa_\epsilon = \kappa_\epsilon(\theta^0)$ and $\kappa_i = \kappa_i(\theta^0)$.

We require the following moment, memory and tail properties.

**Assumption A.**

1. $z_t(\theta) \in \{\epsilon_t(\theta), w_{i,t}(\theta)\}$ *have for each* $\theta \in \Theta$ *strictly stationary, ergodic, and absolutely continuous non-degenerate finite dimensional distributions that are uniformly bounded:* $\sup_{a\in\mathbb{R},\theta\in\Theta}\{(\partial/\partial a)P(z_t(\theta) \leq a)\} < \infty$ *and* $\sup_{a\in\mathbb{R},\theta\in\Theta} ||(\partial/\partial\theta)P(z_t(\theta) \leq a)|| < \infty$.

2. $\kappa_i > 1$ *and* $\kappa_\epsilon > 2$. *If* $\kappa_\epsilon \leq 4$ *then* $P(|\epsilon_t| > a) = da^{-\kappa_\epsilon}(1 + o(1))$ *where* $d \in (0, \infty)$. *If* $\Theta_{1,i}$ *is*

[8] Consider an ARCH(1) model $\sigma_t^2 = \omega^0 + \alpha^0 y_{t-1}^2$ with $\omega^0, \alpha^0 > 0$. Then, for example, the weights $x_t(\theta) = [s_t(\theta)', s_{t-1}(\theta)']'$ are bounded since $s_t(\theta)$ is uniformly bounded.

*not empty such that* $\kappa_i(\theta) \leq 1$ *for some* $\theta$*, then* $P(|w_{i,t}(\theta)| > c)\} = d_i(\theta)c^{-\kappa_i(\theta)}(1+o(1))$ *where* $\inf_{\theta\in\Theta_{1,i}} d_i(\theta) > 0$*,* $\inf_{\theta\in\Theta_{1,i}} \kappa_i(\theta) > 0$ *and* $o(1)$ *is not a function of* $\theta$*.*

3. $w_t(\theta)$ *is* $\Im_{t-1}$*-measurable, continuous, differentiable, and* $E[\sup_{\theta\in\Theta} |w_{i,t}(\theta)|^\iota] < \infty$ *for some tiny* $\iota > 0$*.*

4. $k_n/n^\iota \to \infty$ *for some tiny* $\iota > 0$*.*

**Remark 1** Distribution continuity and differentiability of $m_t(\theta) = (\epsilon_t^2(\theta) - 1)x_t(\theta)$ ensure a unique solution to the GELITT estimation problem exists (cf. Cizek, 2008; Hill, 2015a, 2013).

**Remark 2** Paretian tails in the heavy tail case simplify characterizing tail-trimmed moments by Karamata's Theorem, while tail-trimmed moments arise in the GELITT estimator scale, defined below. We impose a Paretian tail on $w_{i,t}(\theta)$ when $\kappa_i(\theta) \leq 1$ since the mapping $w_{i,t} : \Theta \to \mathbb{R}$ is not here defined. If the mapping were known then in principle we would only need to consider $w_{i,t}$.

**Remark 3** We impose a lower bound on how fast the number of trimmed extremes $k_n$ increases in order to simplify proving a uniform law of large numbers for tail-trimmed dependent data. See Lemma A.4 in the appendix, and its proof in Hill and Prokhorov (2014).

**Remark 4** If $w_t(\theta)$ only contains lags of $s_t(\theta)$ then $\sup_{\theta\in\Theta} ||w_t(\theta)||$ is $L_2$-bounded in view of $\alpha + \beta > 0$ (Francq and Zakoïan, 2004), hence $\Theta_{1,i}$ is empty and A.3 holds.

We now state the main results. Let $\mathbf{0}$ be a $q \times 1$ vector of zeros. Define all parameters

$$\vartheta^0 \equiv [\theta^{0\prime}, \mathbf{0}']' \in \mathbb{R}^{q+3} \quad \text{and} \quad \hat{\vartheta}_n \equiv [\hat{\theta}_n', \hat{\lambda}_n']' \in \mathbb{R}^{q+3},$$

and define covariance and scale matrices

$$\Sigma_n(\theta) \equiv E\left[m_{n,t}^*(\theta)m_{n,t}^*(\theta)'\right] \in \mathbb{R}^{q\times q} \tag{8}$$

$$\mathcal{J}_n(\theta) \equiv -E\left[\left(x_{n,t}^*(\theta) - E\left[x_{n,t}^*(\theta)\right]\right)\left(s_t(\theta) - E\left[s_t(\theta)\right]\right)'\right] \in \mathbb{R}^{q\times 3}$$

$$\mathcal{V}_n(\theta) \equiv n\mathcal{J}_n(\theta)'\Sigma_n^{-1}(\theta)\mathcal{J}_n(\theta) \in \mathbb{R}^{3\times 3}$$

$$\mathcal{A}_n \equiv \begin{bmatrix} \mathcal{V}_n & 0 \\ 0 & n\mathcal{P}_n^{-1} \end{bmatrix} \in \mathbb{R}^{(q+3)\times(q+3)} \quad \text{where} \quad \mathcal{P}_n \equiv \Sigma_n^{-1} - \Sigma_n^{-1}\mathcal{J}_n\left(\mathcal{J}_n'\Sigma_n^{-1}\mathcal{J}_n\right)^{-1}\mathcal{J}_n'\Sigma_n^{-1} \in \mathbb{R}^{q\times q}.$$

The mean-centered Jacobian $\mathcal{J}_n$ arises from the re-centered error in the estimating equations $\hat{m}^*_{n,t}(\theta) = (\hat{\epsilon}^{*2}_{n,t}(\theta) - 1/n \sum_{t=1}^n \hat{\epsilon}^{*2}_{n,t}(\theta)) \times \hat{x}^*_{n,t}(\theta)$, since this is asymptotically equivalent to $m^*_{n,t}(\theta) = (\epsilon^{*2}_{n,t}(\theta) - E\left[\epsilon^{*2}_{n,t}(\theta)\right])) \times (x^*_{n,t}(\theta) - E\left[x^*_{n,t}(\theta)\right])$.

We first prove consistency from first principles, since a standard first order expansion for asymptotic normality involves an estimator of $\mathcal{J}_n$. We can only analyze the latter asymptotically if we first know $\hat{\theta}_n \xrightarrow{p} \theta^0$. See Appendix A for all proofs.

**Theorem 2.1** *Under Assumption A* $\hat{\theta}_n \xrightarrow{p} \theta^0$ *and* $n^{1/2}\Sigma_n^{1/2}\hat{\lambda}_n = O_p(1)$.

Second, $\hat{\theta}_n$ and $\hat{\lambda}_n$ are jointly asymptotically normal.

**Theorem 2.2** *Under Assumption A* $\mathcal{A}_n^{1/2}(\hat{\vartheta}_n - \vartheta^0) \xrightarrow{d} N(0, I_{q+3})$, *in particular* $\mathcal{V}_n^{1/2}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N(0, I_3)$.

**Remark 5** The GELITT scales $A_n$ and $V_n$ are identical in form to the scales for the conventional GEL estimator (Newey and Smith, 2004).

**Remark 6** By the martingale difference property, $E[\epsilon_t^2] = 1$ and dominated convergence, it follows

$$\begin{aligned}\Sigma_n &= E\left[\left(\epsilon^{*2}_{n,t} - E\left[\epsilon^{*2}_{n,t}\right]\right)^2\right] \times E\left[\left(x^*_{n,t} - E\left[x^*_{n,t}\right]\right)\left(x^*_{n,t} - E\left[x^*_{n,t}\right]\right)'\right] \\ &\sim \left(E\left[\epsilon^{*4}_{n,t}\right] - 1\right) \times E\left[\left(x^*_{n,t} - E\left[x^*_{n,t}\right]\right)\left(x^*_{n,t} - E\left[x^*_{n,t}\right]\right)'\right].\end{aligned}$$

Hence, in the case of exact identification $x_t(\theta) = s_t(\theta)$ we have $J_n = E[(s_t - E[s_t])(s_t - E[s_t])']$ and therefore

$$\mathcal{V}_n \sim n\frac{1}{E\left[\epsilon^{*4}_{n,t}\right] - 1} E\left[\left(s_t - E\left[s_t\right]\right)\left(s_t - E\left[s_t\right]\right)'\right].$$

Similarly, when $x_t(\theta)$ contains only $s_t(\theta)$ and its lags then

$$\|\mathcal{V}_n\| \sim Kn\frac{1}{E\left[\epsilon^{*4}_{n,t}\right]}.$$

The same order applies whenever $x_t$ is square integrable, e.g. it only contains $s_t$ and its lags. In this case if $X_t \equiv x_t - E[x_t]$ and $S_t \equiv s_t - E[s_t]$ then:

$$\mathcal{V}_n \sim n\frac{1}{E\left[\epsilon^{*4}_{n,t}\right] - 1}\mathcal{V} \text{ where } \mathcal{V} = \mathcal{J}'\Sigma_x^{-1}\mathcal{J},\ \mathcal{J} = -E\left[X_t S_t'\right] \text{ and } \Sigma_x = E\left[X_t X_t'\right].$$

Hence $(n/(E[\epsilon^{*4}_{n,t}] - 1))^{1/2}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N(0, \mathcal{V}^{-1})$.

**Remark 7** If $E[\epsilon_t^4] < \infty$ and $x_t$ is square integrable then GELITT obtains the same asymptotic distribution as the untrimmed GEL estimator: $n^{1/2}(\hat{\theta}_n - \theta^0) \xrightarrow{d} N(0, (E[\epsilon_t^4] - 1)\mathcal{V}^{-1})$, with $\mathcal{V}$ defined above.

**Remark 8** Notice

$$n\mathcal{P}_n^{-1} = n\Sigma_n \left( I - \mathcal{J}_n \left( \mathcal{J}_n' \Sigma_n^{-1} \mathcal{J}_n \right)^{-1} \mathcal{J}_n' \Sigma_n^{-1} \right)^{-1} \sim Kn\Sigma_n$$

hence $\hat{\lambda}_n$ has a faster rate of convergence than $\hat{\theta}_n$ when $E[\epsilon_t^4] = \infty$. Indeed, by Theorem 2.1 the rate is $n^{1/2}||\Sigma_n||^{1/2} \sim Kn^{1/2}E[\epsilon_{n,t}^{*4}] \times ||E[x_{n,t}^* x_{n,t}^{*\prime}]||$ which is greater than $n^{1/2}$ when $E[\epsilon_t^4] = \infty$.

The rate of convergence can be easily obtained if over-identifying weights $w_t$ are square integrable, e.g. $w_t$ only contain lags of the score $s_t$, since then $x_t$ is $L_2$-bounded and the Jacobian $\mathcal{J}_n = -E[(x_{n,t}^* - E[x_{n,t}^*])(s_t - E[s_t])']$ is uniformly bounded: $\limsup_{n\to\infty} ||\mathcal{J}_n|| \leq K$. In order to see this, by construction of the thresholds and power law Assumption A.2, if $\kappa_\epsilon \in (2, 4]$ then $c_n^{(\epsilon)} = d^{1/\kappa_\epsilon}(n/k_n^{(\epsilon)})^{1/\kappa_\epsilon}$. Therefore if $E[\epsilon_t^4] = \infty$ then by Karamata's Theorem[9]

$$\kappa_\epsilon \in (2, 4) : E\left[\epsilon_{n,t}^{*4}\right] \sim \frac{4}{4 - \kappa_\epsilon} \left(c_n^{(\epsilon)}\right)^4 P\left(|\epsilon_t| > c_n^{(\epsilon)}\right) = \frac{4}{4 - \kappa_\epsilon} d^{4/\kappa_\epsilon} \left(\frac{n}{k_n^{(\epsilon)}}\right)^{4/\kappa_\epsilon - 1} \tag{9}$$

$$\kappa_\epsilon = 4 : E\left[\epsilon_{n,t}^{*4}\right] \sim d\ln(n).$$

In either case $\kappa_\epsilon = 4$ or $\kappa_\epsilon \in (2, 4)$ it follows

$$E\left[\epsilon_{n,t}^{*4}\right] - 1 = E\left[\epsilon_{n,t}^{*4}\right] \times (1 + o(1)). \tag{10}$$

Combine Theorem 2.2 with (9) and (10) to deduce the next result.

**Corollary 2.3** *Let Assumption A hold, and if $q > 3$ then let $w_t$ be square integrable. Then*

$$\kappa_\epsilon \in (2, 4) : \frac{n^{1/2}}{\left(n/k_n^{(\epsilon)}\right)^{2/\kappa_\epsilon - 1/2}} \left(\hat{\theta}_n - \theta^0\right) \xrightarrow{d} N\left(0, \frac{4}{4 - \kappa_\epsilon} d^{4/\kappa_\epsilon} \times \mathcal{V}^{-1}\right)$$

$$\kappa_\epsilon = 4 : \left(\frac{n}{\ln(n)}\right)^{1/2} \left(\hat{\theta}_n - \theta^0\right) \xrightarrow{d} N\left(0, d \times \mathcal{V}^{-1}\right)$$

[9]See Theorem 0.6 in Resnick (1987). The case $\kappa_\epsilon = 4$ follows by observing if $\kappa_\epsilon = 4$ then $c_n^{(\epsilon)} = d^{1/4}(n/k_n^{(\epsilon)})^{1/4}$, hence for finite $a > 0$ there exists $K > 0$ such that $E[\epsilon_t^4 I_{n,t}^{(\epsilon)}] = \int_0^{(c_n^{(\epsilon)})^4} P(|\epsilon_t| > u^{1/4})du = K + \int_a^{(c_n^{(\epsilon)})^4} u^{-1}du = K + 4d\ln(c_n^{(\epsilon)}) \sim K + d\ln(n) \sim d\ln(n)$.

*where* $\mathcal{V} \equiv \mathcal{J}'\Sigma_x^{-1}\mathcal{J}$ *with* $\mathcal{J} \equiv -E[(x_t - E[x_t])(s_t - E[s_t])']$ *and* $\Sigma_x \equiv E[(x_t - E[X_t])(x_t - E[x_t])']$.

As long as $\epsilon_t$ has an unbounded fourth moment $\kappa_\epsilon \in (2, 4]$, the rate of convergence is $o(n^{1/2})$. If $\kappa_\epsilon \in (2, 4)$ then by maximizing the trimming amount $k_n^{(\epsilon)}$ and therefore making $k_n^{(\epsilon)}$ arbitrarily close to a fixed portion $\zeta n$ of $n$ where $\zeta \in (0, 1)$, we can optimize the rate of convergence. Simply let $k_n^{(\epsilon)} \sim n/g_n$ for $g_n \to \infty$ at a slow rate to deduce $\hat{\theta}_n$ can be made as close to $n^{1/2}$-convergent as we choose. A parametric rule for $k_n^{(\epsilon)}$ is convenient, for example

$$k_n^{(\epsilon)} = [\zeta n / \ln(n)] \quad \text{where} \quad \zeta \in (0, 1]. \tag{11}$$

Then for any $\kappa_\epsilon \in (2, 4)$ we have

$$\frac{n^{1/2}}{(\ln(n))^{2/\kappa_\epsilon - 1/2}} \left(\hat{\theta}_n - \theta^0\right) \xrightarrow{d} N(0, \mathcal{V}(\zeta, \kappa_\epsilon, d)), \text{ with } \mathcal{V}(\zeta, \kappa_\epsilon, d) \equiv \frac{1}{\zeta^{4/\kappa_\epsilon - 1}} \frac{4}{4 - \kappa_\epsilon} d^{4/\kappa_\epsilon} \times \mathcal{V}^{-1}. \tag{12}$$

In this case the rate of convergence is identical to Quasi-Maximum Tail-Trimmed Likelihood in Hill (2015a) since the estimating equations are identical or similar to QML score equations. Thus, when $\kappa_\epsilon \in (2, 4]$ the GELITT estimator converges faster than QML as long as $k_n^{(\epsilon)} \sim n/g_n$ for slow $g_n \to \infty$ (see Hill, 2015a).

Notice that by letting $\zeta$ be large we can diminish the asymptotic variance $V(\zeta, \kappa_\epsilon, d)$. By first order asymptotics, it is always better to trim more extreme values per sample since we achieve both a higher rate of convergence and lower asymptotic variance. However, in Section 4 we exploit higher order asymptotics and show that the higher order bias of GELITT is smaller when trimming is reduced.[10] In the case of EL or exact identification, the bias monotonically decreases as trimming is reduced. Indeed, it is easily revealed by simulation that a greater amount of trimming induces small sample bias for standard GEL criterion, e.g. EL, CUE, and ET. Thus, while first order efficiency and the rate of convergence are augmented with a trimming rule like (11) with large $\zeta$, higher order bias is reduced by setting $\zeta$ small, e.g. $\zeta = .05$ as we do in the Section 8 simulation study.

In principle, there is an optimal trimming rule implied by the combination of the first and higher order asymptotic arguments. However, a higher order mean-squared-error will favor efficiency in heavy tailed cases since the higher order variance will dominate the squared bias. Minimizing this mean-squared-error is not practical since it will simply lead to setting $k_n^{(\epsilon)}$ close

[10] We thank a referee for suggesting that second order asymptotics can be useful in justifying optimal trimming rules.

to $n$. Nevertheless, the preceding points to a dominant strategy: *elevate the rate of convergence while controlling higher order bias* by elevating the rate $k_n^{(\epsilon)} \to \infty$ as $n \to \infty$ and, for a given sample, by setting $k_n^{(\epsilon)}$ as a small value relative to $n$.

Finally, although the GELITT rate is optimized to its upper bound $n^{1/2}$ when $k_n^{(\epsilon)} = [\zeta n]$, we cannot use a fixed portion since $\hat{\theta}_n$ need not be consistent for $\theta^0$. This follows since $1/n \sum_{t=1}^n \hat{\epsilon}_{n,t}^{*2} \xrightarrow{p} [0, 1)$ under Assumption A, hence the centered error $\hat{\epsilon}_{n,t}^{*2}(\theta) - 1/n \sum_{t=1}^n \hat{\epsilon}_{n,t}^{*2}(\theta)$ in $\hat{m}_{n,t}^*(\theta)$ may not identify $\theta^0$ (see, e.g., Sakata and White, 1998; Mancini, Ronchetti, and Trojani, 2005). If the distribution of $\epsilon_t$ were assumed, this bias can in theory be removed by simulation-based indirect inference, as in Cantoni and Ronchetti (2001) and Ronchetti and Trojani (2001).

## 2.4 Feasible GELITT

In practice $\sigma_t^2(\theta)$ cannot be computed for $t \leq 1$, so an iterated approximation must be used. Define

$$h_t(\theta) = \tilde{\omega} > 0 \text{ for } t = 0, \text{ and } h_t(\theta) = \omega + \alpha y_{t-1}^2 + \beta h_{t-1}(\theta) \text{ for } t = 1, 2, ... \tag{13}$$

where $\tilde{\omega}$ is not necessarily an element of $\theta^0$. Write $h_t^\theta(\theta) \equiv (\partial/\partial\theta) h_t(\theta)$ and $h_t^{\theta,\theta}(\theta) \equiv (\partial/\partial\theta) h_t^\theta(\theta)$. Under Assumption A it can be shown that stationary and ergodic solutions to (13) and the corresponding equations for $h_t^\theta(\theta)$ and $h_t^{\theta,\theta}(\theta)$ exist (see Lemma A.7 in Hill, 2014a, cf. Meitz and Saikkonen, 2011).

Now replace $\sigma_t^2(\theta)$ with $h_t(\theta)$ and define

$$\mathring{\epsilon}_t(\theta) \equiv \frac{y_t}{h_t(\theta)^{1/2}} \quad \text{and} \quad \mathring{s}_t(\theta) \equiv \frac{1}{h_t(\theta)} h_{i,t}^\theta(\theta) \quad \text{and} \quad \mathring{x}_t(\theta) \equiv \left[\mathring{s}_t(\theta)', \mathring{w}_t(\theta)'\right]'.$$

We write $\mathring{w}_t(\theta$ since the added instruments may be a function of $h_t(\theta)$, for example when $\mathring{w}_t(\theta)$ contains lags of $\mathring{s}_t(\theta)$. The tail-trimmed versions are

$$\widehat{\mathring{\epsilon}}_{n,t}^*(\theta) \equiv \mathring{\epsilon}_t(\theta) I\left(|\mathring{\epsilon}_t(\theta)| \leq \mathring{\epsilon}_{(k_n^{(\epsilon)})}^{(a)}(\theta)\right) \quad \text{and} \quad \widehat{\mathring{x}}_{n,t}^*(\theta) \equiv \left[\mathring{s}_t(\theta)', \widehat{\mathring{w}}_{n,t}^*(\theta)'\right]'$$

$$\mathring{\epsilon}_{n,t}^*(\theta) \equiv \mathring{\epsilon}_t(\theta) I\left(|\mathring{\epsilon}_t(\theta)| \leq c_n^{(\epsilon)}(\theta)\right) \quad \text{and} \quad \mathring{x}_{n,t}^*(\theta) \equiv \left[\mathring{s}_t(\theta)', \mathring{w}_{n,t}^*(\theta)'\right],$$

hence the equations are

$$\widehat{\mathring{m}}_{i,n,t}^*(\theta) \equiv \left(\widehat{\mathring{\epsilon}}_{n,t}^*(\theta) - \frac{1}{n}\sum_{t=1}^n \widehat{\mathring{\epsilon}}_{n,t}^*(\theta)\right) \widehat{\mathring{x}}_{i,n,t}^*(\theta)$$

$$\mathring{m}^*_{i,n,t}(\theta) \equiv \left(\mathring{\epsilon}^*_{n,t}(\theta) - E\left[\mathring{\epsilon}^*_{n,t}(\theta)\right]\right)\left(\mathring{x}^*_{n,t}(\theta) - E\left[\mathring{x}^*_{n,t}(\theta)\right]\right),$$

and the feasible estimators are

$$\widehat{\mathring{\theta}}_n = \arg\min_{\theta\in\Theta} \sup_{\lambda\in\hat{\Lambda}_n(\theta)} \left\{\frac{1}{n}\sum_{t=1}^n \rho\left(\lambda'\widehat{\mathring{m}}^*_{n,t}(\theta)\right)\right\} \text{ and } \widehat{\mathring{\lambda}}_n = \arg\sup_{\lambda\in\hat{\Lambda}_n(\widehat{\mathring{\theta}}_n)} \left\{\frac{1}{n}\sum_{t=1}^n \rho\left(\lambda'\widehat{\mathring{m}}^*_{n,t}(\widehat{\mathring{\theta}}_n)\right)\right\}.$$

Define $\widehat{\mathring{\vartheta}}_n \equiv [\widehat{\mathring{\theta}}'_n, \widehat{\mathring{\lambda}}'_n]'$. The feasible and infeasible estimators have the same limit distribution. The proof is similar to the proof of Theorem 2.3 in Hill (2015a) and is therefore omitted.

**Lemma 2.4** *Under Assumption A* $\mathcal{A}_n^{1/2}(\widehat{\mathring{\vartheta}}_n - \hat{\vartheta}_n) \xrightarrow{p} 0$.

We only work with the infeasible $\hat{\vartheta}_n$ in all that follows for the sake of notational ease.

# 3 Extremal Information of Implied Probabilities

Recall $\rho^{(1)}(u) = (\partial/\partial u)\rho(u)$. By the GELITT first order condition it is easy to show the implied probabilities or profiles have a classic form (Antoine, Bonnal, and Renault, 2007; Newey and Smith, 2004)

$$\hat{\pi}^*_{n,t}(\theta) = \frac{\rho^{(1)}\left(\hat{\lambda}'_n\hat{m}^*_{n,t}(\theta)\right)}{\sum_{t=1}^n \rho^{(1)}\left(\hat{\lambda}'_n\hat{m}^*_{n,t}(\theta)\right)} \text{ where } \hat{\lambda}_n = \arg\sup_{\lambda\in\hat{\Lambda}_n(\hat{\theta}_n)} \left\{\frac{1}{n}\sum_{t=1}^n \rho\left(\lambda'\hat{m}^*_{n,t}(\hat{\theta}_n)\right)\right\}. \tag{14}$$

See Appendix A.3 for derivation of the first order condition, equation (A.8). The profiles $\hat{\pi}^*_{n,t}(\theta)$ promote an empirical counterpart to the GELITT identification condition $E[m^*_{n,t}(\theta^0)] = 0$ since $\hat{\pi}^*_{n,t}(\theta) \in [0,1]$, $\sum_{t=1}^n \hat{\pi}^*_{n,t}(\theta) = 1$, and by the first order condition $\sum_{t=1}^n \hat{\pi}^*_{n,t}(\theta)\hat{m}^*_{n,t}(\hat{\theta}_n) = 0$.

We begin by gleaning information about extremes from $\hat{\pi}^*_{n,t}(\theta)$ in the case of tail-trimmed CUE due to its tractability. Since $\rho$ is quadratic in this case we have (Antoine, Bonnal, and Renault, 2007)

$$\hat{\pi}^*_{n,t}(\theta) = \frac{1 + \hat{\lambda}'_n\hat{m}^*_{n,t}(\theta)}{\sum_{t=1}^n \left\{1 + \hat{\lambda}'_n\hat{m}^*_{n,t}(\theta)\right\}}. \tag{15}$$

Now define the set of time indices at which an error is trimmed:

$$\widehat{\mathcal{I}}^*_n(\theta) \equiv \left\{t : \hat{\epsilon}^*_{n,t}(\theta) = 0\right\} \text{ and } \widehat{\mathcal{I}}^*_n \equiv \widehat{\mathcal{I}}^*_n(\theta^0).$$

Thus, since $\hat{\epsilon}^*_{n,t}(\theta) \equiv \epsilon_t(\theta)\hat{I}^{(\epsilon)}_{n,t}(\theta)\prod_{i=1}^{q-3}\hat{I}^{(w)}_{i,n,t}(\theta)$, then $t \in \widehat{\mathcal{I}}^*_n(\theta)$ when $\epsilon_t$ is large, or any over-identifying weight $w_{i,t}(\theta)$ is large. Then for any $t \in \widehat{\mathcal{I}}^*_n(\theta)$ we have $\hat{m}^*_{n,t}(\theta) = -(1/n\sum_{s=1}^n \hat{\epsilon}^{*2}_{n,s}(\theta)) \times \hat{x}^*_{n,t}(\theta)$ *a.s.*, hence by dominated convergence and limit theory developed in the appendix:

$$\hat{m}^*_{n,t}(\theta) = -\hat{x}^*_{n,t}(\theta) \times (1 + o_p(1)) \,. \tag{16}$$

By imitating arguments in Antoine, Bonnal, and Renault (2007, Theorem 3.1),$\hat{\pi}^*_{n,t}(\theta)$ has the decomposition

$$\hat{\pi}^*_{n,t}(\theta) = \frac{1}{n} - \frac{1}{n}\hat{m}^*_n(\theta)'\check{\Sigma}^{-1}_n(\theta) \times \left\{\hat{m}^*_{n,t}(\theta) - \hat{m}^*_n(\theta)\right\} \tag{17}$$

where

$$\hat{m}^*_n(\theta) \equiv \frac{1}{n}\sum_{t=1}^n \hat{m}^*_{n,t}(\theta) \;\text{ and }\; \check{\Sigma}_n(\theta) \equiv \frac{1}{n}\sum_{t=1}^n \left\{\hat{m}^*_{n,t}(\theta) - \hat{m}^*_n(\theta)\right\}\hat{m}^*_{n,t}(\theta)'.$$

Since $\hat{m}^*_n\check{\Sigma}^{-1}_n\hat{m}^*_n > 0$ *a.s.* and $\hat{m}^*_n \xrightarrow{p} 0$, it follows by (16) and (17) that periods with a trimmed error have an elevated profile $\hat{\pi}^*_{n,t}$:

$$\hat{\pi}^*_{n,t} = \frac{1}{n} + \frac{1}{n}\hat{m}^{*\prime}_n\check{\Sigma}^{-1}_n\hat{m}^*_n + \frac{1}{n}\hat{m}^{*\prime}_n\check{\Sigma}^{-1}_n\hat{x}^*_{n,t} \times (1 + o_p(1)) = \frac{1}{n} + \frac{1}{n}\hat{m}^{*\prime}_n\check{\Sigma}^{-1}_n\hat{m}^*_n\,(1 + o_p(1)) > \frac{1}{n}\ a.s.$$

**Lemma 3.1** *We have $\hat{\pi}^*_{n,t} > 1/n$ with probability approaching one for each period $t$ with a trimmed error (due to a large error and/or large over-identifying weight).*

We can go further by applying limit theory presented in the appendix to (17) to obtain

$$\begin{aligned}\hat{\pi}^*_{n,t} &= \frac{1}{n} + \frac{1}{n^2}\left\{\frac{1}{n^{1/2}}\Sigma_n^{-1/2}\sum_{t=1}^n \hat{m}^*_{n,t}\right\}'\left\{\frac{1}{n^{1/2}}\Sigma_n^{-1/2}\sum_{t=1}^n \hat{m}^*_{n,t}\right\}(1 + o_p(1)) \\ &= \frac{1}{n} + \frac{1}{n^2} \times \mathcal{X}^2_q \times (1 + o_p(1)) = \frac{1}{n}\left(1 + \frac{1}{n} \times \mathcal{X}^2_q \times (1 + o_p(1))\right)\end{aligned}$$

where $t \in \widehat{\mathcal{I}}^*_n$, where $\mathcal{X}^2_q$ is a chi-squared random variable with $q$ degrees of freedom. Since such $\hat{\pi}^*_{n,t}$ satisfy $n^2(\hat{\pi}^*_{n,t} - 1/n) \xrightarrow{d} \mathcal{X}^2_q$ and $\hat{\pi}^*_{n,t} = n^{-1} + n^{-2}\mathcal{X}^2_q(1 + o_p(1)) \in [0,1]$, apply the Helly-Bray Theorem to deduce on average $\hat{\pi}^*_{n,t}$ is $1/n + q/n^2 + o_p(1/n^2)$ in periods in which an extreme error occurs.

**Lemma 3.2** $E[\hat{\pi}^*_{n,t} \mid t \in \widehat{\mathcal{I}}^*_n] = 1/n + q/n^2 + o_p(1/n^2))$.

Although periods with extremes are deemed damaging for asymptotics, this does not imply they are uninformative. Indeed, they do not receive the *least* informative, or *uniform*, profile

value $1/n$. Rather, tail-trimmed CUE assigns periods with exceptionally large errors or weights an *elevated* (relative to uniform $1/n$) probability, roughly on average $1/n + q/n^2$ for large $n$.

But this begs the question regarding which periods are being assigned smaller or larger profiles in general. Decomposition (17) and limit theory in the appendix reveal in any period $t$

$$\begin{aligned}\hat{\pi}^*_{n,t} &= \frac{1}{n}\left(1+\frac{1}{n}\mathcal{X}_q^2\left(1+o_p(1)\right)\right) - \frac{1}{n}\left\{\frac{1}{n^{1/2}}\Sigma_n^{-1/2}\sum_{t=1}^n \hat{m}^*_{n,t}\right\}' \times \frac{1}{n^{1/2}}\Sigma_n^{-1/2}\hat{m}^*_{n,t}\left(1+o_p(1)\right) \\ &= \frac{1}{n}\left\{1+\frac{1}{n}\mathcal{X}_q^2 - \mathcal{Z}' \times \frac{1}{n^{1/2}}\Sigma_n^{-1/2}\hat{m}^*_{n,t}\right\}\left(1+o_p(1)\right),\end{aligned}$$

where $\mathcal{Z}$ is a standard normal random variable on $\mathbb{R}^q$ that satisfies identically $\mathcal{X}_q^2 = \mathcal{Z}'\mathcal{Z}$. Now assume $n$ is sufficiently large that $1/n\sum_{t=1}^n \hat{\epsilon}^{*2}_{n,t} \approx 1$ hence $\hat{m}^*_{n,t} \approx (\hat{\epsilon}^{*2}_{n,t} - 1)\hat{x}^{*2}_{n,t}$.

An asymptotic random draw $\{y_t\}_{t=1}^\infty$ with a propensity for large errors $\epsilon_t$ and therefore large $\hat{m}^*_{n,t} > 0$ implies a larger likelihood that $\mathcal{Z}' \times \Sigma_n^{-1/2}\hat{m}^*_{n,t} > 0$. But this implies $\hat{\pi}^*_{n,t} < n^{-1}\{1 + n^{-1}\mathcal{X}_q^2\}$ for many periods $t$ when a large error occurs. Thus, in an asymptotic draw when a large error is not particularly rare then any given $t$ with a large error is not especially informative: the ascribed profile weight is closer to the flat weighted value $n^{-1}$ than in periods of extreme values. Put differently, a period $t$ that "*goes with the flow*" is not particularly useful for efficient moment estimation by profiling weighting. In fact, in a sample with many large $\epsilon_t$, any period with a *very large* $\epsilon_t$ that is not so large as to be trimmed is, in probability, the *least* useful in the sense of receiving the smallest $\hat{\pi}^*_{n,t}$.

Contrariwise, periods that go "*against the flow*," that is, periods when $\hat{m}^*_{n,t} < 0$, are assigned the largest $\hat{\pi}^*_{n,t}$. This arises either when $\epsilon_t$ is small and $w_{i,t}$ are not extreme values such that $\hat{\epsilon}^{*2}_{n,t} < 1$, or $\epsilon_t$ and/or $w_{i,t}$ are so large that $\epsilon_t$ is trimmed hence $\hat{m}^*_{n,t} \approx -\hat{x}^{*2}_{n,t}$. Intuitively, large values are useful only if they portray dispersion or leverage: a large $\hat{m}^*_{n,t} > 0$ amongst many large positive $\hat{m}^*_{n,t}$ does not provide much useful information. See also Back and Brown (1993) for a classic interpretation of $\hat{\pi}^*_{n,t}$.

# 4 Higher Order Asymptotics and Fractile Choice

In Appendix A.3 we derive the first order expansion:

$$\mathcal{A}_n^{1/2}\left(\hat{\vartheta}_n - \vartheta^0\right) = -\mathcal{I}_n\Sigma_n^{-1/2}\frac{1}{n^{1/2}}\sum_{t=1}^n m^*_{n,t}\left(1+o_p(1)\right), \tag{18}$$

where $\mathcal{I}_n \in \mathbb{R}^{3\times q}$ satisfies $\mathcal{I}_n'\mathcal{I}_n = I_3$. The expansion with $o_p(1)$ replaced with 0 is identical to the GEL first order expansion in Newey and Smith (2004, eq. (A.8)). Since $m^*_{n,t}$ is a martingale difference with $E[m^*_{n,t}m^{*\prime}_{n,t}] = \Sigma_n$ for any fractile sequences $\{k_n^{(\epsilon)}, k_{i,n}^{(w)}\}$, expansion (18) is not helpful for understanding how $k_n^{(\epsilon)}$ influences small bias. Further, in terms of efficiency for the GARCH parameter estimator $\hat{\theta}_n$, a choice of $k_n^{(\epsilon)}$ nearly equal to $\zeta n$ for $\zeta \in (0,1)$ will minimize $\mathcal{V}_n$ by Corollary 2.3. Thus, by first order asymptotics the best guidance we have is to use $k_n^{(\epsilon)} \sim n/g_n$ for essentially any slowly increasing $g_n \to \infty$, e.g. $k_n^{(\epsilon)} = [\zeta n/\ln(n)]$. In this case Corollary 2.3 shows that larger $\zeta$ is associated with a lower asymptotic variance. In simulation experiments, however, it is easily seen that a small $\zeta$ leads to sharp inference since only then is the small sample bias reduced.

We now shed some light on bias by formally deriving a higher order expansion and use higher order bias to gauge what an optimal number of trimmed observations $k_n^{(\epsilon)}$ should be. We also propose a bias-corrected estimator that corrects for bias due to the GEL structure and due to tail-trimming.

In order to reduce the number of trimming fractiles considered, and without affecting the applicability of our derivations, assume over-identifying instruments $w_t$ are square integrable (e.g. $x_t$ contains only lags of $s_t$) and therefore need not be trimmed:

$$m^*_{n,t}(\theta) \equiv \left(\epsilon^{*2}_{n,t}(\theta) - E\left[\epsilon^{*2}_{n,t}(\theta)\right]\right)(x_t(\theta) - E\left[x_t(\theta)\right]) \quad \text{where} \quad \epsilon^*_{n,t}(\theta) \equiv \epsilon_t(\theta) I^{(\epsilon)}_{n,t}(\theta).$$

Allowing for trimming on the error and instruments would substantially complicate the expansion, but the salient features of our analysis below would still carry over: trimming few observations promotes smaller higher order bias.

## 4.1 Higher Order Expansion

Similar to (18), we need only look to arguments in Newey and Smith (2004) to obtain a higher order expansion. Let $\{z^*_{n,t}\}$ be a tail-trimmed random variable. In order to express an asymptotically valid derivative of a tail-trimmed object, let $z^*_{n,t}(\theta) \equiv z_t(\theta) I_{n,t}(\theta)$ where $z_t(\theta)$ is differentiable, $I_{n,t}(\theta) \in \{0,1\}$ and $\inf_{\theta\in\Theta} I_{n,t}(\theta) \xrightarrow{p} 1$, and define[11]

$$\frac{\mathring{\partial}}{\mathring{\partial}\theta} z^*_{n,t}(\theta) \equiv \left(\frac{\partial}{\partial\theta} z_t(\theta)\right) \times I_{n,t}(\theta).$$

[11]The asymptotic theory supporting the use of such a derivative can be found in the appendices Hill (2013, 2015a).

Define

$$\mathfrak{M}^*_{n,t}(\vartheta) \equiv \rho^{(1)}\left(\lambda' m^*_{n,t}(\theta)\right) \times \begin{bmatrix} \frac{\mathring{\partial}}{\mathring{\partial}\theta} m^*_{n,t}(\theta)'\lambda \\ m^*_{n,t}(\theta) \end{bmatrix}$$

$$\mathfrak{G}^*_n(\vartheta) \equiv E\left[\frac{\mathring{\partial}}{\mathring{\partial}\vartheta}\mathfrak{M}^*_{n,t}(\vartheta)\right],\ \mathfrak{G}^*_{j,n}(\vartheta) \equiv E\left[\frac{\mathring{\partial}^2}{\mathring{\partial}\vartheta_j\mathring{\partial}\vartheta}\mathfrak{M}^*_{n,t}(\vartheta)\right],\ \mathfrak{G}^*_{j,k,n}(\vartheta) \equiv E\left[\frac{\mathring{\partial}^3}{\mathring{\partial}\vartheta_j\mathring{\partial}\vartheta_k\mathring{\partial}\vartheta}\mathfrak{M}^*_{n,t}(\vartheta)\right]$$

$$\mathfrak{A}^*_{n,t} \equiv \frac{\mathring{\partial}}{\mathring{\partial}\vartheta}\mathfrak{M}^*_{n,t} - \mathfrak{G}^*_n \text{ and } \psi^*_{n,t} \equiv -\mathfrak{G}^{*-1}_n \mathfrak{M}^*_{n,t}.$$

Since arguments merely mimic the proof of Lemma A.4 and Theorem 3.1 in Newey and Smith (2004), we prove the following claim in Hill and Prokhorov (2014). Write $\tilde{z}_n \equiv 1/n^{1/2}\sum_{t=1}^n z^*_{n,t}$.

**Theorem 4.1** *Under Assumption A and* $||E[w_t w_t']|| < \infty$*:*

$$\hat{\vartheta}_n - \vartheta^0 = \frac{1}{n^{1/2}}\tilde{\psi}^*_n + \frac{1}{n}Q_1\left(\tilde{\psi}^*_n\right) + \frac{1}{n^{3/2}}Q_2\left(\tilde{\psi}^*_n\right) + O_p\left(\frac{\left(E\left[\epsilon^{*4}_{n,t}\right]\right)^2}{n^2}\right), \tag{19}$$

*where* $Q_1(\tilde{\psi}^*_n) \equiv -\mathfrak{G}^{*-1}_n\{\widetilde{\mathfrak{A}}^*_n\tilde{\psi}^*_n + 1/2\sum_{i=1}^{q+3}\tilde{\psi}^*_{i,n}\mathfrak{G}^*_{i,n}\tilde{\psi}^*_n\}$ *and* $Q_2(\tilde{\psi}^*_n) \equiv -\mathfrak{G}^{*-1}_n\mathfrak{Q}_n$*, with*

$$\mathfrak{Q}_n = \widetilde{\mathfrak{A}}^*_n Q_1\left(\tilde{\psi}^*_n\right) + \frac{1}{2}\sum_{i=1}^{q+3}\left\{\tilde{\psi}^*_{i,n}\mathfrak{G}^*_{i,n}Q_1(\tilde{\psi}^*_n) + Q_{i,1}(\tilde{\psi}^*_n)\mathfrak{G}^*_{i,n}\tilde{\psi}^*_n + \tilde{\psi}^*_{i,n}\mathfrak{G}^*_{i,n}\tilde{\psi}^*_n\right\} + \frac{1}{6}\sum_{i,j=1}^{q+3}\tilde{\psi}^*_{i,n}\tilde{\psi}^*_{j,n}\mathfrak{G}^*_{i,j,n}\tilde{\psi}^*_n.$$

*If* $k^{(\epsilon)}_n \sim n/L(n)$ *for some slowly varying* $L(n) \to \infty$ *then for any* $\kappa_\epsilon > 2$*:*

$$\hat{\vartheta}_n - \vartheta^0 = \frac{1}{n^{1/2}}\tilde{\psi}^*_n + \frac{1}{n}Q_1\left(\tilde{\psi}^*_n\right) + O_p\left(\frac{L(n)}{n^{3/2}}\right) \text{ for slowly varying } L(n) \to \infty \tag{20}$$

*hence the asymptotic (higher order) bias for any* $\kappa_\epsilon > 2$ *is* $Bias(\hat{\vartheta}_n) = n^{-1}E[Q_1(\tilde{\psi}^*_n)]$.

**Remark 9** Since $\tilde{\psi}^*_n$ is a function of $\epsilon^{*2}_{n,t}$ and $\widetilde{\mathfrak{A}}^*_n$ is a function of $\epsilon^{*4}_{n,t}$, it is easily verified that $||E[Q_1(\tilde{\psi}^*_n)]|| \sim KE[\epsilon^{*6}_{n,t}]$ and $||E[Q_2(\tilde{\psi}^*_n)]|| \sim KE[\epsilon^{*10}_{n,t}]$. If we were to disband with trimming and use a third order expansion as above, then we need $E[\epsilon^{10}_t] < \infty$ just to deduce $E[Q_1]$ represents asymptotic (higher order) bias, cf. Rothenberg (1984) and Newey and Smith (2004). The analysis in Newey and Smith (2004) of higher order GEL properties, like bias and efficiency, therefore presumes the existence of substantially higher moments than may in fact exist for many macroeconomic and financial time series. Of course, expansion (19) relies on a third order Taylor

expansion with a remainder: using only a second order expansion reduces the higher moment burden for GEL to $E[\epsilon_t^6] < \infty$. Negligible tail-trimming, however, allows us to impose only $E[\epsilon_t^2] < \infty$ and still retain the *same* structure of higher order terms for GELITT.

**Remark 10** The higher order terms are complicated by tail trimming. Notice $\hat{\vartheta}_n$ exhibits two forms of dynamics: one due to the GEL structure itself, and one due to trimming:

$$\begin{aligned}\hat{\vartheta}_n - \vartheta^0 &= \left\{ \frac{1}{n^{1/2}}\tilde{\psi}_n + \frac{1}{n}Q_1\left(\tilde{\psi}_n\right) + \frac{1}{n^{3/2}}Q_2\left(\tilde{\psi}_n\right) \right\} + O_p\left( \frac{\left(E\left[\epsilon_{n,t}^{*4}\right]\right)^2}{n^2} \right) \\ &\quad + \frac{1}{n^{1/2}}\left(\tilde{\psi}_n^* - \tilde{\psi}_n\right) + \frac{1}{n}\left(Q_1\left(\tilde{\psi}_n^*\right) - Q_1\left(\tilde{\psi}_n\right)\right) + \frac{1}{n^{3/2}}\left(Q_2\left(\tilde{\psi}_n^*\right) - Q_2\left(\tilde{\psi}_n\right)\right),\end{aligned}$$

where terms without "$*$" do not have trimming. Notice $\{\cdot\}$ contains GEL higher order terms (Newey and Smith, 2004, Theorem 3.4), and the remaining terms describe the impact of trimming. Thus if $E[\epsilon_t^{10}] < \infty$ then the GELITT (higher order) bias is $E[Q_1(\tilde{\psi}_n^*)]/n = E[Q_1(\tilde{\psi}_n)]/n + \{E[Q_1(\tilde{\psi}_n^*)] - E[Q_1(\tilde{\psi}_n)]\}/n$, hence

$$\text{Bias}(GELITT) \;=\; \text{Bias}(GEL) \;+\; \text{Bias}(trimming).$$

**Remark 11** Result (20) shows $n^{-1}E[Q_1(\tilde{\psi}_n^*)]$ expresses higher order bias when $k_n^{(\epsilon)} \sim n/L(n)$ for slowly varying $L(n) \to \infty$, ultimately due to Karamata theory. Recall that such a trimming rate optimizes the rate of convergence.

## 4.2 Higher Order Bias and Fractile Choice

In principle a higher order mean-squared-error can be computed and this can be minimized, or at least inspected, in order to select the trimming fractile. We focus on bias $n^{-1}E[Q_1(\tilde{\psi}_n^*)]$ in order to conserve space since the (higher order) variance is a tedious function of trimmed moments, even if only based on $n^{-1/2}\tilde{\psi}_n^* + n^{-1}Q_1(\tilde{\psi}_n^*)$. See also Newey and Smith (2004, p. 234). Nevertheless, bias reveals salient features that will carry over to (higher order) mean-squared-error computation.

Recall the criterion function notation $\rho^{(i)}(u) = (\partial/\partial u)^i \rho(u)$, and now assume $\rho^{(3)}(u)$ exists, as it does for EL, CUE and ET. Independence of the errors implies that $E[Q_1(\tilde{\psi}_n^*)]$ for GELITT has the same form as $E[Q_1(\tilde{\psi}_n)]$ for GEL. The proof of the following result closely follows arguments in Newey and Smith (2004, proof of Theorem 4.2), and otherwise uses easily derived forms for tail-trimmed GEL components for GARCH model estimation. See Hill and Prokhorov (2014) for a proof.

**Theorem 4.2** *Write* $X_t \equiv x_t - E[x_t]$ *and* $S_t \equiv s_t - E[s_t]$, *and define* $\mathcal{E}_n^{(1)} \equiv E[\epsilon_{n,t}^{*2}]$, $\mathcal{E}_n^{(i)} \equiv E[(\epsilon_{n,t}^{*2} - E\left[\epsilon_{n,t}^{*2}\right])^i]$ *for* $i = 2, 3$, $\mathcal{J} = -E[X_t' S_t]$, $\Sigma_x \equiv E[X_t X_t']$, $\mathcal{H} \equiv (\mathcal{J}'\Sigma_x^{-1}\mathcal{J})^{-1}\mathcal{J}'\Sigma_x^{-1} \in \mathbb{R}^{3\times q}$, $\mathcal{P} \equiv \Sigma_x^{-1} - \Sigma_x^{-1}\mathcal{J}(\mathcal{J}'\Sigma_x^{-1}\mathcal{J})^{-1}\mathcal{J}'\Sigma_x^{-1}$ *and* $a \equiv [a_j]_{j=1}^q$ *where*

$$a_j \equiv \frac{1}{2} tr\left\{ \left(\mathcal{J}'\Sigma_x^{-1}\mathcal{J}\right)^{-1} \times E\left[ \frac{\partial^2}{\partial\theta\partial\theta'} \left\{ \left(\epsilon_t^2 - 1\right) X_{j,t} \right\} \right] \right\}.$$

*Under Assumption A,* $||E[w_t w_t']|| < \infty$ *and* $k_n^{(\epsilon)} \sim n/L(n)$ *for slowly varying* $L(n) \to \infty$*:*

$$Bias\left(\hat{\vartheta}_n\right) = \frac{1}{n} \begin{bmatrix} \frac{1}{\mathcal{E}_n^{(1)}}\mathcal{H}\left\{ \frac{\mathcal{E}_n^{(2)}}{\mathcal{E}_n^{(1)}} \left( -\left(\mathcal{E}_n^{(1)}\right)^3 a + E\left[S_t X_t' \mathcal{H} X_t\right] \right) + \frac{\mathcal{E}_n^{(3)}}{\mathcal{E}_n^{(2)}} \left(1 + \frac{\rho_3}{2}\right) E\left[X_t' X_t \mathcal{P} X_t\right] \right\} \\ \frac{1}{\mathcal{E}_n^{(2)}}\mathcal{P}\left\{ \frac{\mathcal{E}_n^{(2)}}{\mathcal{E}_n^{(1)}} \left( -\left(\mathcal{E}_n^{(1)}\right)^3 a + E\left[S_t X_t' \mathcal{H} X_t\right] \right) + \frac{\mathcal{E}_n^{(3)}}{\mathcal{E}_n^{(2)}} \left(1 + \frac{\rho_3}{2}\right) E\left[X_t' X_t \mathcal{P} X_t\right] \right\} \end{bmatrix}.$$

This implies a decomposition for Bias$(\hat{\theta}_n)$ depending on whether $\epsilon_t$ has higher moments.

**Corollary 4.3** *Under Assumption A,* $||E[w_t w_t']|| < \infty$ *and* $k_n^{(\epsilon)} \sim n/L(n)$ *for slowly varying* $L(n) \to \infty$ *we have Bias*$(\hat{\theta}_n) = \mathcal{B}_n^{(GMTTM)} + \mathcal{B}_n^{(\Sigma TT)}$, *where*

$$\mathcal{B}_n^{(GMTTM)} \equiv \frac{1}{n} \frac{\mathcal{E}_n^{(2)}}{\left(\mathcal{E}_n^{(1)}\right)^2} \mathcal{H}\left(-a + E\left[S_t X_t' \mathcal{H} X_t'\right]\right) \tag{21}$$

$$\mathcal{B}_n^{(\Sigma TT)} \equiv \frac{1}{n} \frac{\mathcal{E}_n^{(3)}}{\mathcal{E}_n^{(1)}\mathcal{E}_n^{(2)}} \mathcal{H}\left(1 + \frac{\rho_3}{2}\right) E\left[X_t' X_t \mathcal{P} X_t\right].$$

*If* $E[\epsilon_t^4] < \infty$, *such that* $\mathcal{E}^{(2)} \equiv E[(\epsilon_t^2 - 1)^2] < \infty$, *then* $\mathcal{B}_n^{(GMTTM)} = \mathcal{B}_n^{(GMM)} + \mathcal{B}_n^{(TT_{GMM})}$, *where*

$$\mathcal{B}_n^{(GMM)} \equiv \frac{1}{n} \mathcal{E}^{(2)} \mathcal{H}\left(-a + E\left[S_t X_t' \mathcal{H} X_t'\right]\right) \tag{22}$$

$$\mathcal{B}_n^{(TT_{GMM})} \equiv \frac{1}{n} \left( \frac{\mathcal{E}_n^{(2)}}{\left(\mathcal{E}_n^{(1)}\right)^2} - \mathcal{E}^{(2)} \right) \mathcal{H}\left(-a + E\left[S_t X_t' \mathcal{H} X_t'\right]\right).$$

*If* $E[\epsilon_t^6] < \infty$, *such that* $\mathcal{E}^{(3)} \equiv E[(\epsilon_t^2 - 1)^3] < \infty$, *then* $\mathcal{B}_n^{(\Sigma TT)} = \mathcal{B}_n^{(\Sigma)} + \mathcal{B}_n^{(TT_\Sigma)}$, *where*

$$\mathcal{B}_n^{(\Sigma)} \equiv \frac{1}{n} \frac{\mathcal{E}^{(3)}}{\mathcal{E}^{(2)}} \mathcal{H}\left(1 + \frac{\rho_3}{2}\right) E\left[X_t' X_t \mathcal{P} X_t\right] \text{ and } \mathcal{B}_n^{(TT_\Sigma)} \equiv \frac{1}{n} \left\{ \frac{\mathcal{E}_n^{(3)}}{\mathcal{E}_n^{(1)}\mathcal{E}_n^{(2)}} - \frac{\mathcal{E}^{(3)}}{\mathcal{E}^{(2)}} \right\} \mathcal{H}\left(1 + \frac{\rho_3}{2}\right) E\left[X_t' X_t \mathcal{P} X_t\right].$$

**Remark 12** The first term $\mathcal{B}_n^{(GMTTM)}$ in (21) is the bias associated with optimal (one-step) Generalized Method of Tail-Trimmed Moments [GMTTM], hence the estimating equations are $(\partial/\partial\theta')E[m_{n,t}^*(\theta)]|_{\theta^0}\Sigma_n^{-1}m_{n,t}(\theta)$, cf. Hansen (1982) and Hill and Renault (2010). The second term $\mathcal{B}_n^{(\Sigma TT)}$ is the bias associated with estimating the tail-trimmed estimating equation covariance. GELITT and GEL therefore have identical higher order bias forms: when $\rho_3 = -2$ (e.g. EL), or in the exactly identified case (hence $\mathcal{P} = 0$), then $\text{Bias}(\hat{\theta}_n) = \mathcal{B}_n^{(GMTTM)}$ (notice in a GARCH framework in general $E[S_t'S_tS_{i,t}] \neq 0$). Thus, under exact identification or tail-trimmed EL, it is logical to expect GELITT bias to be comparatively small. In simulation experiments, however, tail-trimmed EL performs well, but CUE leads to even lower bias in many cases, evidently due to the fact that its quadratic criterion is far easier to handle computationarlly (cf. Bonnal and Renault, 2004; Antoine, Bonnal, and Renault, 2007). See Section 8.

**Remark 13** If higher moments exist then GELITT bias decomposes into GEL bias and bias due solely to trimming. For example, if $E[\epsilon_t^4] < \infty$ such that standard asymptotics apply (since $x_t$ is square integrable), then $\mathcal{B}_n^{(GMTTM)}$ is simply bias $\mathcal{B}_n^{(GMM)}$ for optimal (one-step) GMM, plus bias $\mathcal{B}_n^{(TT_{GMM})}$ that arises from tail-trimming. Since GELITT bias can be estimated as in Newey and Smith (2004, Section 5), the bias-corrected estimator both removes higher order GEL bias (when it exists), and bias due to tail-trimming. See Section 4.3

Exactly how the amount of trimming impacts estimator's (higher order) bias depends intimately on tail decay and therefore on the tail-trimmed moments $\mathcal{E}_n^{(i)}$ as $n$ increases, as well as on the moments $E[X_tX_t']$, $E[X_t(-s_{j,t}s_t + (\partial/\partial\theta_j)s_t)]$, and $E[X_tX_t'x_{i,t}]$, and the moment functions $\mathcal{H}$ and $\mathcal{P}$. A general understanding is therefore not available, but details can be gleaned if the errors have Paretian tails. In this case, a choice of a smaller $k_n^{(\epsilon)}$ results in a smaller bias.

**Lemma 4.4** *Let $P(|\epsilon_t| \geq a) = da^{-\kappa_\epsilon}(1 + o(1))$ for $d > 0$ and $\kappa_\epsilon > 2$, let Assumption A hold, and assume $||E[w_tw_t']|| < \infty$ and $k_n^{(\epsilon)} \sim n/L(n)$ for slowly varying $L(n) \to \infty$. Then, $\mathcal{B}_n^{(GMTTM)}$ and $\mathcal{B}_n^{(\Sigma TT)}$ are small for small $k_n^{(\epsilon)}$. Therefore Bias$(\hat{\theta}_n)$ is relatively small when $k_n^{(\epsilon)}$ is small. Moreover, if higher order moments of the error term exist then the bias due to trimming is close to zero when $k_n^{(\epsilon)}$ is small.*

In order to know whether $\mathcal{B}_n^{(GMTTM)}$ and $\mathcal{B}_n^{(\Sigma TT)}$ move in the same or opposite direction as $k_n^{(\epsilon)}$ increases, we require the signs of $-a + E[S_tX_t'\mathcal{H}X_t']$ and $(1 + \rho_3/2)E[X_t'X_t\mathcal{P}X_t]$, which is difficult to determine except in special cases. If the criterion is EL such that $\rho_3 = -2$, or if there is exact identification such that $\mathcal{P} = 0$, then $\mathcal{B}_n^{(\Sigma TT)} = 0$. This gives us the next result.

**Corollary 4.5** *Let* $P(|\epsilon_t| \geq a) = da^{-\kappa_\epsilon}(1 + o(1))$ *for* $d > 0$ *and* $\kappa_\epsilon > 2$*, let Assumption A hold, and assume* $||E[w_t w_t']|| < \infty$ *and* $k_n^{(\epsilon)} \sim n/L(n)$ *for slowly varying* $L(n) \to \infty$*. Let the criterion be EL or assume* $x_t = s_t$*. Then,* $Bias(\hat{\theta}_n) = \mathcal{B}_n^{(GMTTM)}$ *monotonically decreases as* $k_n^{(\epsilon)}$ *decreases. If higher order moments of the error term exist, then bias due to trimming is monotonically closer to zero for smaller* $k_n^{(\epsilon)}$*.*

**Remark 14** Recall the dual conclusions that by first order asymptotics when $k_n^{(\epsilon)}$ is close to $\zeta n$ then the GELITT scale $\mathcal{V}_n$ is increased such that efficiency is augmented, and that $n^{-1}E[Q_1(\tilde{\psi}_n^*)]$ represents (higher order) bias. So the (higher order) bias is reduced and (first order) efficiency is augmented when, for example, $k_n^{(\epsilon)} = [\zeta n/\ln(n)]$ and $\zeta$ is small. In order for trimming to have any impact at all in terms of producing an approximately normal GELITT estimator for a particular sample when the errors are heavy tailed, clearly $k_n^{(\epsilon)} \geq 1$ for each $n$, hence $\zeta$ cannot be too small. We find $\zeta \in [.025, .075]$ works well, and in the simulation study below we focus on $\zeta = .05$, translating to $k_n^{(\epsilon)} = 1$ when $n = 100$ and $k_n^{(\epsilon)} = 2$ when $n = 250$. We also show that a variety of trimming fractile rules lead to similar results, but in general a small but rapidly increasing $k_n^{(\epsilon)}$ is best for higher order bias reduction both in theory and in practice.

## 4.3 Bias-Corrected GELITT

In general, setting $k_n^{(\epsilon)}$ small relative to $n$ will lead to a relatively small bias. There is, however, always the bias due to the higher order terms depicted in Theorem 4.1, cf. Newey and Smith (2004). We now estimate the bias using implied probabilities, but the empirical distribution may also be used. Define Jacobian, Hessian, and covariance estimators:

$$\widehat{\mathcal{J}}_n^{(\pi)} \equiv -\sum_{s=1}^{n} \hat{\pi}_{n,t}^*(\hat{\theta}_n) \left( x_t(\hat{\theta}_n) - \sum_{s=1}^{n} \hat{\pi}_{n,t}^*(\hat{\theta}_n) x_t(\hat{\theta}_n) \right) \times \left( s_t(\hat{\theta}_n) - \sum_{s=1}^{n} \hat{\pi}_{n,t}^*(\hat{\theta}_n) s_t(\hat{\theta}_n) \right)'$$

$$\hat{\Sigma}_x^{(\pi)} \equiv \sum_{s=1}^{n} \hat{\pi}_{n,t}^*(\hat{\theta}_n) \left( x_t(\hat{\theta}_n) - \sum_{s=1}^{n} \hat{\pi}_{n,t}^*(\hat{\theta}_n) x_t(\hat{\theta}_n) \right) \left( x_t(\hat{\theta}_n) - \sum_{s=1}^{n} \hat{\pi}_{n,t}^*(\hat{\theta}_n) x_t(\hat{\theta}_n) \right)'$$

$$\widehat{\mathcal{H}}_n^{(\pi)} \equiv \left( \widehat{\mathcal{J}}_n^{(\pi)\prime} \hat{\Sigma}_x^{(\pi)-1} \widehat{\mathcal{J}}_n^{(\pi)} \right)^{-1} \widehat{\mathcal{J}}_n^{(\pi)\prime} \hat{\Sigma}_x^{(\pi)-1} \text{ and } \widehat{\mathcal{P}}_n^{(\pi)} = \hat{\Sigma}_x^{(\pi)-1} - \hat{\Sigma}_x^{(\pi)-1} \widehat{\mathcal{J}}_n^{(\pi)} \widehat{\mathcal{H}}_n^{(\pi)}$$

$$\hat{a}_{j,n}^{(\pi)} \equiv \frac{1}{2} tr \left\{ \left( \widehat{\mathcal{J}}_n^{(\pi)\prime} \hat{\Sigma}_x^{(\pi)-1} \widehat{\mathcal{J}}_n^{(\pi)} \right)^{-1} \times \sum_{s=1}^{n} \hat{\pi}_{n,t}^*(\hat{\theta}_n) \frac{\partial^2}{\partial\theta\partial\theta'} \left\{ \left( \epsilon_t^2(\hat{\theta}_n) - 1 \right) s_{j,t}(\hat{\theta}_n) \right\} \right\} \text{ and } \hat{a}_n^{(\pi)} = \left[ \hat{a}_{j,n}^{(\pi)} \right]_{j=1}^{3}$$

$$\widehat{\mathcal{E}}_{1,n}^{(\pi)} \equiv \sum_{t=1}^{n} \hat{\pi}_{n,t}^*(\hat{\theta}_n) \hat{\epsilon}_{n,t}^{*2}(\hat{\theta}_n) \text{ and } \widehat{\mathcal{E}}_{i,n}^{(\pi)} \equiv \sum_{t=1}^{n} \hat{\pi}_{n,t}^*(\hat{\theta}_n) \left( \hat{\epsilon}_{n,t}^{*2}(\hat{\theta}_n) - \widehat{\mathcal{E}}_{1,n}^{(\pi)} \right)^i \text{ for } i = 2, 3.$$

Define the bias estimator components:

$$\hat{\mathcal{B}}_n^{(GMTTM)} \equiv \frac{1}{n}\frac{\widehat{\mathcal{E}}_{2,n}^{(\pi)}}{\left(\widehat{\mathcal{E}}_{1,n}^{(\pi)}\right)^2}\widehat{\mathcal{H}}_n^{(\pi)}\left(-\hat{a}_n^{(\pi)} + \frac{1}{n}\sum_{t=1}^{n} S_t X_t' \widehat{\mathcal{H}}_n^{(\pi)} X_t'\right)$$

$$\hat{\mathcal{B}}_n^{(\Sigma TT)} \equiv \frac{1}{n}\frac{\widehat{\mathcal{E}}_{3,n}^{(\pi)}}{\widehat{\mathcal{E}}_{1,n}^{(\pi)}\widehat{\mathcal{E}}_{2,n}^{(\pi)}}\widehat{\mathcal{H}}_n^{(\pi)}\left(1+\frac{\rho_3}{2}\right)\frac{1}{n}\sum_{t=1}^{n} X_t' X_t \widehat{\mathcal{P}}_n^{(\pi)} X_t.$$

The GELITT bias estimator is $\hat{\mathcal{B}}_n(\hat{\theta}_n) = \hat{\mathcal{B}}_n^{(GMTTM)} + \hat{\mathcal{B}}_n^{(\Sigma TT)}$, and in the case of EL or exact identification we use $\hat{\mathcal{B}}(\hat{\theta}_n) = \hat{\mathcal{B}}_n^{(GMTTM)}$. The bias-corrected GELITT estimator is then:

$$\hat{\theta}_n^{(bc)} = \hat{\theta}_n - \hat{\mathcal{B}}_n(\hat{\theta}_n).$$

The estimator $\hat{\theta}_n^{(bc)}$ has the same limit distribution as $\hat{\theta}_n$, and is higher order unbiased provided $k_n^{(\epsilon)} \sim n/L(n)$.

**Theorem 4.6** *Under Assumption A,*$||E[w_t w_t']|| < \infty$ *and* $k_n^{(\epsilon)} \sim n/L(n)$ *for slowly varying* $L(n) \to \infty$ *we have* $Bias(\hat{\theta}_n^{(bc)}) = 0$ *and* $\mathcal{V}_n^{1/2}(\hat{\theta}_n^{(bc)} - \theta^0) \xrightarrow{d} N(I_3)$.

# 5 Robust Testing

We now use GELITT theory to construct a scale estimator, and robust versions of tests of over-identifying restrictions. A natural estimator of the GELITT scale $\mathcal{V}_n \equiv n\mathcal{J}_n'\Sigma_n^{-1}\mathcal{J}_n$ is $\hat{\mathcal{V}}_n(\theta) \equiv n\widehat{\mathcal{J}}_n(\theta)'\hat{\Sigma}_n^{-1}(\theta)\widehat{\mathcal{J}}_n(\theta)$ where

$$\widehat{\mathcal{J}}_n(\theta) \equiv -\frac{1}{n}\sum_{t=1}^{n}\left(x_{n,t}^*(\theta) - \widehat{\mathcal{X}}_n(\theta)\right)\left(s_t(\theta) - \widehat{\mathcal{S}}_n(\theta)\right)' \quad \text{and} \quad \hat{\Sigma}_n(\theta) \equiv \frac{1}{n}\sum_{t=1}^{n}\hat{m}_{n,t}^*(\theta)\hat{m}_{n,t}^*(\theta)',$$

with $\widehat{\mathcal{X}}_n(\theta) \equiv 1/n\sum_{t=1}^{n} x_{n,t}^*(\theta)$ and $\widehat{\mathcal{S}}_n(\theta) \equiv 1/n\sum_{t=1}^{n} s_t(\theta)$. In the case of exact identification a more compact estimator is possible since $\mathcal{J}_n = -E[(s_t - E[s_t]) \times (s_t - E[s_t])']$, and by dominated convergence and independence $\Sigma_n \sim E[(\hat{\epsilon}_{n,t}^{*4} - 1)] \times \mathcal{J}_n$, hence $\mathcal{V}_n \sim n\mathcal{J}_n/(E[\hat{\epsilon}_{n,t}^{*4}] - 1)$. In this case we can use $\hat{\mathcal{V}}_n(\theta) = n\widehat{\mathcal{J}}_n(\theta)/(1/n\sum_{s=1}^{n}\hat{\epsilon}_{n,t}^{*4}(\theta) - 1)$. Efficient versions of these estimators substitute the empirical probabilities $1/n$ for the implied probabilities $\hat{\pi}_{n,t}^*(\hat{\theta}_n)$: see Section 6.

**Theorem 5.1** *Under Assumption A,* $\hat{\mathcal{V}}_n(\tilde{\theta}_n) = \mathcal{V}_n(1 + o_p(1))$ *for any* $\tilde{\theta}_n \xrightarrow{p} \theta^0$.

Next, recall the GEL weights have two parts $x_t(\theta) = [s_t(\theta)', w_t(\theta)']$, so that the proposed over-identifying moment conditions are based on $w_t(\theta) : \Theta \to \mathbb{R}^{q-3}$. It is therefore interesting to test the assumption $E[(\epsilon_t^2 - 1)w_t] = 0$ without imposing higher moments on $\epsilon_t$ or $w_t$. A theory for heavy tail robust moment condition tests is presented in Hill (2012) and Hill and Aguilar (2013), but those papers treat the plug-in estimator as not necessarily using those moment conditions for estimation, and they do not exploit empirical information about the data generating process for efficient moment estimation. Define the GELTT criterion function $\hat{Q}_n(\theta, \lambda) \equiv 1/n \sum_{t=1}^n \rho(\lambda' \hat{m}_{n,t}^*(\theta))$. Recalling $\rho(0) = 0$, the heavy tail robust trilogy test statistics are Likelihood Ratio $\mathcal{LR}_n = 2n\hat{Q}_n(\hat{\theta}_n, \hat{\lambda}_n)$, score $\mathcal{S}_n = n\hat{m}_n^*(\hat{\theta}_n)'\hat{\Sigma}_n^{-1}(\hat{\theta}_n)\hat{m}_n^*(\hat{\theta}_n)$ and Lagrange Multiplier $\mathcal{LM}_n = n\hat{\lambda}_n'\hat{\Sigma}_n(\hat{\theta}_n)^{-1}\hat{\lambda}_n$. The score statistic $\mathcal{S}_n$ is identical in form to the heavy tail robust test statistic in Hill and Aguilar (2013), while all three statistics are equivalent under the null with probability approaching one. See Smith (1997) for original contributions in the GEL literature, cf. Hansen (1982).

**Theorem 5.2** *Under Assumption A and $q > 3$ with $E[(\epsilon_t^2 - 1)w_t] = 0$ we have $\mathcal{LR}_n, \mathcal{S}_n, \mathcal{LM}_n \xrightarrow{d} \chi^2(q-3)$ hence all three statistics are asymptotically equivalent under the null. Further, if $E[(\epsilon_t^2 - 1)w_t] \neq 0$ then $\mathcal{LR}_n, \mathcal{S}_n, \mathcal{LM}_n \xrightarrow{p} \infty$.*

A classical Wald statistic for linear or nonlinear restrictions is also easily constructed. Let $\mathcal{R} : \Theta \to \mathbb{R}^J$ for $J \geq 1$ be a continuous, differentiable function such that $\mathcal{D}(\theta) \equiv (\partial/\partial\theta)\mathcal{R}(\theta)$ is continuous and has full column rank, and $\varphi \in \mathbb{R}^J$. The null hypothesis is $\mathcal{R}(\theta^0) = \varphi$, and the Wald statistic is $\mathcal{W}_n \equiv (\mathcal{R}(\hat{\theta}_n) - \varphi)'[\mathcal{D}(\hat{\theta}_n)\hat{\mathcal{V}}_n(\hat{\theta}_n)^{-1}\mathcal{D}(\hat{\theta}_n)']^{-1}(\mathcal{R}(\hat{\theta}_n) - \varphi)$.

**Theorem 5.3** *Under Assumption A and $\mathcal{R}(\theta^0) = 0$ we have $\mathcal{W}_n \xrightarrow{d} \chi^2(J)$, and if $\mathcal{R}(\theta^0) \neq 0$ then $\mathcal{W}_n \xrightarrow{p} \infty$.*

**Remark 15** In a more general setting, standard asymptotic tests for GMM and GEL estimators are overly sized in small samples (see, e.g., Hall and Horowitz, 1996; Inoue and Shintani, 2006), and bootstrap methods are possibly invalid when over-identifying restrictions are present (Hall and Horowitz, 1996). Various bootstrap techniques have been suggested to improve on the small sample performance of Wald tests and tests of over-identification (e.g., Hall and Horowitz, 1996), and for QML inference for GARCH models with heavy tailed errors (e.g. Hall and Yao, 2003). The latter is key since the bootstrap is valid for thin tailed and exceptionally heavy tailed data (i.e. heavier than a power law), but not necessarily when the data have power law tails and unbounded higher moments (see Hall, 1990). In the present setting under the null, our Wald statistic is, to a first order approximation, a quadratic form of a self-standardized sum of tail-trimmed estimating equations: $\mathcal{W}_n = \mathcal{D}\mathcal{H}_n\Sigma_n^{1/2}\mathcal{Z}_n\mathcal{Z}_n'\Sigma_n^{1/2}\mathcal{H}_n'\mathcal{D}' + o_p(1)$ where $\mathcal{D} = \mathcal{D}(\theta^0)$, $\mathcal{H}_n =$

$(\mathcal{J}_n'\Sigma_n^{-1}\mathcal{J}_n)^{-1}\mathcal{J}_n'\Sigma_n^{-1}$ and $\mathcal{Z}_n = [\mathcal{Z}_{i,n}]_{i=1}^q = \Sigma_n^{-1/2} n^{-1/2} \sum_{t=1}^n m_{n,t}^*$. Although self-standardization ensures standard asymptotics since $\mathcal{Z}_n \xrightarrow{d} N(0, I_q)$, this is hairline: the self-standardized tail-trimmed equations $\mathcal{Z}_{i,n}$ have a unit variance $E[\mathcal{Z}_{i,n}^2] = 1$, but asymptotically have unbounded moments greater than two when $E[\epsilon_t^4] = \infty$ since $E|\mathcal{Z}_{i,n}|^{2+\iota} \to \infty$ for $\iota > 0$. Whether bootstrap techniques are valid in this case is unknown, and therefore not tackled in this paper.

# 6 Robust and Efficient Moment Estimation

In this section we estimate a set of moments $E[g_t(\theta^0)]$, where $g_t = [g_{i,t}]_{i=1}^h : \Theta \to \mathbb{R}^h$ for $h \geq 1$ is $\Im_t$-measurable, integrable, stationary, ergodic, *a.s.* continuous and differentiable on $\Theta$-*a.e.* Implicitly $g_t$ may depend on other parameters although we do not express it. Examples are the Jacobian and covariance matrices used for test statistic constructions; unconditional moments of $y_t$, $\sigma_t^2$ or $\epsilon_t$; conditional moments like the *expected shortfall* of a financial asset; and tail moments including those used to characterize tail indices (see Hill, 2010, for theory and references). We show that the use of $\hat{\pi}_{n,t}^*(\theta)$, rather than the empirical probabilities $1/n$, leads to a non-trivial efficiency improvement for a heavy tail robust moment estimator, mimicking classic results in Back and Brown (1993), Brown and Newey (1998) and Smith (2011).

Consider heavy tail robust estimation under the premise that $E[g_{i,t}^2(\theta^0)] < \infty$ is unknown. Define tail specific observations $g_{i,t}^{(-)}(\theta) \equiv g_{i,t}(\theta) I(g_{i,t}(\theta) < 0)$ and $g_{i,t}^{(+)}(\theta) \equiv g_{i,t}(\theta) I(g_{i,t}(\theta) \geq 0)$, let $g_{i,(j)}^{(\cdot)}(\theta)$ be the order statistics $g_{i,(1)}^{(+)}(\theta) \geq g_{i,(2)}^{(+)}(\theta) \geq \cdots$ and $g_{i,(1)}^{(-)}(\theta) \leq g_{i,(2)}^{(-)}(\theta) \leq \cdots$ and let $k_{1,i,n}^{(g)}$ and $k_{2,i,n}^{(g)}$ be intermediate order statistics. Similar to methods in Hill (2012, 2015b) and Hill and Aguilar (2013), for heavy tail robust estimation we tail-trim $g_{i,t}$:

$$\hat{g}_{i,n,t}^*(\theta) \equiv g_{i,t}(\theta) \hat{I}_{i,n,t}^{(g)}(\theta) = g_{i,t}(\theta) I\left( g_{i,(k_{1,i,n}^{(g)})}^{(-)}(\theta) \leq g_{i,t}(\theta) \leq g_{i,(k_{2,i,n}^{(g)})}^{(+)}(\theta) \right).$$

The uniform (or flat) and profile weighted sample mean estimators are

$$\overline{\hat{g}}_n^*(\theta) \equiv \frac{1}{n} \sum_{t=1}^n \hat{g}_{n,t}^*(\theta) \quad \text{and} \quad \overline{\hat{g}}_n^{*(\pi)}(\theta) \equiv \sum_{t=1}^n \hat{\pi}_{n,t}^*(\theta) \hat{g}_{n,t}^*(\theta).$$

In the tail-trimmed CUE case we can use the profile formulas (15)-(17) to deduce that $\overline{\hat{g}}_n^{*(\pi)}(\theta)$ is a sample version of an unbiased minimum variance estimator $E[\hat{g}_{n,t}^*(x)]$, that is $\overline{\hat{g}}_n^{*(\pi)}(\theta) = \overline{\hat{g}}_n^*(\theta) - \overline{\hat{m}}_n^*(\theta)' \check{\Sigma}_n(\theta)^{-1} \times \widehat{cov}(\hat{g}_{n,t}^*(\theta), \hat{m}_{n,t}^*(\theta))$, where $\widehat{cov}(a,b) \equiv 1/n \sum_{t=1}^n a_t \{b_t - \bar{b}\}$. Thus, $\overline{\hat{g}}_n^{*(\pi)}(\theta)$ is asymptotically best in the class of estimators with the form $\overline{\hat{g}}_n^*(\theta) - \overline{\hat{m}}_n^*(\theta)' A$. See Bonnal and Renault (2004, Corollary 3.5).

The asymptotic theory for $\overline{\hat{g}}_n^{*(\pi)}(\theta)$ requires the non-stochastic positive functions $\{c_{1,i,n}^{(g)}(\theta), c_{2,i,n}^{(g)}(\theta)\}$ that $g_{i,(k_{1,i,n}^{(g)})}^{(-)}(\theta)$ and $g_{i,(k_{2,i,n}^{(g)})}^{(+)}(\theta)$ estimate:

$$P\left(g_{i,t}^{(-)}(\theta) < -c_{1,i,n}^{(g)}(\theta)\right) = \frac{k_{1,i,n}^{(g)}}{n} \text{ and } P\left(g_{i,t}^{(+)}(\theta) > c_{2,i,n}^{(g)}(\theta)\right) = \frac{k_{2,i,n}^{(g)}}{n}.$$

Define a deterministically trimmed version

$$g_{i,n,t}^{*}(\theta) \equiv g_{i,t}(\theta) I_{i,n,t}^{(g)}(\theta) = g_{i,t}(\theta) I\left(-c_{1,i,n}^{(g)}(\theta) \leq g_{i,t}(\theta) \leq c_{2,i,n}^{(g)}(\theta)\right),$$

and associated Jacobian, covariance and scale matrices

$$\Upsilon_n \equiv \frac{1}{n}\sum_{s,t=1}^{n} E\left[\left(g_{n,s}^{*} - E\left[g_{n,s}^{*}\right]\right)\left(g_{n,t}^{*} - E\left[g_{n,t}^{*}\right]\right)'\right] \text{ and } \Gamma_n \equiv \frac{1}{n}\sum_{s,t=1}^{n} E\left[g_{n,s}^{*} m_{n,t}^{*\prime}\right]$$

$$G_{i,j,n} \equiv \frac{\partial}{\partial \theta_j} E\left[g_{i,t}(\theta) I_{i,n,t}^{(g)}(\theta)\right] |_{\theta^0}$$

$$\begin{aligned}\mathfrak{V}_n \equiv \Upsilon_n &- G_n'\Sigma_n^{-1}\mathcal{J}_n\left(\mathcal{J}_n'\Sigma_n^{-1}\mathcal{J}_n\right)^{-1}\Gamma_n' - \Gamma_n\left(\mathcal{J}_n'\Sigma_n^{-1}\mathcal{J}_n\right)^{-1}\mathcal{J}_n'\Sigma_n^{-1}G_n \\ &+ G_n'\left(\mathcal{J}_n'\Sigma_n^{-1}\mathcal{J}_n\right)^{-1}G_n - \Gamma_n\mathcal{P}_n\Gamma_n'.\end{aligned}$$

Notice $\Gamma_n = 1/n\sum_{s\geq t=1}^{n} E[g_{n,s}^{*} m_{n,t}^{*\prime}]$ by the martingale difference property of $m_{n,t}^{*}$.

Asymptotic theory is again expedited if we assume $g_{i,t}(\theta)$ have power law tails when $E[g_{i,t}^2(\theta)] = \infty$. Define $\Theta_{2,i}^{(g)} = \{\theta \in \Theta : E[g_{i,t}^2(\theta)] = \infty\}$.

**Assumption B.** *If* $\sup_{\theta\in\Theta} E[g_{i,t}^2(\theta)] = \infty$ *then* $g_{i,t}(\theta)$ *has for each* $t$ *a common power-law tail* $P(|g_{i,t}(\theta)| > m) = d_i^{(g)}(\theta) c^{-\kappa_i^{(g)}(\theta)}(1 + o(1))$ *where* $\inf_{\theta\in\Theta_{2,i}^{(g)}} \kappa_i^{(g)}(\theta) > 0$, $\kappa_i^{(g)} = \kappa_i^{(g)}(\theta^0) > 1$, $\inf_{\theta\in\Theta_{2,i}^{(g)}} d_i^{(g)}(\theta) > 0$ *and* $o(1)$ *is not a function of* $\theta$.

**Theorem 6.1** *Let* $\{y_t, \epsilon_t, \sigma_t^2, w_t, g_t\}$ *satisfy Assumptions A and B, and assume* $n^{1/2}\mathfrak{V}_n^{-1/2}\{E[g_{n,t}^{*}] - E[g_t]\} \to 0$. *Then* $n_n^{1/2}\mathfrak{V}_n^{-1/2}\{\overline{\hat{g}}_n^{*(\pi)}(\hat{\theta}_n) - E[g_t]\} \xrightarrow{d} N(0, I_h)$. *If* $\max\{\kappa_1^{(g)}, \kappa_2^{(g)}\} \geq 2$ *and* $k_{i,n}^{(g)} \to \infty$ *at a slowly varying rate then* $n^{1/2}\mathfrak{V}_n^{-1/2}\{E[g_{n,t}^{*}] - E[g_t]\} \to 0$ *holds.*

**Remark 16** The scale $\mathfrak{V}_n$ has a classic form, denoting long-run dispersion of $g_{n,t}^{*}$ by $\Upsilon_n$, amplified by sampling error due to $\hat{\theta}_n$, and corrected by the efficiency improvement afforded by $\hat{\pi}_{n,t}^{*}(\hat{\theta}_n)$. In the nonparametric case $g_t(\theta) = g_t$ and we have $G_n = 0$. Hence the scale reduces to $\mathfrak{V}_n = \Upsilon_n - \Gamma_n\mathcal{P}_n\Gamma_n'$ revealing a pure efficiency gain by exploiting the profile probabilities with over-identification rather than empirical probabilities (see Antoine, Bonnal, and Renault, 2007;

Smith, 2011). Under exact identification $\mathcal{P}_n = 0$, so of course there is no efficiency gain when $g_t(\theta) = g_t$.

**Remark 17** Consistent estimators of $G_n$, $\Upsilon_n$ and $\Gamma_n$ are easy to derive as in Section 5. A quadratic form $\overline{\hat{g}}_n^{*(\pi)}(\hat{\theta}_n)'\hat{\mathfrak{V}}_n^{-1}\overline{\hat{g}}_n^{*(\pi)}(\hat{\theta}_n)$ can then be used to test $E[g_t] = 0$. If we simply use $\overline{\hat{g}}_n^{*}(\theta)$ then
$\overline{\hat{g}}_n^{*}(\hat{\theta}_n)'\hat{\Upsilon}_n^{-1}(\hat{\theta}_n)\overline{\hat{g}}_n^{*}(\hat{\theta}_n)$ with a consistent HAC estimator $\hat{\Upsilon}_n(\hat{\theta}_n)$ is identical to the tail-trimmed moment condition test statistic in Hill and Aguilar (2013).

**Remark 18** Consider the scalar case $h = 1$ for simplicity. The identification assumption $n^{1/2}\mathfrak{V}_n^{-1/2}\{E[g_{n,t}^*] - E[g_t]\} \to 0$ is superfluous if tails are not too heavy and trimming is fairly light. Otherwise, the assumption implies that we assume asymmetric trimming is set such that $E[g_{n,t}^*] \to E[g_t]$ rapidly enough for asymptotic unbiasedness in the limit distribution of $\overline{\hat{g}}_n^{*(\pi)}(\hat{\theta}_n)$. An alternative method is to use intrinsically easier symmetric trimming $\hat{g}_{i,n,t}^*(\theta) = g_{i,t}(\theta)I(|g_{i,t}(\theta)| \leq g^{(a)}_{i,(k_{i,n}^{(g)})}(\theta))$ coupled with a bias correction estimator such that identification $n^{1/2}\mathfrak{V}_n^{-1/2}\{E[g_{n,t}^*] - E[g_t]\} \to 0$ is not needed. See Section 7, and see Hill (2015b) for further results and references.

**Remark 19** If each $E[g_{i,t}^2] < \infty$ then trimming for $g_t$ is not required. We can, however, still use the GELITT profiles for a more efficient moment estimator since $n^{1/2}\mathfrak{V}_n^{-1/2}(\sum_{t=1}^n \hat{\pi}_{n,t}(\hat{\theta}_n)g_t(\hat{\theta}_n) - E[g_t]) \xrightarrow{d} N(0, I_h)$, where $G_{i,j,n} \equiv (\partial/\partial\theta_j)E[g_{i,t}(\theta)]|_{\theta^0}$, $\Upsilon_n \equiv 1/n\sum_{s,t=1}^n E[g_s g_t']$, $\Gamma_n(\theta) \equiv 1/n\sum_{s,t=1}^n E\left[g_s m_{n,t}^{*\prime}\right]$ and so on.

**Remark 20** The profiles can be exploited for an efficient GELITT scale estimator $\hat{\mathcal{V}}_n^{(\pi)}(\theta) \equiv n\widehat{\mathcal{J}}_n^{(\pi)}(\theta)'\hat{\Sigma}_n^{(\pi)}(\theta)^{-1}\widehat{\mathcal{J}}_n^{(\pi)}(\theta)$. Define $\hat{X}_n^{(\pi)}(\theta) \equiv \sum_{s=1}^n \hat{\pi}_{n,t}^*(\theta)x_{n,t}^*(\theta)$, $\widehat{\mathcal{S}}_n^{(\pi)}(\theta) \equiv \sum_{s=1}^n \hat{\pi}_{n,t}^*(\theta)s_t(\theta)$ and $\widehat{\mathcal{E}}_n^{(\pi)2}(\theta) \equiv \sum_{s=1}^n \hat{\pi}_{n,t}^*(\theta)\hat{\epsilon}_{n,t}^{*2}$. Define equations $\hat{m}_{n,t}^*(\theta) \equiv (\hat{\epsilon}_{n,t}^{*2} - \widehat{\mathcal{E}}_n^{*2}(\theta))x_{n,t}^*(\theta)$. Then use $\widehat{\mathcal{J}}_n^{(\pi)}(\theta) \equiv -\sum_{s=1}^n \hat{\pi}_{n,t}^*(\theta)(x_{n,t}^*(\theta) - \widehat{\mathcal{X}}_n^{(\pi)}(\theta)) \times (s_t(\theta) - \widehat{\mathcal{S}}_n^{(\pi)}(\theta))'$ and $\hat{\Sigma}_n^{(\pi)}(\theta) \equiv \sum_{t=1}^n \hat{\pi}_{n,t}^*(\theta)\hat{m}_{n,t}^*(\theta)\hat{m}_{n,t}^*(\theta)'$.

# 7 Example - Expected Shortfall

There are many interesting examples of efficient and robust moment estimation for GARCH processes. We present one concerning the *expected shortfall* [ES] of an asset, which has not evidently been treated in the GEL literature.

Recall the ES of $y_t \in \mathbb{R}$ with $E|y_t| < \infty$ is the conditional expected loss $ES_\alpha \equiv -E[y_t|y_t \leq q_\alpha]$ $= -\alpha^{-1}E[y_t I(y_t \leq q_\alpha)] > 0$, where $-q_\alpha > 0$ is the Value-at-Risk for risk level $\alpha \in (0,1)$. If $E[y_t^2]$ $< \infty$ then an efficient and asymptotically normal estimator is based on the GELITT profiles: $\widehat{ES}_{n,\alpha}^{(\pi)}(\theta) \equiv -\alpha^{-1}\sum_{t=1}^n \hat{\pi}_{n,t}^*(\theta)y_t I(y_t \leq \hat{q}_{n,\alpha})$ where $\hat{q}_{n,\alpha}$ consistently estimates $q_\alpha$. Hill (2015b)

uses tail-trimming to deliver asymptotically normal and unbiased ES estimators for possibly infinite variance processes. We extend that theory here to allow for profile weighting.[12] We first apply Theorem 6.1 to a biased, profile-weighted tail-trimmed ES estimator, and then present a new result for a bias-corrected estimator.

## 7.1 Profile-Weighted Tail-Trimmed ES

The heavy tail robust profile-weighted version is

$$\widehat{ES}_{n,\alpha}^{*(\pi)} \equiv -\frac{1}{\alpha}\sum_{t=1}^{n}\hat{\pi}_{n,t}^{*}y_t I\left(y_{(k_n^{(y)})}^{(-)} \leq y_t \leq y_{[\alpha n]}\right) \text{ where } \hat{\pi}_{n,t}^{*} \equiv \hat{\pi}_{n,t}^{*}(\hat{\theta}_n),$$

where $y_t^{(-)} \equiv y_t I(y_t < 0)$, $k_n^{(y)} \to \infty$, and $k_n^{(y)}/n \to 0$. Trivially $y_{(k_n^{(y)})}^{(-)} < y_{[\alpha n]}$ $a.s.$ as $n \to \infty$ since $k_n^{(y)}/n \to 0$, so assume $n$ is large enough that $y_{(k_n^{(y)})}^{(-)} < y_{[\alpha n]}$ $a.s.$ Define positive deterministic thresholds $\{l_n^{(y)}\}$ by $P(-l_n^{(y)} \leq y_t) = k_n^{(y)}/n$, hence by dominated convergence:

$$-\frac{1}{\alpha}E\left[y_{n,t}^{*}\right] = -\frac{1}{\alpha}E\left[y_t I\left(-l_n^{(y)} \leq y_t \leq q_\alpha\right)\right] \to ES_\alpha.\text{where } y_{n,t}^{*} \equiv y_t I\left(-l_n^{(y)} \leq y_t \leq q_\alpha\right).$$

It is easy to alter Theorem 6.1 to allow for a central order upper bound $y_{[\alpha n]}$, since under Assumption A $y_t$ is stationary and geometrically $\beta$-mixing (e.g. Nelson, 1990; Carrasco and Chen, 2002), hence $y_{[\alpha n]} = q_\alpha + O_p(1/n^{1/2})$. See, e.g., Mehra and Rao (1975). Define

$$\Upsilon_n \equiv \frac{1}{n}\sum_{s,t=1}^{n} E\left[\left(y_{n,s}^{*} - E\left[y_{n,s}^{*}\right]\right)\left(y_{n,t}^{*} - E\left[y_{n,t}^{*}\right]\right)\right] \text{ and } \Gamma_n \equiv \frac{1}{n}\sum_{s\geq t=1}^{n} E\left[y_{n,s}^{*}m_{n,t}^{*\prime}\right]$$

$$\mathfrak{V}_n \equiv \Upsilon_n - \Gamma_n\mathcal{P}_n\Gamma_n' \text{ and } \mathcal{B}_n \equiv -\frac{1}{\alpha}E\left[y_t I\left(y_t \leq -l_n^{(y)}\right)\right].$$

As long as $y_t$ satisfies Assumption A, and since Assumption B is superfluous by measurability, it follows by Theorem 6.1

$$\frac{n^{1/2}}{\mathfrak{V}_n^{1/2}}\left\{\widehat{ES}_{n,\alpha}^{*(\pi)} + \mathcal{B}_n - ES_\alpha\right\} \xrightarrow{d} N\left(0, \alpha^{-2}\right).$$

The scale form $\mathfrak{V}_n$follows since $\hat{\theta}_n$ only enters $\hat{\pi}_{n,t}^{*}$. Thus, we can only achieve an efficiency gain

[12]We use the central order statistic $\hat{q}_{n,\alpha} = y_{[\alpha n]}$ for simplicity, similar to Chen (2008) and Hill (2015b). See Scaillet (2004) and Linton and Xiao (2013) for smoothed kernel estimators. See Linton and Xiao (2013) for non-standard limit theory for conventional ES estimators when $y_t$ has a regularly varying distribution tail with index $\kappa \in (1, 2)$.

if over-identifying conditions are used, since otherwise $\mathfrak{V}_n = \Upsilon_n$ and hence $\widehat{ES}^{*(\pi)}_{n,\alpha}$ has the same asymptotic properties as the flat-weighted estimator of Hill (2015b).

## 7.2 Bias-Corrected Profile-Weighted Tail-Trimmed ES

Unless $\kappa_1 \geq 2$, and trimming is light $k_n^{(y)} = O(\ln(n))$, the bias does not vanish: $(n^{1/2}/\mathfrak{V}_n^{1/2})|\mathcal{B}_n| \to \infty$ (Hill, 2015b, Section 1). Hill (2015b) presents a bias corrected version of the flat weighted ES estimator $\widehat{ES}^{*}_{n,\alpha} \equiv \alpha^{-1}n^{-1}\sum_{t=1}^{n} y_t I(y^{(-)}_{(k_n^{(y)})} \leq y_t \leq y_{[\alpha n]})$. The same methods and theory can be easily applied to $\widehat{ES}^{*(\pi)}_{n,\alpha}$ in view of $n^{3/2}||\Sigma_n||^{1/2}$-consistency of the profiles $\hat{\pi}^*_{n,t}$, cf. Lemma A.12 in the appendix. We present the bias correction here and refer the reader to Hill (2015b) for theory details on the bias form.[13]

Let $\kappa_1$ be the left tail index, $P(y_t \leq -c) = d_1 c^{-\kappa_1}(1 + o(1))$, cf. Basrak, Davis, and Mikosch (2002). The expected shortfall exists only if $\kappa_1 > 1$ (for risk measure theory in the very heavy tailed case, see, e.g. Garcia, Renault, and Tsafack, 2007; Ibragimov, 2009). Hill (1975)'s estimator of $\kappa_1$ is $\hat{\kappa}_{1,m_n} \equiv (1/m_n \sum_{i=1}^{m_n} \ln(y^{(-)}_{(i)}/y^{(-)}_{(m_n)}))^{-1}$, where $\{m_n\}$ is an intermediate order sequence. The bias estimator is

$$\hat{\mathcal{B}}_n \equiv -\frac{1}{\alpha}\left(\frac{\hat{\kappa}_{1,m_n}}{\hat{\kappa}_{1,m_n} - 1}\frac{k_n^{(y)}}{n} y^{(-)}_{(k_n^{(y)})}\right)$$

and the bias-corrected estimator is $\widehat{ES}^{(bc)(\pi)}_{n,\alpha} \equiv \widehat{ES}^{*(\pi)}_{n,\alpha} + \hat{\mathcal{B}}_n$. If $y_t$ were known to be symmetrically distributed, then $\kappa_1$ can be estimated using $|y_t|$, allowing for more observations and therefore a sharper estimator. As in Hill (2015b), we select $m_n$ from a window of such fractiles such that $\widehat{ES}^{(bc)(\pi)}_{n,\alpha}$ is close to the asymptotically unbiased untrimmed estimator, provided $\hat{\kappa}_{1,m_n} > 1$. Write $m_n(\xi) \equiv [\xi m_n]$ where $0 < \underline{\xi} \leq \xi \leq \bar{\xi}$ for some chosen $\{\underline{\xi}, \bar{\xi}\} \in (0, \infty)$, and write $\hat{\mathcal{B}}_n(\xi)$ to show dependence on $\xi$. Then the "optimally" bias corrected estimator is $\widehat{ES}^{(bc*)(\pi)}_{n,\alpha} \equiv \widehat{ES}^{*(\pi)}_{n,\alpha} + \hat{\mathcal{B}}_n(\hat{\xi}_n)$, where

$$\hat{\xi}_n = \underset{\underline{\xi}\leq\xi\leq\bar{\xi}:\hat{\kappa}_{1,m_n(\xi)}>1}{\arg\inf} \left\{\left|\widehat{ES}^{*(\pi)}_{n,\alpha} + \hat{\mathcal{B}}_n(\xi) - \widetilde{ES}^{(\pi)}_{n,\alpha}\right|\right\}$$

with untrimmed $\widetilde{ES}^{(\pi)}_{n,\alpha} \equiv -\alpha^{-1}\sum_{t=1}^{n} \hat{\pi}^*_{n,t} y_t I(y_t \leq y_{[\alpha n]})$. As long as $y_t$ satisfies a second order power law property in order to ensure $\hat{\kappa}_{1,m_n} = \kappa_1 + O_p(1/m_n^{1/2})$, and $m_n/k_n^{(y)} \to \infty$, then $\hat{\kappa}_{1,m_n}$ does not affect asymptotics (similar to Hill, 2015b, Theorem 2.2).

Hill (2015b) only considers a flat weighted version of $\widehat{ES}^{(bc*)(\pi)}_{n,\alpha}$. The bias estimator $\hat{\mathcal{B}}_n(\hat{\xi}_n)$, however, may exhibit enough sampling error that $\widehat{ES}^{*(\pi)}_{n,\alpha}$ is closer to $\widetilde{ES}^{(\pi)}_{n,\alpha}$ than is the bias

[13] See also Peng (2001), cf. Csörgo, Horváth, and Mason (1986), who evidently originally proposed a different version of this bias-correction for iid data.

corrected $\widehat{ES}_{n,\alpha}^{(bc*)(\pi)}$. In practice we therefore use whichever estimator is best:

$$\begin{aligned}\widehat{ES}_{n,\alpha}^{(obc)(\pi)} \equiv\ & \widehat{ES}_{n,\alpha}^{(bc*)(\pi)} I\left(\left|\widehat{ES}_{n,\alpha}^{(bc*)(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right| < \left|\widehat{ES}_{n,\alpha}^{*(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right|\right) \\ & + \widehat{ES}_{n,\alpha}^{*(\pi)} I\left(\left|\widehat{ES}_{n,\alpha}^{(bc*)(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right| > \left|\widehat{ES}_{n,\alpha}^{*(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right|\right).\end{aligned} \tag{23}$$

In Theorem 7.1, we show that if $k_n^{(y)} = o((\ln(n))^a)$ for some $a > 0$, then $\widehat{ES}_{n,\alpha}^{*(\pi)}$ is chosen with probability approaching one *if and only if* $\kappa_1 \geq 2$, since only then is $\widehat{ES}_{n,\alpha}^{*(\pi)}$ unbiased in its limit distribution.

The limit distribution of the flat weight ES estimator is based on the joint asymptotic behavior of the tail-trimmed $y_t I(-l_n^{(y)} \leq y_t \leq q_\alpha)$ and the tail process $\{I(y_t \leq -l_n^{(y)}) - E[I(y_t \leq -l_n^{(y)})]\}$ which governs the order statistic $y_{(k_n^{(y)})}^{(-)}$ in the bias estimator $\hat{\mathcal{B}}_n$. Under profile weighting clearly $\hat{\pi}_{n,t}^* = \hat{\pi}_{n,t}^*(\hat{\theta}_n)$, and therefore $m_{n,t}^*$, will also affect asymptotics. In addition to the long-run variance $\Upsilon_n$ and covariance $\Gamma_n$, we therefore need the following. Recall $\Sigma_n \equiv E[m_{n,t}^* m_{n,t}^{*\prime}]$, define variables:

$$\mathcal{W}_{n,t} \equiv \left[\mathcal{Y}_{n,t}^*, m_{n,t}^{*\prime}, \mathcal{I}_{n,t}\right]' \text{ where } \mathcal{Y}_{n,t}^* = y_{n,t}^* - E\left[y_{n,t}^*\right], \mathcal{I}_{n,t} \equiv \left(\frac{n}{k_n^{(y)}}\right)^{1/2} \left(I\left(y_t \leq -l_n\right) - E\left[\left(y_t \leq -l_n\right)\right]\right),$$

and define long run variances and covariances:

$$\begin{aligned}\mathfrak{I}_n &\equiv \frac{1}{n}\sum_{s,t=1}^{n} E\left[\mathcal{I}_{n,s}\mathcal{I}_{n,t}\right] \quad \text{and} \quad \Psi_n \equiv \frac{1}{n}\sum_{s\geq t=1}^{n} E\left[\mathcal{I}_{n,s} m_{n,t}^*\right] \\ \Gamma_n &\equiv \frac{1}{n}\sum_{s\geq t=1}^{n} E\left[y_{n,s}^* m_{n,t}^{*\prime}\right] \quad \text{and} \quad \Phi_n \equiv \frac{1}{n}\sum_{s,t=1}^{n} E\left[\mathcal{Y}_{n,s}^* \mathcal{I}_{n,t}\right] \\ \mathfrak{W}_n &\equiv \frac{1}{n}\sum_{s,t=1}^{n} E\left[\mathcal{W}_{n,s}\mathcal{W}_{n,t}'\right] = \begin{bmatrix} \Upsilon_n & \Gamma_n & \Phi_n \\ \Gamma_n' & \Sigma_n & \Psi_n' \\ \Phi_n & \Psi_n & \mathfrak{I}_n \end{bmatrix}.\end{aligned}$$

Define a scale $\mathfrak{S}_n \equiv \mathcal{D}_n' \mathfrak{W}_n \mathcal{D}_n$ where $\mathcal{D}_n \equiv [1, -\Gamma_n \mathcal{P}_n, (\kappa_1 - 1)^{-1}(k_n^{(y)}/n)^{1/2} l_n^{(y)}]'$, and define a linear combination of scales:

$$\begin{aligned}\mathfrak{SV}_n \equiv\ & \mathfrak{S}_n I\left(\left|\widehat{ES}_{n,\alpha}^{(bc*)(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right| < \left|\widehat{ES}_{n,\alpha}^{*(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right|\right) \\ & + \mathfrak{V}_n I\left(\left|\widehat{ES}_{n,\alpha}^{(bc*)(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right| > \left|\widehat{ES}_{n,\alpha}^{*(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right|\right).\end{aligned}$$

**Theorem 7.1** *Let Assumption A hold, let* $P(y_t \leq -c) = d_1 c^{-\kappa_1}(1 + O(c^{-\xi_1}))$ *for some* $d_1, \xi_1 > 0$ *and* $\kappa_1 > 1$*, and let* $m_n \to \infty$*,* $m_n = O((\ln(n))^a)$ *for any chosen* $a > 0$*, and* $m_n/k_n^{(y)} \to \infty$*. Then (a).* $(n/\mathfrak{S}_n)^{1/2}(\widehat{ES}_{n,\alpha}^{(bc*)(\pi)} - ES_\alpha) \xrightarrow{d} N(0, \alpha^{-2})$*; (b).* $(n/\mathfrak{SV}_n)^{1/2}(\widehat{ES}_{n,\alpha}^{(obc)(\pi)} - ES_\alpha) \xrightarrow{d} N(0, \alpha^{-2})$*; and (c).* $\mathfrak{SV}_n = \mathfrak{S}_n + o_p(1)$ if and only if $\kappa_1 < 2$ *and* $\mathfrak{SV}_n = \mathfrak{V}_n + o_p(1)$ if and only if $\kappa_1 \geq 2$.

**Remark 21** Under second order power law tail decay $P(y_t \leq -c) = d_1 c^{-\kappa_1}(1 + O(c^{-\xi_1}))$ we need observations from sufficiently far out in the tails $m_n = O(n^{2\xi_1/(2\xi_1+\kappa_1)})$ to ensure $\hat{\kappa}_{1,m_n} = \kappa_1 + O_p(1/m_n^{1/2})$. See Haeusler and Teugels (1985). Since $\xi_1$ and $\kappa_1$ are unknown, we impose $m_n = O((\ln(n))^a)$ as a viable sufficient condition. The bound $k_n = o(m_n)$ ensures tail exponent estimators do no affect the limit distribution of $\widehat{ES}_{n,\alpha}^{(bc*)(\pi)}$ and $\widehat{ES}_{n,\alpha}^{(obc)(\pi)}$. However, $k_n = o((\ln(n))^a)$ also implies the untrimmed estimator $\widetilde{ES}_{n,\alpha}^{(\pi)}$ used to determine $\mathfrak{SV}_n$ does not affect asymptotics.

A flat-weighted estimator estimator $\widehat{ES}_{n,\alpha}^{(obc)}$ can similarly be defined. We also present the limit theory for $\widehat{ES}_{n,\alpha}^{(obc)}$ since this also contains a bias estimation improvement over Hill's (2015b) $\widehat{ES}_{n,\alpha}^{(bc*)}$. Define $\tilde{\mathfrak{S}}_n = \tilde{\mathcal{D}}_n' \mathfrak{W}_n \tilde{\mathcal{D}}_n$ where $\tilde{\mathcal{D}}_n = [1, 0, (\kappa_1 - 1)^{-1}(k_n^{(y)}/n)^{1/2} l_n^{(y)}]'$, and:

$$\tilde{\mathfrak{S}}\Upsilon_n = \tilde{\mathfrak{S}}_n I\left(\left|\widehat{ES}_{n,\alpha}^{(bc*)} - \widetilde{ES}_{n,\alpha}\right| < \left|\widehat{ES}_{n,\alpha}^{*} - \widetilde{ES}_{n,\alpha}\right|\right) + \Upsilon_n I\left(\left|\widehat{ES}_{n,\alpha}^{(bc*)} - \widetilde{ES}_{n,\alpha}\right| > \left|\widehat{ES}_{n,\alpha}^{*} - \widetilde{ES}_{n,\alpha}\right|\right),$$

with untrimmed $\widetilde{ES}_{n,\alpha} \equiv -\alpha^{-1} n^{-1} \sum_{t=1}^n y_t I(y_t \leq y_{[\alpha n]})$. We omit a proof of the following since it is similar to the proof of Theorem 7.1.

**Theorem 7.2** *Let Assumption A hold, let* $P(y_t \leq -c) = d_1 c^{-\kappa_1}(1 + O(c^{-\xi_1}))$ *for some* $d_1, \xi_1 > 0$ *and* $\kappa_1 > 1$*, and let* $m_n \to \infty$*,* $m_n = O((\ln(n))^a)$ *for any chosen* $a > 0$*, and* $m_n/k_n^{(y)} \to \infty$*. Then (a).* $(n/\tilde{\mathfrak{S}}_n)^{1/2}(\widehat{ES}_{n,\alpha}^{(bc*)} - ES_\alpha) \xrightarrow{d} N(0, \alpha^{-2})$*; (b).* $(n/\tilde{\mathfrak{S}}\Upsilon_n)^{1/2}(\widehat{ES}_{n,\alpha}^{(obc)} - ES_\alpha) \xrightarrow{d} N(0, \alpha^{-2})$*; and (c).* $\tilde{\mathfrak{S}}\Upsilon_n = \tilde{\mathfrak{S}}_n + o_p(1)$ if and only if $\kappa_1 < 2$*, and* $\tilde{\mathfrak{S}}\Upsilon_n = \Upsilon_n + o_p(1)$ if and only if $\kappa_1 \geq 2$.

The scales $\mathfrak{V}_n$, $\mathfrak{S}_n$ and $\mathfrak{SV}_n$ are easily estimated. Construct $\hat{\mathcal{P}}_n^{(\pi)}$ using $\hat{\Sigma}_n^{(\pi)}$ and $\widehat{\mathcal{J}}_n^{(\pi)}$. Let $\hat{\Upsilon}_n$, $\widehat{\mathfrak{I}}_n$, $\hat{\Gamma}_n$,, $\hat{\Psi}_n$ and $\hat{\Phi}_n$ be consistent estimators of the long-run variances $\Upsilon_n$ and $\mathfrak{I}_n$ and covariances $\Gamma_n$, $\Psi_n$ and $\Phi_n$, e.g. $\hat{\Gamma}_n = \sum_{s \geq t=1}^n \mathcal{K}_n((s-t)/\gamma_n) y_s I(y_{(k_n^{(y)})}^{(-)} \leq y_s \leq y_{[\alpha n]}) \hat{m}_{n,t}^*(\hat{\theta}_n)'$ where $\mathcal{K}_n(\cdot)$ is the kernel function with bandwidth $\gamma_n \to \infty$, $\gamma_n = o(n)$. Further, we require

$$\hat{\mathcal{D}}_n \equiv \left[1, -\hat{\Gamma}_n \hat{\mathcal{P}}_n^{(\pi)}, -\frac{1}{\hat{\kappa}_{1,m_n} - 1}\left(\frac{k_n^{(y)}}{n}\right)^{1/2} y_{(k_n^{(y)})}^{(-)}\right]' \text{ and } \widehat{\mathcal{I}}_{n,t} \equiv \left(\frac{n}{k_n^{(y)}}\right)^{1/2}\left\{I\left(y_t \leq -y_{(k_n^{(y)})}^{(-)}\right) - \frac{k_n^{(y)}}{n}\right\}.$$

Notice $-y^{(-)}_{(k_n^{(y)})}$ estimates $l_n^{(y)}$ in $\mathcal{D}_n$. Now compute $\widehat{\mathfrak{W}}_n$ from the above estimators, and:

$$\hat{\mathfrak{V}}_n \equiv \hat{\Upsilon}_n - \hat{\Gamma}_n \hat{\mathcal{P}}_n^{(\pi)} \hat{\Gamma}_n' \text{ and } \widehat{\mathfrak{S}}_n \equiv \hat{\mathcal{D}}_n' \widehat{\mathfrak{W}}_n \hat{\mathcal{D}}_n \tag{24}$$

$$\widehat{\mathfrak{SV}}_n \equiv \widehat{\mathfrak{S}}_n I\left(\left|\widehat{ES}_{n,\alpha}^{(bc*)(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right| < \left|\widehat{ES}_{n,\alpha}^{*} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right|\right) + \hat{\mathfrak{V}}_n I\left(\left|\widehat{ES}_{n,\alpha}^{(bc*)(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right| > \left|\widehat{ES}_{n,\alpha}^{*(\pi)} - \widetilde{ES}_{n,\alpha}^{(\pi)}\right|\right).$$

Similarly $\widehat{\tilde{\mathfrak{S}}\Upsilon}_n \equiv \hat{\tilde{\mathfrak{S}}}_n I(|\widehat{ES}_{n,\alpha}^{(bc*)} - \widetilde{ES}_{n,\alpha}| < |\widehat{ES}_{n,\alpha}^{*} - \widetilde{ES}_{n,\alpha}|) + \hat{\Upsilon}_n I(|\widehat{ES}_{n,\alpha}^{(bc*)} - \widetilde{ES}_{n,\alpha}| > |\widehat{ES}_{n,\alpha}^{*} - \widetilde{ES}_{n,\alpha}|)$, where $\hat{\tilde{\mathfrak{S}}}_n$ is constructed like $\widehat{\mathfrak{S}}_n$.

Consistency $\hat{\mathfrak{V}}_n/\mathfrak{V}_n \xrightarrow{p} 1$ and $\widehat{\mathfrak{S}}_n/\mathfrak{S}_n \xrightarrow{p} 1$ follow from Assumption A and limit theory arguments in the appendix. See Hill and Aguilar (2013) and Hill (2015b) for limit theory for kernel variance estimators under tail-trimming for a large class of kernels, and see Hill (2015b) for a similar scale estimator result under flat weighing. Last, $\widehat{\mathfrak{SV}}_n/\mathfrak{SV}_n \xrightarrow{p} 1$ follows from $\hat{\mathfrak{V}}_n/\mathfrak{V}_n \xrightarrow{p} 1$ and $\widehat{\mathfrak{S}}_n/\mathfrak{S}_n \xrightarrow{p} 1$, and $\widehat{\tilde{\mathfrak{S}}\Upsilon}_n/\tilde{\mathfrak{S}}\Upsilon_n \xrightarrow{p} 1$ can likewise be shown.

# 8 Simulation Study

In this section we study the small sample behavior of the GELITT estimators. We draw $10,000$ samples $\{y_t\}_{t=1}^n$ of size $n \in \{100, 250\}$ from a GARCH(1,1) process $y_t = \sigma_t \epsilon_t$ with $\sigma_t^2 = 1 + .3y_{t-1}^2 + .6\sigma_{t-1}^2$. The starting value is $\sigma_1^2 = 1$, and we simulate $2n$ observations and retain the last $n$ for estimation. The errors $\epsilon_t$ are iid with either a standard normal distribution, or a symmetric Pareto distribution $P(\epsilon_t > \epsilon) = P(\epsilon_t < -\epsilon) = (1/2)(1+\epsilon)^{-\kappa}$ with tail index $\kappa \in \{2.5, 4.5\}$. In the latter case we standardize $\epsilon_t$ to ensure $E[\epsilon_t^2] = 1$.

## 8.1 Base-Case

We estimate $\theta^0 = [1, .3, .6]'$ by GELITT and non-trimmed GEL using empirical likelihood, CUE and exponential tilting criteria $\rho(\cdot)$. The iterated volatility process used for estimation is $h_1(\theta) = \omega$ and $h_t(\theta) = \omega + \alpha y_{t-1}^2 + \beta h_{t-1}(\theta)$. In order to reduce notation, we simply write feasible variables as $\epsilon_t(\theta) \equiv y_t/h_t(\theta)$ and $s_t(\theta) \equiv (\partial/\partial\theta)\ln(h_t(\theta))$, etc. The estimating equations are $m_t(\theta) \equiv (\epsilon_t^2(\theta) - 1)x_t(\theta)$ with $x_t(\theta) = s_t(\theta)$ or $x_t(\theta) = [s_t'(\theta), s_{t-1}'(\theta)]'$ hence $q = 3$ or 6.

As discussed following Corollaries 2.3 and 4.5, and Lemma 4.4, the GELITT rate of convergence is optimized with $k_n^{(\epsilon)}$ close to $\zeta n$ for $\zeta \in (0,1)$, while higher order bias is reduced by using a small $\zeta$. Further, lightly trimming the score equations $s_t(\theta)$ improves finite sample performance,

although it is not needed in theory since $||E[s_t s_t']|| < \infty$. In the *base-case* we therefore trim $\epsilon_t(\theta)$ using a fractile $k_n^{(\epsilon)} = \max\{1, [.05n/\ln(n)]\}$, and we trim $s_t(\theta)$ based on extremes of $y_{t-1}$ generating the trimmed variable $\widehat{s}^*_{n,t}(\theta) = s_t(\theta) I(|y_t| \leq y^{(a)}_{(k_n^{(y)})})$ with $k_n^{(y)} = \max\{1, [.2\ln(n)]\}$. Since $n \in \{100, 250\}$ the fractiles are just $\{k_{100}^{(\epsilon)}, k_{100}^{(y)}\} = \{1, 1\}$ and $\{k_{250}^{(\epsilon)}, k_{250}^{(y)}\} = \{2, 1\}$. This combination promotes excellent over-all small sample results.

In Sections 8.2 and 8.3 we inspect how our estimator responds to variations from these specifications by studying parameter values for IGARCH and explosive GARCH models, and variations on the trimming fractiles.

Solving the GEL optimization problem posses well known problems due to the saddle point construction. We therefore roughly follow Guggenberger (2008) and search over a fine grid within $\Theta$. We uniformly randomly select 100,000 $\{\lambda, \theta\}$ from $[-.1, .1]^q \times [0, 1]^3$ and use only those points $\{\lambda, \theta\}$ that satisfy $\alpha + \beta \leq 1$ to ensure a stationary solution. This leads to roughly 3500 $\lambda' s$ and $\theta' s$, thus the typical grid has over 12,000,000 couplets $\{\lambda, \theta\}$. Except for CUE, for each $\theta$ we do a grid search for the "inner" optimization problem to find $\hat{\lambda}_n(\theta) = \arg\sup_{\lambda \in \hat{\Lambda}_n(\theta)} \{1/n \sum_{t=1}^n \rho(\lambda' \hat{m}^*_{n,t}(\theta))\}$ where only EL restricts $\hat{\Lambda}_n(\theta)$ above and beyond the grid $\Lambda$. Since CUE is quadratic, we use its analytic solution $\hat{\lambda}_n(\theta) = -(\sum_{t=1}^n \hat{m}^*_{n,t}(\theta)\hat{m}^*_{n,t}(\theta)')^{-1} \times \sum_{t=1}^n \hat{m}^*_{n,t}(\theta)$, cf. Bonnal and Renault (2004, eq. (3.3)). Then for the "outer" optimization problem we do a grid search to find $\hat{\theta}_n = \arg\min_{\theta \in \Theta} \{1/n \sum_{t=1}^n \rho(\hat{\Lambda}_n(\theta)' \hat{m}^*_{n,t}(\theta))\}$.[14]

We also compute $\theta^0$ by QML, and by Hill's (2015a) Quasi-Maximum Tail-Trimmed Likelihood [QMTTL], Peng and Yao's (2003) Log-LAD and Zhu and Ling's (2011) Weighted Laplace QML [WLQML]. The QMTTL criterion is $\sum_{t=2}^n \{\ln h_t(\theta) + \epsilon_t(\theta)\} \hat{I}_{n,t}(\theta)$ where $\hat{I}_{n,t}(\theta) \equiv I(|\epsilon_t(\theta)| \leq \epsilon^{(a)}_{(k_n^{(\epsilon)})}(\theta)) \times I(|y_{t-1}| \leq y^{(a)}_{(k_n^{(y)})})$ with $k_n^{(\epsilon)} = [.05n/\ln(n)]$ and $k_n^{(y)} = [.2\ln(n)]$. The Log-LAD criterion is $\sum_{t=2}^n |\ln \epsilon_t^2(\theta)|$. The WLQML criterion is $\sum_{t=2}^n \{\ln h_t^{1/2}(\theta) + |\epsilon_t(\theta)|\} w_t$ where the weights $w_t$ are computed as Zhu and Ling (2011): $w_t = (\max\{1, C^{-1} \sum_{i=1}^\infty i^{-9} |y_{t-i} I(|y_{t-i}| > C)|\})^{-4}$ where $C = y^{(a)}_{(.05n)}$ and $y_{t-i} = 0 \ \forall i \geq t$. In these cases we use a grid search over 10,000 uniformly randomly selected points $\theta \in [0, 1]^3$ subject to $\alpha + \beta \leq 1$.

We report the simulation bias, mean squared error and 95% confidence region for $\theta_3^0 = \beta^0 = .6$ across the 10,000 sample paths. The confidence region is computed by evaluating the profile empirical likelihood ratio function $2 \sum_{t=1}^n \rho(\hat{\lambda}_n' \hat{m}^*_{n,t}(\theta))$ evaluated at $\hat{\theta}_n$, with increments $\pm .005$ on $\hat{\theta}_{n,3}$, and choosing the endpoints based on when we reject the empirical likelihood ratio test

[14] Guggenberger (2008) focuses on a scalar iid regression model where the parameter is unrestricted in theory. He uses a gradient-Hessian method for the inner optimization problem to solve for $\hat{\lambda}_n(\theta)$ due to global concavity, and a grid search to find $\hat{\theta}_n$. We have a multivariate problem where $\theta^0$ is naturally bounded. Further, due to the iterative and therefore nonlinear nature of $h_t(\theta) = \omega + \alpha y_{t-1} + \beta h_{t-1}(\theta)$, we simply use a grid search for both inner and outer optimization problems by selecting entire vector points $\lambda$ and $\theta$. In view of computing $h_t(\theta)$ for each $\theta$, this is quite computationally intensive.

null hypothesis. The simulation results for $\theta_2^0 = \alpha = .3$ are qualitatively similar, and in general estimation results are similar for a range of values of $(\alpha^0, \beta^0)$. The intercept $\omega^0$, however, is more challenging and generally results in an estimator with greater dispersion. This becomes particularly accute when $\omega^0$ is close to zero as typically arises with financial data, and as is commonly encountered with QML and related estimators.

We also report the Kolmogorov-Smirnov statistic scaled by its 5% critical value. The statistic is computed from the standardized sequence $\{(\hat{\theta}_{n,3}^{(r)} - \theta_3^0)/s_R\}_{r=1}^R$ where $\{\hat{\theta}_{n,3}^{(r)}\}_{r=1}^R$ is the sequence of $R = 10,000$ independent estimates of $\theta_3^0$, and $s_R^2 \equiv 1/R \sum_{r=1}^R (\hat{\theta}_{n,3}^{(r)} - \theta_3^0)^2$ is a simulation estimator of $E[(\hat{\theta}_{n,3}^{(r)} - \theta_3^0)^2]$. Finally, we perform t-tests of the hypotheses that $\theta_3^0$ is $\theta_3 \in \{.6, .5, .35, 0\}$ and we report rejection frequencies at the 5% level. We reject the null hypothesis when $|(\hat{\theta}_{n,3}^{(r)} - \theta_3)/s_R| > 1.96$, hence the test is performed under the assumption the estimator is asymptotically normal. This fails to be true for GEL and QML when $E[\epsilon_t^4] = \infty$ hence size distortions are expected.

Simulation results for the *base-case* are reported in Tables 1 and 2. In the GEL and GELITT cases we only show results using over-identifying restrictions $q = 6$ since the exact identification results are similar. QML, WLQML, and Log-LAD exhibit comparatively large bias, where the small sample problems with QML are well known and lead to large t-test size distortions (see Section 1). Further, although Log-LAD and Weighted Laplace QML are robust in theory to heavy tails, since they are asymptotically normal when $E[\epsilon_t^2] < \infty$ and $E[\epsilon_t^4] = \infty$, they are not robust in small samples (see also Hill, 2015a). Indeed, each non-GEL estimator in this study, with the exception of QMTTL, deviates from normality and exhibits t-test size distortions. QMTTL compares well with the robust GEL counterparts, but relative to tail-trimmed CUE has a larger bias and mean squared error.

The GEL estimators by comparison are sharper than the non-GEL estimators, and trimming leads to estimators that are closer to normally distributed and have accurate t-test size. The most promising estimator is tail-trimmed CUE: in most cases it has the lowest bias and mse, and is closest to normally distributed. A plausible explanation is the quadratic criterion form: the estimator can be computed more easily which leads overall to small computation error, while trimming improves any estimator's approximate normality (cf. Hill, 2013, 2015a). It is also substantially faster to compute.

These findings are key since GELITT estimators have the same first order asymptotics, and GELITT and GEL are identical asymptotically when $E[\epsilon_t^4] < \infty$. Moreover, the EL criterion (with or without trimming) promotes smaller higher order bias. Thus, the simplicity of the CUE criterion form, and the sampling improvement associated with trimming a few sample extremes, leads to a dominant estimator.

## 8.2 IGARCH and Explosive GARCH

Our next experiment uses different GARCH parameter values such that $\alpha^0 + \beta^0 \geq 1$. We consider IGARCH $\{\alpha^0, \beta^0\} = \{.4, .6\}$ or $\{.3, .7\}$ and explosive GARCH $\{\alpha^0, \beta^0\} = \{.45, .6\}$ or $\{.35, .7\}$ and focus on the CUE criterion due to its dominant performance above. The explosive cases are easily verified to be stationary.[15] The search grid is now restricted to $\alpha + \beta \leq 1.1$.

We use the same trimming fractile for $\epsilon_t$ as above, $k_n^{(\epsilon)} = \max\{1, [.05n/\ln(n)]\}$. However, since $y_t$ now has heavier tails, the score weights $s_t(\theta) \equiv (\partial/\partial\theta)\ln(h_t(\theta))$ are more volatile in small samples, which leads to greater small sample bias then when $\{\alpha^0, \beta^0\} = \{.3, .6\}$.[16] We therefore increase the fractile $k_n^{(y)} = \max\{1, [.5\ln(n)]\}$ which implies $\{k_n^{(\epsilon)}, k_n^{(y)}\} = \{1, 2\}$ when $n = 100$ and $\{k_n^{(\epsilon)}, k_n^{(y)}\} = \{2, 3\}$ when $n = 250$. We show in Section 8.3 that related fractile values also lead to competitive GELITT results when base-case values $\{\alpha^0, \beta^0\} = \{.3, .6\}$ are used, hence the preceeding fractiles $\{k_n^{(\epsilon)}, k_n^{(y)}\}$ may be used in general. Tables 3 and 4 show the GELITT estimator works well, even when $y_t$ is very heavy tailed.

## 8.3 Trimming Variations

We now alter the trimming specifications for GELITT in order to see how various rules impact our estimator. We use the same base-case parameter values $\alpha^0 = .3$ and $\beta^0 = .6$. In view of the redundance of some results, and the relatively strong performance of CUE under tail-trimming as reported above, we only coincide the CUE criterion.

We do two experiments. In the first, we compute bias and Kolmogorov-Smirnov [KS] statistics over a grid of trimming fractiles $\{k_n^{(\epsilon)}, k_n^{(y)}\}$. In this case, we only use Paretian $\epsilon_t$ with index $\kappa = 2.5$ and sample size $n = 100$. In the second we fix either $k_n^{(\epsilon)}$ *or* $k_n^{(y)}$ and inspect bias, mse, the KS test and t-tests for each $\epsilon_t$ distribution and sample size $n$. Since the former reveals the essential details that we desire, we present the latter in the supplemental material Hill and Prokhorov (2014).

See Figures 1 and 2 for a plot of simulation bias and the KS statistic scaled by its 5% critical value. The plots are over a grid $\{k_n^{(\epsilon)}, k_n^{(y)}\}$ ranging from $\{1, 1\}$ to $\{12, 23\}$. Smaller $k_n^{(\epsilon)}$ aligns with lower bias and KS values for evidently any $k_n^{(y)}$. Furthermore, for fixed $k_n^{(\epsilon)}$, bias and the KS statistic increase noticeably only when $k_n^{(y)}$ is fairly large.

[15] We drew $R = 1,000,000$ observations of iid $\epsilon_t$ from normal and Paretian distributions, and computed $1/R\sum_{t=1}^{R}\ln(\alpha^0 + \beta^0\epsilon_t^2)$. The 99.99% asymptotic confidence bands are below zero, providing evidence of stationarity (cf. Nelson, 1990).

[16] A possible reason is the iterated volatility process $h_1^2(\theta) = \omega$ and $h_t^2(\theta) = \omega + \alpha y_{t-1}^2 + \beta h_{t-1}^2(\theta)$ tends to under-approximate $\sigma_t^2(\theta)$ in small samples, hence standardized GARCH processes in small samples tend to be heavier tailed than the true process. See Hill (2015c) for evidence.

# 9 Empirical Application - Expected Shortfall

We estimate the parameters of a GARCH model and the expected shortfall for financial returns series. We use the same data studied in Hill (2015b) in order to compare results. The data are the Hang Seng Index [HSI] for June 3, 1996 - May 31, 1998, and the Russian Ruble - U.S. Dollar exchange rate for Jan. 1, 1999 - Oct. 29, 2008. The Ruble period lies between major financial crises in Russia, and globally. See Hill (2013, 2015b) and Ibragimov (2009) for evidence that these series are heavy tailed, and likely have an infinite variance over the chosen sample period. We take each series $\{x_t\}$ and compute the daily log returns $y_t = \ln(x_t) - \ln(x_{t-1})$, resulting in 489 and 2449 returns for the HSI and Ruble, respectively. See Figure 1 in Hill (2015b) for plots of returns, and tail index confidence bands.

We pass each series through a GARCH(1,1) filter using tail-trimmed CUE with $k_n^{(\epsilon)} = \max\{1, [.05n/\ln(n)]\}$ and $k_n^{(y)} = \max\{1, [.2\ln(n)]\}$, as in the base-line simulation experiment. We compute the optimal bias-corrected profile weighted expected shortfall $\widehat{ES}_{n,\alpha}^{(obc)(\pi)}$ and flat weighted $\widehat{ES}_{n,\alpha}^{(obc)}$ at risk levels $\alpha = .05$. The estimates are computed over rolling sub-samples of size 250 days, hence there are 2,200 and 240 windows for the Ruble and HSI, respectively. We use the same fractiles as in Hill (2015b, Section 3): tail trimming with $k_n^{(y)} = \min\{1, [.25n^{2/3}/(\ln(n))^{2\iota}]\}$ and tail index estimation with $m_n(\phi) = \min\{1, \phi[k_n^{(y)}(\ln(n))^{\iota}]\}$ and $\phi \in [.05, \mathcal{M}]$ where $\mathcal{M} = 20$ for the HSI and $\mathcal{M} = 7$ for the Ruble. Hill (2015b) uses $\mathcal{M} = 7$ in both cases, but we find using a much larger upper bound improves our bias corrected estimator during the most volatility periods.

We compute 95% asymptotic confidence bands $\widehat{ES}_{n,\alpha}^{(obc)} \pm 1.96 \times (\widehat{\tilde{\mathfrak{S}}\mathfrak{V}}_n/n)^{1/2}$ and $\widehat{ES}_{n,\alpha}^{(obc)(\pi)} \pm 1.96 \times (\widehat{\mathfrak{S}\mathfrak{V}}_n/n)^{1/2}$ using estimators $\widehat{\tilde{\mathfrak{S}}\mathfrak{V}}_n$ and $\widehat{\mathfrak{S}\mathfrak{V}}_n$ detailed in, and following, (24). As in Hill (2015b, Section 4), where appropriate for variance and covariance estimators, we use a Barlett kernel and bandwidth $\gamma_n = n^{.25}$. We also compute the non-trimmed expected shortfall estimator with flat and profile weighting, where the latter is based on tail-trimmed CUE. We use the same kernel method for computing the asymptotic scale, and compute asymptotic confidences bands under the assumption a second moment exists.

Figures 3-5 contain the rolling window results. We focus the discussion on the HSI in Figures 3 and 4 since the results for the Ruble are similar. There are four noticeable outcomes. First, Figure 3 shows the flat *or* profile weighted convex combinations $\{\widehat{ES}_{n,\alpha}^{(obc)}, \widehat{ES}_{n,\alpha}^{(obc)(\pi)}\}$ are nearly equivalent to the untrimmed ES estimator with flat or profile weighting. Although our plots do not show this, we find this occurs primarily from the expanded range $m_n(\phi)$ on $\phi \in [.05, 20]$ relative to Hill's (2015b) $[.05, 7]$. The estimator $\widehat{ES}_{n,\alpha}^{(bc*)}$ used in Hill (2015b), and the profile weighted version $\widehat{ES}_{n,\alpha}^{(bc*)(\pi)}$, both with $\phi \in [.05, 7]$, deviate from the untrimmed estima-

tors during later windows, windows that contain the most volatile periods. When these extremes are trimmed, they can render a trimmed estimator comparatively more biased. Further, it is harder to estimate large bias well since large bias by construction implies a greater trimmed mean distance from the tails, while the bias estimator is based on a tail moment approximation that is sharper in the extreme tails by construction (i.e. it is sharper when the trimmed mean portion is comparatively small). The difficulty in estimating bias during volatile periods is ameliorated primarily by optimizing bias over a greater range of tail fractiles, but also by using $\{\widehat{ES}_{n,\alpha}^{(obc)}, \widehat{ES}_{n,\alpha}^{(obc)(\pi)}\}$ since they cannot be farther from the untrimmed mean.

The same outcome occurs with the Ruble during the crisis year 1999, in which volatility was at its highest during the sample period. The estimates in Hill (2015b) deviate from the untrimmed estimator more than the estimates computed here, but all roughly converge during the low volatility period starting roughly in 2000.

Second, Figure 3 shows $\{\widehat{ES}_{n,\alpha}^{(obc)}, \widehat{ES}_{n,\alpha}^{(obc)(\pi)}\}$ are slightly more volatile than the untrimmed estimator, precisely due to the bias estimator.

Third, from Figure 4 we see that the confidence bands for the untrimmed estimator are very large, indicating greater dispersion in the non-tail trimmed data. The estimated variance for the untrimmed estimator, with or without profile weighting, is roughly 100 to 1000 times larger due to the exceptionally large values that remain when tail-trimming is not used, and the greatest discrepancy occurs during the later windows since these have the largest sample values.

Fourth , the use of profile weights leads to slightly tighter confidence bands, as theory predicts. Although it is difficult to see, the variance estimates are roughly $1\% - 5\%$ smaller with profile weights in the case of tail-trimming, but only roughly $.5\%. - 1\%$ smaller when tail-trimming is not used due to the large dispersion of this estimator.

# 10 Concluding Remarks

We develop heavy tail robust Generalized Empirical Likelihood estimators for GARCH models by tail-trimming the errors in QML-type estimating equations. Feedback erodes the rate of convergence below $n^{1/2}$ when the errors have an infinite fourth moment, but tail-trimming permits asymptotically standard inference. In heavy tailed cases, the rate can always be pushed as close to $n^{1/2}$ as we choose by using a simple rule of thumb for trimming. Tail-trimming in a GEL framework offers both heavy tail robustness and implied probabilities for efficient and robust moment estimation and inference, and we show how the profile weights in the CUE case augment weight on observations based on whether the error is very large or not. A higher order bias

characterization coupled with first order asymptotics gives new details about what a reasonable trimming strategy should be. We use the profiles for efficient and heavy tail robust expected shortfall estimation, and propose an improved bias correction, with new limit theory and scale estimation. The GEL estimator works well in controlled experiments, where tail-trimmed CUE is especially promising. Finally, improvements to the bias-corrected tail-trimmed expected shortfall estimator lead to a superb approximation to the sample mean with low dispersion, made evident by an application to financial returns. Future work should focus on the bootstrap or related sub-sampling techniques for tail-trimmed heavy tailed data, in order to ease anticipated size distortions from GEL-related test statistics. Further, although we present a (higher order) bias corrected tail-trimmed GEL estimator, we leave for future research a study of its finite sample performance.

# 11 Acknowledgements

We especially thank two referees and editor Yacine Aït Sahalia for helpful comments. We also gratefully acknowledge helpful comments from participants of the 2nd Humboldt-Copenhagen Conference in Financial Econometrics, the 7th International Symposium on Econometric Theory and Applications and the 2011 NBER-NSF Time-Series Conference, as well as seminar participants at University of New South Wales, New Economic School, University of Auckland and Kyoto University. This research has been supported by the Social Sciences and Humanities Research Council of Canada.

# A Appendix: Proofs of Main Results

We first introduce notation used in the proofs. We then present supporting lemmas used to prove the main results. Finally, we prove the main theorems.

## A.1 Notation

Throughout $o_p(1)$ does not depend on $\theta$ and $\lambda$, unless otherwise specified. "*w.p.a.1*" means "*with probability approaching one*".

In order to reduce the number of cases and to keep notation simple, we assume $x_t$ is square integrable and not trimmed, and whenever useful we assume exact identification $x_t = s_t$. Hence

$$\hat{\epsilon}^*_{n,t}(\theta) = \epsilon_t(\theta)\hat{I}^{(\epsilon)}_{n,t}(\theta) \text{ and } \hat{m}^*_{n,t}(\theta) = \left(\hat{\epsilon}^{*2}_{n,t}(\theta) - \frac{1}{n}\sum_{t=1}^{n}\hat{\epsilon}^{*2}_{n,t}(\theta)\right) \times x_t(\theta).$$

The proofs below extend to the over-identification case where $w_t$ contains lags of $s_t$, and can be easily generalized to allow for other $\Im_{t-1}$-measurable $w_t$ that require trimming. Similarly, we augment Assumption A.2 and impose power law tails on $\epsilon_t$ in general:

$$P\left(|\epsilon_t| > a\right) = d_\epsilon a^{-\kappa_\epsilon}\left(1 + o\left(1\right)\right) \quad \text{where} \quad d_\epsilon \in (0, \infty) \text{ and } \kappa_\epsilon \in (2, \infty). \tag{A.1}$$

We compactly write throughout:

$$d = d_\epsilon, \;\; \kappa = \kappa_\epsilon, \;\; k_n = k_n^{(\epsilon)} \;\text{ and }\; c_n = c_n^{(\epsilon)}.$$

Recall

$$\hat{\Lambda}_n(\theta) = \left\{\lambda : \lambda' \hat{m}_{n,t}^*(\theta) \in \mathcal{D}, \; t = 1, 2, ..., n\right\} \;\text{ and }\; \Lambda_n = \left\{\lambda : \sup_{\theta \in \Theta} \left\|\lambda' \Sigma_n^{1/2}(\theta)\right\| \leq K n^{-1/2}\right\}.$$

We require a criterion and moments based on the trimmed equations $m_{n,t}^*(\theta)$ that use non-stochastic thresholds:

$$\begin{aligned}
&\hat{Q}_n(\theta, \lambda) \equiv \frac{1}{n}\sum_{t=1}^n \rho\left(\lambda' \hat{m}_{n,t}^*(\theta)\right) \;\text{ and }\; \tilde{Q}_n(\theta, \lambda) \equiv \frac{1}{n}\sum_{t=1}^n \rho\left(\lambda' m_{n,t}^*(\theta)\right) \\
&\Lambda_n \equiv \left\{\lambda : \sup_{\theta \in \Theta} \left\|\lambda' \Sigma_n^{1/2}(\theta)\right\| \leq K n^{-1/2}\right\} \\
&m_n^*(\theta) \equiv \frac{1}{n}\sum_{t=1}^n m_{n,t}^*(\theta) \;\text{ and }\; \hat{m}_n^*(\theta) \equiv \frac{1}{n}\sum_{t=1}^n \hat{m}_{n,t}^*(\theta) \;\text{ and }\; \mathfrak{m}_n \equiv \sup_{\theta \in \Theta} \left\|E\left[m_{n,t}^*(\theta)\right]\right\|.
\end{aligned}$$

Asymptotic arguments require covariance and Jacobian components for tail-trimmed equations:

$$\begin{aligned}
&\hat{\Sigma}_n(\theta) \equiv \frac{1}{n}\sum_{t=1}^n \hat{m}_{n,t}^*(\theta)\hat{m}_{n,t}^*(\theta)' \;\text{ and }\; \tilde{\Sigma}_n(\theta) \equiv \frac{1}{n}\sum_{t=1}^n m_{n,t}^*(\theta) m_{n,t}^*(\theta)' \qquad (A.2) \\
&\widehat{\mathcal{J}}_{n,t}(\theta) \equiv \left(\frac{\partial}{\partial\theta}\epsilon_t^2(\theta) \times \hat{I}_{n,t}^{(\epsilon)}(\theta) - \frac{1}{n}\sum_{t=1}^n \frac{\partial}{\partial\theta}\epsilon_t^2(\theta) \times \hat{I}_{n,t}^{(\epsilon)}(\theta)\right) x_t(\theta) \\
&\qquad\qquad + \left(\epsilon_t^2(\theta)\hat{I}_{n,t}^{(\epsilon)}(\theta) - \frac{1}{n}\sum_{t=1}^n \epsilon_t^2(\theta)\hat{I}_{n,t}^{(\epsilon)}(\theta)\right)\frac{\partial}{\partial\theta}x_t(\theta)
\end{aligned}$$

Non-negligible trimming, and distribution continuity and non-degeneracy, ensure

$$\liminf_{n\to\infty} \|\mathfrak{m}_n\| > 0 \text{ and } \liminf_{n\to\infty} \|\Sigma_n\| > 0, \text{ and } \Sigma_n^{-1} \text{ exists as } n \to \infty.$$

Assumption A holds throughout. Then $\{y_t, \sigma_t^2(\theta)\}$ on $\Theta$ are stationary, ergodic, and geometrically $\beta$-mixing on $\Theta$ by (2), cf. Nelson (1990) and Carrasco and Chen (2002). Therefore, $w_t(\theta)$ is geometrically $\beta$-mixing since it is $\Im_{t-1}$-measurable, and $\epsilon_t(\theta) = \epsilon_t\sigma_t/\sigma_t(\theta)$ is stationary and ergodic.

Since $E(\sup_{\theta\in\Theta}|\sigma_t^2/\sigma_t^2(\theta)|)^p < \infty$ for any $p > 0$, cf. Francq and Zakoïan (2004, eq. (4.25)), it follows the product convolution $\epsilon_t(\theta) = \epsilon_t\sigma_t/\sigma_t(\theta)$ has a power law tail with the same index $\kappa > 2$ (Breiman, 1965):

$$P\left(|\epsilon_t(\theta)| > a\right) = d(\theta)a^{-\kappa}\left(1 + o\left(1\right)\right), \ \inf_{\theta\in\Theta} d(\theta) \in (0,\infty), \text{ and } o\left(1\right) \text{ does not depend on } \theta. \tag{A.3}$$

By construction of $c_n(\theta)$ in (6), therefore,

$$c_n(\theta) = d(\theta)^{1/\kappa}\left(n/k_n\right)^{1/\kappa}. \tag{A.4}$$

Similarly $\sup_{\theta\in\mathcal{N}_0}|s_{i,t}(\theta)|$ is $L_p$-bounded for any $p > 2$ and some compact subset $\mathcal{N}_0 \subseteq \Theta$ containing $\theta^0$. This follows by a trivial generalization of arguments in Francq and Zakoïan (2004, Section 4.2). Therefore, in the exact identification case by independence $m_{i,t}(\theta) = (\epsilon_t^2(\theta) - 1)s_{i,t}(\theta) = (\epsilon_t^2\sigma_t^2/\sigma_t^2(\theta) - 1)s_{i,t}(\theta)$ has a power-law tail with index $\kappa/2$ (see, e.g., Breiman, 1965):

$$P\left(|m_{i,t}(\theta)| > a\right) = d_i(\theta)a^{-\kappa/2}\left(1 + o\left(1\right)\right), \ \inf_{\theta\in\Theta} d_i(\theta) \in (0,\infty), \text{ and } o\left(1\right) \text{ does not depend on } \theta. \tag{A.5}$$

The trimmed moment $\mathfrak{E}_n(\theta) \equiv E[\epsilon_t^4(\theta)I(|\epsilon_t(\theta)| \leq c_n(\theta))]$ can be characterized by case by invoking (A.3), (A.4) and Karamata's Theorem (cf. Theorem 0.6 in Resnick, 1987):

$$\begin{aligned}
&\text{if } \ \kappa = 4: \ \ (0,\infty) \leftarrow \inf_{\theta\in\Theta}\left\{\frac{\mathfrak{E}_n(\theta)}{\ln(n)}\right\} \leq \sup_{\theta\in\Theta}\left\{\frac{\mathfrak{E}_n(\theta)}{\ln(n)}\right\} \to (0,\infty) \\
&\text{if } \ \kappa < 4: \ \ (0,\infty) \leftarrow \inf_{\theta\in\Theta}\left\{\frac{\mathfrak{E}_n(\theta)}{c_n^4(\theta)(k_n/n)}\right\} \leq \sup_{\theta\in\Theta}\left\{\frac{\mathfrak{E}_n(\theta)}{c_n^4(\theta)(k_n/n)}\right\} \to (0,\infty).
\end{aligned} \tag{A.6}$$

Similarly, by (A.5) and Karamata's Theorem, $\mathfrak{M}_{i,j,n}(\theta) \equiv E[m_{i,n,t}^*(\theta)m_{j,n,t}^*(\theta)]$ satisfies

$$\begin{aligned}
&\text{if } \ \kappa = 4: \ \ (0,\infty) \leftarrow \inf_{\theta\in\Theta}\left\{\frac{\mathfrak{M}_{i,j,n}(\theta)}{\ln(n)}\right\} \leq \sup_{\theta\in\Theta}\left\{\frac{\mathfrak{M}_{i,j,n}(\theta)}{\ln(n)}\right\} \to (0,\infty) \\
&\text{if } \ \kappa < 4: \ \ (0,\infty) \leftarrow \inf_{\theta\in\Theta}\left\{\frac{\mathfrak{M}_{i,j,n}(\theta)}{c_n^4(\theta)(k_n/n)}\right\} \leq \sup_{\theta\in\Theta}\left\{\frac{\mathfrak{M}_{i,j,n}\theta)}{c_n^4(\theta)(k_n/n)}\right\} \to (0,\infty).
\end{aligned} \tag{A.7}$$

## A.2 Preliminary Results

We require several supporting lemmata in order to prove the main theorems. Proofs are presented in the supplemental material Hill and Prokhorov (2014). First, we repeatedly exploit uniform bounds on the thresholds $c_n(\theta)$ and covariance $\Sigma_n(\theta)$, and a uniform law for the intermediate order sequence $\{\epsilon_{(k_n)}^{(a)}(\theta)\}$.

**Lemma A.1 (threshold bound)** $\sup_{\theta\in\Theta}\{c_n^4(\theta)/||\Sigma_n(\theta)||\} = o(n)$.

**Lemma A.2 (covariance bound)** $\sup_{\theta\in\Theta}||\Sigma_n(\theta)|| = o(n)$.

**Lemma A.3 (uniform threshold law)** $\sup_{\theta\in\Theta}|\epsilon_{(k_n)}^{(a)}(\theta)/c_n(\theta) - 1| = O_p(1/k_n^{1/2})$.

Next, we require a variety of laws of large numbers for possibly very heavy tailed random variables. We therefore present a basic result here for general use.

**Lemma A.4 (generic ULLN)** *Let $\{z_t(\theta)\}$ be a strictly stationary geometrically $\beta$-mixing process, with Paretian tail $P(|z_t(\theta)| > z) = d(\theta)z^{-\kappa(\theta)}(1 + o(1))$, $d(\theta), \kappa(\theta) \in (0, \infty)$. Define the tail trimmed version $z^*_{n,t}(\theta) \equiv z_t(\theta)I(|z_t(\theta)| \leq c_n(\theta))$, where $P(|z_t(\theta)| > c_n(\theta)) = k_n/n = o(1)$, and $k_n \to \infty$. Let $k_n/n^\iota \to \infty$ for some tiny $\iota > 0$. Then $\sup_{\theta\in\Theta} |1/n \sum_{t=1}^n \{z^*_{n,t}(\theta) - E[z^*_{n,t}(\theta)]\} \times (1 + o_p(1))| \xrightarrow{p} 0$ where $o_p(1)$ that may be a functions of $\theta$.*

We must show asymptotics are grounded on $m^*_{n,t}(\theta)$, we require consistent covariance and Jacobian estimators, and a central limit theorem for tail-trimmed equations.

**Lemma A.5 (approximation)** $\sup_{\theta\in\Theta} ||n^{-1/2}\Sigma_n^{-1/2}(\theta) \sum_{t=1}^n \{\hat{m}^*_{n,t}(\theta) - m^*_{n,t}(\theta)\}|| = o_p(1)$.

**Lemma A.6 (covariance consistency)** *Recall $\tilde{\Sigma}_n$ and $\hat{\Sigma}_n$ in (A.2), and assume $\tilde{\theta}_n \xrightarrow{p} \theta^0$. a. $\tilde{\Sigma}_n(\tilde{\theta}_n) = \Sigma_n(1 + o_p(1))$; and b. $\hat{\Sigma}_n(\tilde{\theta}_n) = \Sigma_n(1 + o_p(1))$.*

**Lemma A.7 (Jacobian consistency)** $1/n \sum_{t=1}^n \widehat{\mathcal{J}}_{n,t}(\tilde{\theta}_n) = \mathcal{J}_n \times (1 + o_p(1))$ *for any* $\tilde{\theta}_n \xrightarrow{p} \theta^0$.

**Lemma A.8 (CLT)** $n^{-1/2}\Sigma_n^{-1/2} \sum_{t=1}^n m^*_{n,t} \xrightarrow{d} N(0, I_q)$.

The next set of results are classic supporting arguments for GEL asymptotics, cf. Newey and Smith (2004), augmented to account for tail-trimming and heavy tails.

**Lemma A.9 (uniform GEL argument)** $\sup_{\theta\in\Theta,\lambda\in\Lambda_n}\{\max_{1\leq t\leq n} |\lambda' m^*_{n,t}(\theta)|\} \xrightarrow{p} 0$, $\sup_{\theta\in\Theta,\lambda\in\Lambda_n} \{\max_{1\leq t\leq n} |\lambda'\hat{m}^*_{n,t}(\theta)|\} \xrightarrow{p} 0$ *and* $\Lambda_n \subseteq \hat{\Lambda}_n(\theta)$ *w.p.a.1.* $\forall\theta \in \Theta$. *In particular* $\sup_{\theta\in\Theta,\lambda\in\Lambda_n} \{\max_{1\leq t\leq n} |\lambda'\{\hat{m}^*_{n,t}(\theta) - m^*_{n,t}(\theta)\}|\} \xrightarrow{p} 0$.

**Lemma A.10 (constrained GEL)** *Consider any sequence $\{\tilde{\theta}_n\}$, $\tilde{\theta}_n \in \Theta$, $\tilde{\theta}_n \xrightarrow{p} \theta^0$, such that $||m^*_n(\tilde{\theta}_n)|| = O_p(||\Sigma_n||^{1/2}/n^{1/2})$. Then $\bar{\lambda}_n \equiv \arg\max_{\lambda\in\hat{\Lambda}_n(\tilde{\theta}_n)}\{\hat{Q}_n(\tilde{\theta}_n, \lambda)\}$ exists w.p.a.1, $\bar{\lambda}_n = O_p(||\tilde{\Sigma}_n(\tilde{\theta}_n)||^{-1/2}n^{-1/2}) = o_p(1)$, and $\sup_{\lambda\in\hat{\Lambda}_n(\tilde{\theta}_n)}\{\hat{Q}_n(\tilde{\theta}_n, \lambda)\} \leq \rho^{(0)} + O_p(||\tilde{\Sigma}_n(\tilde{\theta}_n)||^{-1}n^{-1})$.*

**Lemma A.11 (equation limit)** $m^*_n(\hat{\theta}_n) = O_p(||\Sigma_n||^{1/2}/n^{1/2}) = o_p(1)$.

**Lemma A.12 (profile weight)** *Let $\tilde{\pi}^*_{n,t}(\theta) \equiv \rho^{(1)}(\tilde{\lambda}_n'\hat{m}^*_{n,t}(\theta))/\sum_{t=1}^n \rho^{(1)}(\tilde{\lambda}_n'\hat{m}^*_{n,t}(\theta))$. If $\tilde{\lambda}_n = O_p(||\Sigma_n||^{-1/2}n^{-1/2})$ where $O_p(\cdot)$ is not a function of $\theta$, then $\sup_{\theta\in\Theta} \max_{1\leq t\leq n} |\hat{\pi}^*_{n,t}(\theta) - 1/n| = O_p(||\Sigma_n||^{-1/2}/n^{3/2})$.*

**Remark 22** $\tilde{\lambda}_n = O_p(||\Sigma_n||^{-1/2}n^{-1/2})$ holds for the GELITT multipliers $\hat{\lambda}_n$ by Theorem 2.1.

## A.3 Proofs of Theorems 2.1 and 2.2

**Proof of Theorem 2.1.**

**Step 1.** Consider $\hat{\theta}_n$. By ULLN Lemma A.4 $|m_n^*(\hat{\theta}_n) - E[m_{n,t}^*(\hat{\theta}_n)] \times (1 + o_p(1))|| \xrightarrow{p} 0$ and by Lemma A.11 $m_n^*(\hat{\theta}_n) \xrightarrow{p} 0$. Hence, by the triangle inequality $||E[m_{n,t}^*(\hat{\theta}_n)] \times (1 + o_p(1))|| \xrightarrow{p} 0$, therefore $E[m_{n,t}^*(\hat{\theta}_n)] \to 0$.

Now, observe that $E[m_{n,t}^*(\theta)]$ is continuous, by dominated convergence $E[m_{n,t}^*(\theta)] \to 0$ *if and only if* $\theta = \theta^9$, and by construction $E[m_{n,t}^*(\theta^0)] = 0$ for $1 \leq t \leq n$ and $n \geq 1$. At any other $\tilde{\theta} \neq \theta^0$ it follows by the definition of a limit that $||E[m_{n,t}^*(\tilde{\theta})]|| > 0$ for all $n \geq N$ and some $N \geq 1$. Therefore $\theta^0$ is the unique point that satisfies $E[m_{n,t}^*(\theta^0)] = 0$ for all $n \geq N$. Combine $E[m_{n,t}^*(\hat{\theta}_n)] \to 0$ and $E[m_{n,t}^*(\theta)] = 0$ *if and only if* $\theta = \theta^0$ for all $n \geq N$ to deduce by continuity that $||\hat{\theta}_n - \theta^0|| \leq \delta$ for any $\delta > 0$ with probability approaching one. Hence $\hat{\theta}_n \xrightarrow{p} \theta^0$.

**Step 2.** Now consider $\hat{\lambda}_n$. In view of $\hat{\theta}_n \xrightarrow{p} \theta^0$ by Step 1, and $m_n^*(\hat{\theta}_n) = O_p(||\Sigma_n||/n^{1/2})$ by Lemma A.11, the conditions of Lemma A.10 are satisfied for $\hat{\theta}_n$. Therefore $\hat{\lambda}_n$ exists and $\hat{\lambda}_n = O_p(||\tilde{\Sigma}_n(\hat{\theta}_n)||^{-1/2} n^{-1/2})$ which is $O_p(||\Sigma_n||^{-1/2} n^{-1/2})$ by covariance consistency Lemma A.6.a. ■

**Proof of Theorem 2.2.** The proof is similar to arguments in Newey and Smith (2004, p. 240-24). Write $\hat{\rho}_{n,t}^{(i)} \equiv \rho^{(i)}(\hat{\lambda}_n' \hat{m}_{n,t}^*(\hat{\theta}_n))$ and $\mathring{\rho}_{n,t}^{(i)} \equiv \rho^{(i)}(\lambda_{n,*}' \hat{m}_{n,t}^*(\hat{\theta}_n))$ for some $0 \leq ||\lambda_{n,*}|| \leq ||\hat{\lambda}_n||$ that may be different in different places. Let $\theta_{n,*}$ satisfy $||\theta_{n,*} - \theta^0|| \leq ||\hat{\theta}_n - \theta^0||$ which may be different in different places. Define

$$\hat{\mathcal{M}}_n(\theta_{n,*}, \lambda_{n,*}) \equiv \begin{bmatrix} 0 & \frac{1}{n}\sum_{t=1}^n \hat{\rho}_{n,t}^{(1)} \widehat{\mathcal{J}}_{n,t}(\hat{\theta}_n)' \\ \frac{1}{n}\sum_{t=1}^n \mathring{\rho}_{n,t}^{(1)} \widehat{\mathcal{J}}_{n,t}(\theta_{n,*}) & \frac{1}{n}\sum_{t=1}^n \mathring{\rho}_{n,t}^{(2)} \hat{m}_{n,t}^*(\theta_{n,*}) \hat{m}_{n,t}^*(\hat{\theta}_n)' \end{bmatrix}.$$

and

$$\mathcal{A}_n \equiv \begin{bmatrix} n(\mathcal{J}_n' \Sigma_n^{-1} \mathcal{J}_n) & 0 \\ 0 & n\mathcal{P}_n^{-1} \end{bmatrix}, \ \mathcal{M}_n \equiv -\begin{bmatrix} 0 & \mathcal{J}_n' \\ \mathcal{J}_n & \Sigma_n \end{bmatrix}, \ \mathcal{M}_n^{-1} = -\begin{bmatrix} -(\mathcal{J}_n' \Sigma_n^{-1} \mathcal{J}_n)^{-1} & \mathcal{H}_n \\ \mathcal{H}_n' & \mathcal{P}_n \end{bmatrix}$$
$$\mathcal{H}_n \equiv (\mathcal{J}_n' \Sigma_n^{-1} \mathcal{J}_n)^{-1} \mathcal{J}_n' \Sigma_n^{-1} \text{ and } \mathcal{P}_n \equiv \Sigma_n^{-1} - \Sigma_n^{-1} \mathcal{J}_n (\mathcal{J}_n' \Sigma_n^{-1} \mathcal{J}_n)^{-1} \mathcal{J}_n' \Sigma_n^{-1} \text{ and } \mathcal{V}_n \equiv n(\mathcal{J}_n' \Sigma_n^{-1} \mathcal{J}_n).$$

Notice $\max_{1 \leq t \leq n} |\mathring{\rho}_{n,t}^{(i)} + 1| \xrightarrow{p} 0$ follows directly from Lemmas A.10 and A.9 since $||\lambda_{n,*}|| \leq ||\hat{\lambda}_n|| = O_p(||\Sigma_n||^{-1/2} n^{-1/2})$ by Theorem 2.1.

**Step 1.** The first-order-condition is

$$\sum_{t=1}^n \rho^{(1)}\left(\hat{\lambda}_n' \hat{m}_{n,t}^*(\hat{\theta}_n)\right) \times \begin{bmatrix} \widehat{\mathcal{J}}_{n,t}(\hat{\theta}_n)' \hat{\lambda}_n \\ \hat{m}_{n,t}^*(\hat{\theta}_n) \end{bmatrix} = 0 \quad a.s. \tag{A.8}$$

This follows by combining classic GEL optimization theory with optimization theory when estimating equations are trimmed. The former is grounded on seminal arguments due to Newey and Smith (2004, p. 240-241) based on the saddle point optimization problem (5). The latter

involves *almost sure* differentiability of trimmed equations that are continuous functions of continuously distributed random variables, developed in Cizek (2008, Appendices). See also Parente and Smith (2011).

Further, for some $||\theta_{n,*} - \theta^0|| \leq ||\hat{\theta}_n - \theta^0||$ and $||\lambda_{n,*}|| \leq ||\hat{\lambda}_n||$ that may be different in different places:

$$0 = \left[ \begin{array}{cc} 0', & \frac{1}{n}\sum_{t=1}^{n} \hat{m}_{n,t}^{*\prime} \end{array} \right]' + \hat{\mathcal{M}}_n \left(\theta_{n,*}, \lambda_{n,*}\right) \times \left(\hat{\vartheta}_n - \vartheta^0\right). \tag{A.9}$$

This can be verified by using theory developed in Hill (2013, Appendix B) and Hill (2015a, Appendix B) for similar first order expansions under tail-trimming, in order to expand the first order equations (A.8) around $\vartheta^0$ as in Newey and Smith (2004, p. 240-241).

**Step 2.** Covariance and Jacobian consistency Lemmas A.6 and A.7 apply in view Theorem 2.1, and $0 \leq ||\lambda_{n,*}|| \leq ||\hat{\lambda}_n||$ and $||\theta_{n,*} - \theta^0|| \leq ||\hat{\theta}_n - \theta^0||$. Combine that with $\rho^{(i)}(0) = -1$ for $i = 1, 2$, and uniform GEL argument Lemma A.9 to obtain $\hat{\mathcal{M}}_n = \mathcal{M}_n(1 + o_p(1))$. Now exploit expansion (A.9) to solve

$$\hat{\vartheta}_n - \vartheta^0 = -\mathcal{M}_n^{-1} \left[ \begin{array}{cc} 0' & \frac{1}{n}\sum_{t=1}^{n} \hat{m}_{n,t}^{*\prime} \end{array} \right]' \times (1 + o_p(1)).$$

By Lemma A.5 $n^{-1/2}\Sigma_n^{-1/2} \sum_{t=1}^{n} \{\hat{m}_{n,t}^* - m_{n,t}^*\} = o_p(1)$, hence by the construction of $\mathcal{A}_n$ and CLT Lemma A.8, we have that:

$$\mathcal{A}_n^{1/2}(\hat{\vartheta}_n - \vartheta^0) = -\mathcal{A}_n^{1/2} \left[ \begin{array}{c} \mathcal{H}_n \Sigma_n^{1/2}/n^{1/2} \\ \mathcal{P}_n \Sigma_n^{1/2}/n^{1/2} \end{array} \right] \Sigma_n^{-1/2} \frac{1}{n^{1/2}} \sum_{t=1}^{n} m_{n,t}^* \left(1 + o_p(1)\right) \xrightarrow{d} N\left(0, I_{q+3}\right). \tag{A.10}$$

This completes the proof. ■

## A.4 Remaining Proofs

**Proof of Lemma 4.4.** We have by direct integration and Karamata theory

$$\begin{aligned}
&\kappa_\epsilon \in (2, 2i) : E\left[\epsilon_t^{2i} I\left(|\epsilon_t| \leq c_n^{(\epsilon)}\right)\right] \sim \frac{2i}{2i - \kappa_\epsilon} d^{2i/\kappa_\epsilon} \times \left(\frac{n}{k_n^{(\epsilon)}}\right)^{2i/\kappa_\epsilon - 1} = \varpi^{(i)} \times \left(\frac{n}{k_n^{(\epsilon)}}\right)^{2i/\kappa_\epsilon - 1} \\
&\kappa_\epsilon = 2i : E\left[\epsilon_t^{2i} I\left(|\epsilon_t| \leq c_n^{(\epsilon)}\right)\right] \sim d \ln(n) \\
&\kappa_\epsilon > 2i : E\left[\epsilon_t^{2i} I\left(|\epsilon_t| \leq c_n^{(\epsilon)}\right)\right] \sim E\left[\epsilon_t^{2i}\right] - \frac{\kappa_\epsilon}{\kappa_\epsilon - 2i} d^{2i/\kappa_\epsilon} \times \left(\frac{k_n^{(\epsilon)}}{n}\right)^{1 - 2i/\kappa_\epsilon} = E\left[\epsilon_t^{2i}\right] - \xi^{(i)} \times \left(\frac{k_n^{(\epsilon)}}{n}\right)^{1 - 2i/\kappa_\epsilon}.
\end{aligned}$$

Now treat $k_n^{(\epsilon)}$ as a continuous argument $k \in [0, n)$, and write $\mathcal{E}_n^{(1)}(k) \equiv 1 - \xi^{(1)}(k/n)^{1 - 2/\kappa_\epsilon}$, etc., and

$$\mathcal{B}_n^{(GMTTM)}(k) \equiv \frac{1}{n} \frac{\mathcal{E}_n^{(2)}(k)}{\left(\mathcal{E}_n^{(1)}(k)\right)^2} \mathcal{H}\left(-a + E\left[S_t X_t' \mathcal{H} X_t'\right]\right)$$

$$\mathcal{B}_n^{(\Sigma TT)}(k) \equiv \frac{1}{n}\frac{\mathcal{E}_n^{(3)}(k)}{\mathcal{E}_n^{(1)}(k)\mathcal{E}_n^{(2)}(k)}\mathcal{H}\left(1+\frac{\rho_3}{2}\right)E\left[X_t'X_t\mathcal{P}X_t\right]$$

We have

$$\begin{aligned}\frac{\partial}{\partial k}\mathcal{B}_n^{(GMTTM)}(k) &= \left\{\frac{1}{n}\frac{\mathcal{E}_n^{(2)}(k)}{\left(\mathcal{E}_n^{(1)}(k)\right)^2}\left(\frac{\partial}{\partial k}\ln\mathcal{E}_n^{(2)}(k)-2\frac{\partial}{\partial k}\ln\mathcal{E}_n^{(1)}(k)\right)\right\}\times\mathcal{H}\left(-a+E\left[S_tX_t'\mathcal{H}X_t'\right]\right)\\ &= \mathcal{D}_n(k)\times\mathcal{H}\left(-a+E\left[S_tX_t'\mathcal{H}X_t'\right]\right).\end{aligned}$$

In order to deduce the sign of $\mathcal{D}_n(k)$, first notice $\mathcal{E}_n^{(1)} \sim 1-\xi^{(1)}(k_n^{(\epsilon)}/n)^{1-2/\kappa_\epsilon}$ and

$$\frac{\partial}{\partial k}\ln\mathcal{E}_n^{(1)}(k) = -\frac{\frac{1}{n}\left(\frac{n}{k}\right)^{2/\kappa_\epsilon}\xi^{(1)}\left(1-\frac{2}{\kappa_\epsilon}\right)}{1-\xi^{(1)}\left(\frac{k}{n}\right)^{1-2/\kappa_\epsilon}} < 0.$$

Now, if $\kappa_\epsilon \in (2,4)$ then $\mathcal{E}_n^{(2)} \sim \varpi^{(2)}(n/k_n^{(\epsilon)})^{4/\kappa_\epsilon-1} - (1-\xi^{(1)}(k_n^{(\epsilon)}/n)^{1-2/\kappa_\epsilon})^2 = o(n)$, hence for all $n \geq N$ and some $N$:

$$\frac{\partial}{\partial k}\mathcal{E}_n^{(2)}(k) = -\frac{1}{n}\left(\frac{n}{k}\right)^{4/\kappa_\epsilon}\left\{\left(\frac{4}{\kappa_\epsilon}-1\right)\varpi^{(2)}-2\left(1-\frac{2}{\kappa_\epsilon}\right)\xi^{(1)}\left(1-\xi^{(1)}\left(\frac{k}{n}\right)^{1-2/\kappa_\epsilon}\right)\left(\frac{k}{n}\right)^{2/\kappa_\epsilon}\right\} < 0.$$

It is easy to check $\mathcal{D}_n(k) > 0$ for all $k$, all $n \geq N$, and some $N$ since $\mathcal{E}_n^{(1)} \nearrow 1$ and $(\partial/\partial k)\mathcal{E}_n^{(i)} < 0$. Therefore $\mathcal{B}_{i,n}^{(GMTTM)}(k)$ and $(\partial/\partial k)\mathcal{B}_{i,n}^{(GMTTM)}(k)$ have the same sign. Similarly:

$$\begin{aligned}\frac{\partial}{\partial k}\mathcal{B}_n^{(\Sigma TT)}(k) &= \left\{\frac{1}{n}\frac{\mathcal{E}_n^{(3)}}{\mathcal{E}_n^{(1)}(k)\mathcal{E}_n^{(2)}(k)}\left(\frac{\partial}{\partial k}\ln\mathcal{E}_n^{(3)}(k)-\frac{\partial}{\partial k}\ln\mathcal{E}_n^{(1)}(k)-\frac{\partial}{\partial k}\ln\mathcal{E}_n^{(2)}(k)\right)\right\}\\ &\quad\times\mathcal{H}\left(1+\frac{\rho_3}{2}\right)E\left[X_t'X_t\mathcal{P}X_t\right]\\ &= \mathcal{F}_n(k)\times\mathcal{H}\left(1+\frac{\rho_3}{2}\right)E\left[X_t'X_t\mathcal{P}X_t\right],\end{aligned}$$

where $\mathcal{F}_n(k) > 0$ for all $k$, $n \geq N$, and some $N$. Thus, in the heavy tail case $\mathcal{B}_n^{(GMTTM)}$ and $\mathcal{B}_n^{(\Sigma TT)}$ can each be made small by using a smaller $k_n^{(\epsilon)}$.

If $\kappa_\epsilon = 4$ such that $\mathcal{E}_n^{(2)} \sim d\ln(n) - (1-\xi^{(1)}(k_n^{(\epsilon)}/n)^{1/2})^2$ then for large enough $n$:

$$\frac{\partial}{\partial k}\mathcal{E}_n^{(2)}(k) = \left(1-\xi^{(1)}\left(\frac{k}{n}\right)^{1/2}\right)\xi^{(1)}\frac{1}{n}\left(\frac{n}{k}\right)^{1/2} > 0.$$

Then $(\partial/\partial k)\ln\mathcal{E}_n^{(2)}(k) < 0$ is of order $O(n^{-1}(n/k)^{1/2}/\ln(n))$, but $(\partial/\partial k)\ln\mathcal{E}_n^{(1)}(k) < 0$ is of order

$O(n^{-1}(n/k)^{1/2})$, hence $\mathcal{D}_n(k) > 0$ for large enough $n$. Similarly $\mathcal{F}_n(k) > 0$ hence again $\mathcal{B}_n^{(GMTTM)}$ and $\mathcal{B}_n^{(\Sigma TT)}$ are small for small $k_n^{(\epsilon)}$.

Next suppose $\kappa_\epsilon > 4$ and consider bias decomposition $\mathcal{B}_n^{(GMTTM)} = \mathcal{B}_n^{(GMM)} + \mathcal{B}_n^{(TT_{GMM})}$ in (22) such that trimming only effects $\mathcal{B}_n^{(TT_{GMM})}$, write

$$\mathcal{B}_n^{(TT_{GMM})}(k) = \frac{1}{n}\left(\frac{\mathcal{E}_n^{(2)}(k)}{\left(\mathcal{E}_n^{(1)}(k)\right)^2} - \mathcal{E}^{(2)}\right)\mathcal{H}\left(-a + E\left[S_t X_t' \mathcal{H} X_t'\right]\right),$$

and note as

$$\frac{\partial}{\partial k}\mathcal{B}_n^{(GMTTM)} = \frac{\partial}{\partial k}\mathcal{B}_n^{(TT_{GMM})}(k) = \mathcal{D}_n(k) \times \mathcal{H}\left(-a + E\left[S_t X_t' \mathcal{H} X_t'\right]\right).$$

In this case $\mathcal{E}_n^{(2)} \sim (E[\epsilon_t^4] - \xi^{(2)}(k_n^{(\epsilon)}/n)^{1-4/\kappa_\epsilon}) - (1 - \xi^{(1)}(k_n^{(\epsilon)}/n)^{1-2/\kappa_\epsilon})^2$. Then

$$\begin{aligned}
&\frac{\partial}{\partial k}\ln \mathcal{E}_n^{(2)}(k) - 2\frac{\partial}{\partial k}\ln \mathcal{E}_n^{(1)}(k) \\
&= -\frac{\frac{1}{n}\left(\frac{n}{k}\right)^{2/\kappa_\epsilon}\left\{\left(1-\frac{4}{\kappa_\epsilon}\right)\xi^{(2)}\left(\frac{n}{k}\right)^{2/\kappa_\epsilon} - 2\left(1-\xi^{(1)}\left(\frac{k}{n}\right)^{1-2/\kappa_\epsilon}\right)\left(1-\frac{2}{\kappa_\epsilon}\right)\xi^{(1)}\right\}}{\mathcal{E}^{(2)} + \left(1-\xi^{(2)}\left(\frac{k}{n}\right)^{1-4/\kappa_\epsilon}\right) - \left(1-\xi^{(1)}\left(\frac{k}{n}\right)^{1-2/\kappa_\epsilon}\right)^2} + 2\frac{\frac{1}{n}\left(\frac{n}{k}\right)^{2/\kappa_\epsilon}\xi^{(1)}\left(1-\frac{2}{\kappa_\epsilon}\right)}{\mathcal{E}^{(1)} - \xi^{(1)}\left(\frac{k}{n}\right)^{1-2/\kappa_\epsilon}} \\
&= -\frac{1}{n}\left(\frac{n}{k}\right)^{2/\kappa_\epsilon}\left(\frac{\left\{\left(1-\frac{4}{\kappa_\epsilon}\right)\xi^{(2)}\left(\frac{n}{k}\right)^{2/\kappa_\epsilon} - 2\left(1-\xi^{(1)}\left(\frac{k}{n}\right)^{1-2/\kappa_\epsilon}\right)\left(1-\frac{2}{\kappa_\epsilon}\right)\xi^{(1)}\right\}}{\mathcal{E}^{(2)} + \left(1-\xi^{(2)}\left(\frac{k}{n}\right)^{1-4/\kappa_\epsilon}\right) - \left(1-\xi^{(1)}\left(\frac{k}{n}\right)^{1-2/\kappa_\epsilon}\right)^2} - \frac{2\xi^{(1)}\left(1-\frac{2}{\kappa_\epsilon}\right)}{\mathcal{E}^{(1)} - \xi^{(1)}\left(\frac{k}{n}\right)^{1-2/\kappa_\epsilon}}\right) \\
&\to -\infty \text{ as } k \to 0.
\end{aligned}$$

Therefore $\mathcal{B}_n^{(TT_{GMM})}(0) = 0$ and $(\partial/\partial k)\mathcal{B}_n^{(TT_{GMM})}(k) \to -\infty$ as $k \to 0$ hence $\mathcal{B}_n^{(TT_{GMM})}(k) < 0$ $\forall n$ in a neighborhood of 0. Further, $(\partial/\partial k)\mathcal{B}_n^{(TT_{GMM})}(k) < 0$ for any fixed $k$ and large enough $n$ hence $\mathcal{B}_n^{(TT_{GMM})}(k) < 0$ for large enough $n$. Since $k_n^{(\epsilon)}/n \to 0$ it therefore follows that $\mathcal{B}_n^{(TT_{GMM})}$ is monotonically closer to zero for smaller $k_n^{(\epsilon)}$.

It remains to characterize $\mathcal{B}_n^{(\Sigma TT)}$. By mimicking the arguments above, first it can be shown that if $\kappa_\epsilon \in (4, 6]$ then $\mathcal{B}_n^{(\Sigma TT)}$ is small for small $k_n^{(\epsilon)}$. Second, if $\kappa_\epsilon > 6$ then $\mathcal{B}_n^{(\Sigma TT)} = \mathcal{B}_n^{(\Sigma)} + \mathcal{B}_n^{(TT_\Sigma)}$where $\mathcal{B}_n^{(TT_\Sigma)}$ is monotonically closer to zero for smaller $k_n^{(\epsilon)}$. ■

**Proof of Theorem 4.6.** Note $||\Sigma_n|| \sim KE[\epsilon_{n,t}^{*4}]$ and $||\mathcal{V}_n|| \sim Kn/E[\epsilon_{n,t}^{*4}]$. Further, $k_n^{(\epsilon)} \sim n/L(n)$ implies $E|\epsilon_{n,t}^*|^p = O(L(n))$ for any $p \geq 2$ and slowly varying $L(n) \to \infty$: the bound is trivial if $E|\epsilon_t|^p < \infty$ and otherwise follows from Paretian tail decay and Karamata theory.

Consider the claim $\text{Bias}(\hat{\theta}_n^{(bc)}) = 0$. Let $\{\widehat{\mathcal{H}}_n, \hat{a}_n, \widehat{\mathcal{P}}_n\}$ denote $\{\widehat{\mathcal{H}}_n^{(\pi)}, \hat{a}_n^{(\pi)}, \widehat{\mathcal{P}}_n^{(\pi)}\}$ with $\hat{\pi}_{n,t}^*(\hat{\theta}_n)$ replaced with $1/n$, and define $\mathcal{E}_{1,n}^* = 1/n\sum_{t=1}^n \epsilon_{n,t}^{*2}(\hat{\theta}_n)$ and $\mathcal{E}_{i,n}^* = 1/n\sum_{t=1}^n(\epsilon_{n,t}^{*2}(\hat{\theta}_n) - \mathcal{E}_{1,n}^*)^i$ for $i = 2, 3$. By the argument used to prove Lemma A.5 we can replace $\hat{\epsilon}_{n,t}$ with $\epsilon_{n,t}^*$, and by Theorem

2.1 and A.12 we can replace $\hat{\pi}^*_{n,t}(\hat{\theta}_n)$ with $1/n$. In particular:

$$\mathcal{V}_n^{1/2}(\hat{\mathcal{B}}_n(\hat{\theta}_n) - \mathcal{B}^*_n(\hat{\theta}_n)) = o_p(1) \tag{A.11}$$

where

$$\mathcal{B}^*_n(\hat{\theta}_n) = \frac{1}{n}\frac{\mathcal{E}^*_{2,n}}{\left(\mathcal{E}^*_{1,n}\right)^2}\widehat{\mathcal{H}}_n\left(-\hat{a}_n + \frac{1}{n}\sum_{t=1}^n S_t X_t'\widehat{\mathcal{H}}_n X_t'\right) + \frac{1}{n}\frac{\mathcal{E}^*_{3,n}}{\mathcal{E}^*_{1,n}\mathcal{E}^*_{2,n}}\widehat{\mathcal{H}}_n\left(1+\frac{\rho_3}{2}\right)\frac{1}{n}\sum_{t=1}^n X_t'X_t\widehat{\mathcal{P}}_n X_t.$$

$\text{Bias}(\hat{\theta}^{(bc)}_n) = 0$ can therefore be shown by applying the method of proof of Theorems 4.1 and 4.2 to the argument used in Newey and Smith (2004, proof of Theorem 5.1).

The remaining claim $\mathcal{V}_n^{1/2}(\hat{\theta}^{(bc)}_n - \theta^0)$ follows if we show $\mathcal{V}_n^{1/2}\hat{\mathcal{B}}_n(\hat{\theta}_n) = o_p(1)$. Define $\mathcal{B}_n \equiv \text{Bias}(\hat{\theta}_n)$. We have

$$\left\|\mathcal{V}_n^{1/2}\hat{\mathcal{B}}_n(\hat{\theta}_n)\right\| \leq \left(\frac{n}{E\left[\epsilon^{*4}_{n,t}\right]}\right)^{1/2}\|\mathcal{B}_n\| + \left(\frac{n}{E\left[\epsilon^{*4}_{n,t}\right]}\right)^{1/2}\left\|\hat{\mathcal{B}}_n(\hat{\theta}_n) - \mathcal{B}_n\right\|. \tag{A.12}$$

By Corollary 4.3 and $E|\epsilon^*_{n,t}|^p = O(L(n))$ for any $p \geq 2$ the first term in (A.12) satisfies:

$$\begin{aligned}\left(\frac{n}{E\left[\epsilon^{*4}_{n,t}\right]}\right)^{1/2}\|\mathcal{B}_n\| &\leq K\left(\frac{n}{E\left[\epsilon^{*4}_{n,t}\right]}\right)^{1/2}\times\frac{1}{n}E\left[\epsilon^{*4}_{n,t}\right] + K\left(\frac{n}{E\left[\epsilon^{*4}_{n,t}\right]}\right)^{1/2}\frac{1}{n}\frac{E\left[\epsilon^{*6}_{n,t}\right]}{E\left[\epsilon^{*4}_{n,t}\right]} \\ &= K\left(\frac{1}{n}E\left[\epsilon^{*4}_{n,t}\right]\right)^{1/2} + K\frac{1}{n^{1/2}}\frac{E\left[\epsilon^{*6}_{n,t}\right]}{\left(E\left[\epsilon^{*4}_{n,t}\right]\right)^{3/2}} = o(1).\end{aligned} \tag{A.13}$$

Next, use (A.11) to deduce the second term in (A.12) satisfies

$$\begin{aligned}&\left(\frac{n}{E\left[\epsilon^{*4}_{n,t}\right]}\right)^{1/2}\left\|\hat{\mathcal{B}}_n(\hat{\theta}_n) - \mathcal{B}_n\right\| \\ &\leq \frac{1}{\left(E\left[\epsilon^{*4}_{n,t}\right]\right)^{1/2}n^{1/2}}\left\|\frac{\mathcal{E}^*_{2,n}}{\left(\mathcal{E}^*_{1,n}\right)^2}\widehat{\mathcal{H}}^{(\pi)}_n\left(-\hat{a}_n + \frac{1}{n}\sum_{t=1}^n S_t X_t'\widehat{\mathcal{H}}^{(\pi)}_n X_t'\right) - \frac{\mathcal{E}^{(2)}_n}{\left(\mathcal{E}^{(1)}_n\right)^2}\mathcal{H}\left(-a + E\left[S_t X_t'\mathcal{H}X_t'\right]\right)\right\| \\ &\quad + \frac{1}{\left(E\left[\epsilon^{*4}_{n,t}\right]\right)^{1/2}n^{1/2}}\left\|\frac{\mathcal{E}^*_{3,n}}{\mathcal{E}^*_{1,n}\mathcal{E}^*_{2,n}}\widehat{\mathcal{H}}^{(\pi)}_n\left(1+\frac{\rho_3}{2}\right)\frac{1}{n}\sum_{t=1}^n X_t'X_t\widehat{\mathcal{P}}^{(\pi)}_n X_t - \frac{\mathcal{E}^{(3)}_n}{\mathcal{E}^{(1)}_n\mathcal{E}^{(2)}_n}\mathcal{H}\left(1+\frac{\rho_3}{2}\right)E\left[X_t'X_t\mathcal{P}X_t\right]\right\| \\ &= \mathcal{A}_{1,n} + \mathcal{A}_{2,n}.\end{aligned}$$

By using the limit theory developed in Appendix A.2 it can be shown $\mathcal{E}^*_{i,n}/\mathcal{E}^{(i)}_n = 1 + o_p(1)$, $\widehat{\mathcal{H}}^{(\pi)}_n = \mathcal{H} + o_p(1)$, $\hat{a}_n = a + o_p(1)$, and

$$\frac{1}{n}\sum_{t=1}^n S_t X_t'\widehat{\mathcal{H}}^{(\pi)}_n X_t' = E\left[S_t X_t'\mathcal{H}X_t'\right] + o_p(1) \text{ and } \frac{1}{n}\sum_{t=1}^n X_t'X_t\widehat{\mathcal{P}}^{(\pi)}_n X_t = E\left[X_t'X_t\mathcal{P}X_t\right] + o_p(1).$$

Therefore, coupled with $E|\epsilon^*_{n,t}|^p = O(L(n))$, it follows:

$$\begin{aligned}\mathcal{A}_{1,n} &\leq K\frac{\mathcal{E}^{(2)}_n}{\left(E\left[\epsilon^{*4}_{n,t}\right]\right)^{1/2} n^{1/2}}\left|\frac{\mathcal{E}^{(1)}_n}{\left(\mathcal{E}^*_{1,n}\right)^2}\frac{\mathcal{E}^*_{2,n}}{\mathcal{E}^{(2)}_n} - 1\right| \qquad (A.14)\\ &\sim K\frac{E\left[\epsilon^{*4}_{n,t}\right]}{\left(E\left[\epsilon^{*4}_{n,t}\right]\right)^{1/2} n^{1/2}}\left|\frac{\mathcal{E}^{(1)}_n}{\left(\mathcal{E}^*_{1,n}\right)^2}\frac{\mathcal{E}^*_{2,n}}{\mathcal{E}^{(2)}_n} - 1\right| = o_p\left(\left(\frac{E\left[\epsilon^{*4}_{n,t}\right]}{n}\right)^{1/2}\right) = o_p\left(1\right).\end{aligned}$$

Similarly $\mathcal{A}_{2,n} = o_p(1)$. Combine (A.12)-(A.14) to prove the required result. ■

**Proof of Theorem 5.1.** The claim follows from covariance and Jacobian consistency Lemmas A.6 and A.7. ■

**Proof of Theorems 5.2 and 5.3.** The claims for $\mathcal{W}_n \equiv \mathcal{R}(\hat{\theta}_n)'[\mathcal{D}(\hat{\theta}_n)\hat{\mathcal{V}}_n(\hat{\theta}_n)^{-1}\mathcal{D}(\hat{\theta}_n)']^{-1}\mathcal{R}(\hat{\theta}_n)$ follow from continuity of $\mathcal{D}(\theta)$ and $\mathcal{R}(\theta)$, Theorems 2.1, 2.2, and 5.1, and the mapping theorem.

Now consider the likelihood ratio statistic $\mathcal{LR}_n = 2n\hat{Q}(\hat{\theta}_n, \hat{\lambda}_n)$ . Define $\hat{m}^*_n(\theta) \equiv 1/n\sum_{t=1}^n \hat{m}^*_{n,t}(\theta)$, $m^*_n(\theta) \equiv 1/n\sum_{t=1}^n m^*_{n,t}(\theta)$, and $\mathcal{H}_n \equiv (\mathcal{J}_n'\Sigma_n^{-1}\mathcal{J}_n)^{-1}\mathcal{J}_n'\Sigma_n^{-1}$. Similar to Newey and Smith (2004, p. 240-241), by a second order Taylor expansion of $\hat{Q}(\hat{\theta}_n, \hat{\lambda}_n)$ around $\lambda = 0$, with $\mathring{\rho}^{(i)}_{n,t} \equiv \rho^{(i)}(\lambda'_{n,*}\hat{m}^*_{n,t}(\hat{\theta}_n))$ and $||\lambda_{n,*}|| \leq ||\hat{\lambda}_n||$,

$$2n\hat{Q}(\hat{\theta}_n, \hat{\lambda}_n) = 2n\left[-\hat{\lambda}_n'\hat{m}^*_n(\hat{\theta}_n) + \frac{1}{2}\hat{\lambda}_n'\frac{1}{n}\sum_{t=1}^n \mathring{\rho}^{(2)}_{n,t}\hat{m}^*_{n,t}(\hat{\theta}_n)\hat{m}^*_{n,t}(\hat{\theta}_n)'\hat{\lambda}_n\right]. \qquad (A.15)$$

Use (A.10) to deduce $n^{1/2}\mathcal{P}_n^{-1/2}\hat{\lambda}_n = -n^{1/2}\mathcal{P}_n^{-1/2}\mathcal{P}_n m^*_n \times (1 + o_p(1)) + o_p(1)$, hence $\hat{\lambda}_n = -\mathcal{P}_n m^*_n \times (1 + o_p(1))$. Further, by the same argument following expansion (A.9), coupled with uniform approximation and Jacobian consistency Lemmas A.5 and A.7, and estimator expansion (A.10):

$$\begin{aligned}\hat{m}^*_n(\hat{\theta}_n) &= m^*_n + \mathcal{J}_n'\left(\hat{\theta}_n - \theta^0\right) \times (1 + o_p(1)) = m^*_n + \mathcal{J}_n'\mathcal{V}_n^{-1/2}\mathcal{V}_n^{1/2}\mathcal{H}_n m^*_n \times (1 + o_p(1))\\ &= m^*_n + \mathcal{J}_n'\mathcal{H}_n m^*_n \times (1 + o_p(1)) = m^*_n + \mathcal{J}_n'\left(\mathcal{J}_n'\Sigma_n^{-1}\mathcal{J}_n\right)^{-1}\mathcal{J}_n'\Sigma_n^{-1}m^*_n, \times (1 + o_p(1)),\end{aligned}$$

hence $\Sigma_n^{-1}\hat{m}^*_n(\hat{\theta}_n) = \mathcal{P}_n m^*_n \times (1 + o_p(1))$. Therefore $\hat{\lambda}_n = -\Sigma_n^{-1}\hat{m}^*_n(\hat{\theta}_n) \times (1 + o_p(1))$. Plug the latter into (A.15) and invoke covariance consistency Lemma A.6 twice to deduce

$$\mathcal{LR}_n = 2n\hat{Q}(\hat{\theta}_n, \hat{\lambda}_n) = n\hat{m}^*_n(\hat{\theta}_n)'\Sigma_n^{-1}\hat{m}^*_n(\hat{\theta}_n) \times (1 + o_p(1)) = n\hat{m}^*_n(\hat{\theta}_n)'\hat{\Sigma}_n^{-1}\hat{m}^*_n(\hat{\theta}_n) \times (1 + o_p(1)).$$

The limit for $\mathcal{LR}_n$ under Assumption A and the null hypothesis $E[(\epsilon_t^2 - 1)w_t] = 0$ now follows from covariance consistency Lemma A.6, and Theorem 2.1 in Hill and Aguilar (2013). Conversely, if $E[(\epsilon_t^2 - 1)w_t] \neq 0$ then it is straightforward to alter the proof of Theorem 2.1 to show under Assumption A that there exists a unique point $\tilde{\theta} \in \Theta$ satisfying $\hat{\theta}_n \xrightarrow{p} \tilde{\theta}$ where $E[m_t(\tilde{\theta})] - m = 0$ for some non-zero $m \in \mathbb{R}^q$. This follows since $\epsilon_t(\tilde{\theta})$ is square integrable, and and $\{\epsilon_t(\tilde{\theta}), x_t(\tilde{\theta})\}$ are stationary and geometrically on $\Theta$. Lemmas A.5 and A.7 can be modified accordingly in view of stationarity. The claim can then be proven along the lines of Theorem

2.2 in Hill and Aguilar (2013). The remaining claims for $\mathcal{LM}_n$ and $\mathcal{S}_n$ follow similarly. ■

**Proof of Theorem 6.1.** The following extends arguments in Bonnal and Renault (2004, Corollary 3.6), Smith (2011, Theorem 3.1), and Hill and Aguilar (2013, proof of Theorem 2.1). Since $g_t$ is $\Im_t$-measurable, stationary, continuous and differentiable on $\Theta$-$a.e.$, it suffices to work with $[g^*_{n,t}(\theta)', m^*_{n,t}(\theta)']'$ throughout in view of approximation theory for tail-trimmed equations developed in Hill (2015a,b, 2013) and Hill and Aguilar (2013). We therefore need only prove $n^{1/2}\mathfrak{V}_n^{-1/2}(\overline{g}^{*(\pi)}_n(\hat{\theta}_n) - E[g^*_{n,t}]) \xrightarrow{d} N(0, I_h)$ where $\overline{g}^{*(\pi)}_n(\theta) = \sum_{t=1}^n \hat{\pi}^*_{n,t}(\hat{\theta}_n) g^*_{n,t}(\hat{\theta}_n)$.

By a Taylor expansion of $\hat{\pi}^*_{n,t}$ around $\lambda = 0$, use Lemma A.9 to deduce

$$\hat{\pi}^*_{n,t} = \frac{1}{n} + \left\{ \frac{\mathring{\rho}^{(2)}_{n,t} \hat{\lambda}_n' m^*_{n,t}(\hat{\theta}_n)}{\sum_{s=1}^n \mathring{\rho}^{(1)}_{n,s}} - \frac{\mathring{\rho}^{(1)}_{n,t} \sum_{s=1}^n \mathring{\rho}^{(2)}_{n,s} m_{n,s}(\hat{\theta}_n)' \hat{\lambda}_n}{\left[\sum_{s=1}^n \mathring{\rho}^{(1)}_{n,s}\right]^2} \right\} \times (1 + o_p(1))$$

where $\mathring{\rho}^{(i)}_{n,t} \equiv \rho^{(i)}(\lambda_{n,*}' \hat{m}^*_{n,t}(\hat{\theta}_n))$ and $||\lambda_{n,*}|| \leq ||\hat{\lambda}_n||$. By virtue of Lemmas A.9 and A.11 and Theorem 2.1 we have $\max_{1 \leq t \leq n} |\mathring{\rho}^{(i)}_{n,t} + 1| \xrightarrow{p} 0$, $||m^*_n(\hat{\theta}_n)|| = O_p(||\Sigma_n||^{1/2}/n^{1/2})$, and $\hat{\lambda}_n = O_p(||\Sigma_n||^{-1/2} n^{-1/2})$, and from Lemma A.12 $\sup_{\theta \in \Theta} |\sum_{t=1}^n \hat{\pi}^*_{n,t}(\theta) - 1| = O_p(||\Sigma_n||^{-1/2} n^{-1/2}) = O_p(n^{-1/2})$. Hence

$$\hat{\pi}^*_{n,t}(\hat{\theta}_n) = \frac{1}{n} + \frac{1}{n}\left\{ \hat{\lambda}_n' m^*_{n,t}(\hat{\theta}_n) \times (1 + o_p(1)) + O_p(n^{-1}) \right\} \quad \text{and} \quad \sum_{t=1}^n \hat{\pi}^*_{n,t}(\hat{\theta}_n) = 1 + O_p(n^{-1/2}). \tag{A.16}$$

Arguments similar to the first order expansion (A.9) in the proof of Theorem 2.2, and covariance consistency Lemma A.6, can be used to verify

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^n \left( g^*_{n,t}(\hat{\theta}_n) - E[g^*_{n,t}] \right) \times m^*_{n,t}(\hat{\theta}_n)' = \Gamma_n \times (1 + o_p(1)) \\
&\frac{1}{n}\sum_{t=1}^n \left( g^*_{n,t}(\hat{\theta}_n) - E[g^*_{n,t}] \right) = \frac{1}{n}\sum_{t=1}^n \left( g^*_{n,t} - E[g^*_{n,t}] \right) + G_n \left( \hat{\theta}_n - \theta^0 \right) \times (1 + o_p(1)).
\end{aligned}$$

Hence

$$\begin{aligned}
&\sum_{t=1}^n \hat{\pi}^*_{n,t}(\hat{\theta}_n) \left\{ g^*_{n,t}(\hat{\theta}_n) - E\left[g^*_{n,t}\right] \right\} \qquad (A.17) \\
&= \frac{1}{n}\sum_{t=1}^n \left( g^*_{n,t}(\hat{\theta}_n) - E\left[g^*_{n,t}\right] \right) + \frac{1}{n}\sum_{t=1}^n \left( g^*_{n,t}(\hat{\theta}_n) - E\left[g^*_{n,t}\right] \right) \times m^*_{n,t}(\hat{\theta}_n)' \hat{\lambda}_n \times (1 + o_p(1)) \\
&\quad + \frac{1}{n}\sum_{t=1}^n \left( g^*_{n,t}(\hat{\theta}_n) - E\left[g^*_{n,t}\right] \right) \times O_p(1/n) \\
&= \frac{1}{n}\sum_{t=1}^n \left( g^*_{n,t} - E\left[g^*_{n,t}\right] \right) + G_n(\hat{\theta}_n - \theta^0) \times (1 + o_p(1)) + \Gamma_n \hat{\lambda}_n \times (1 + o_p(1)).
\end{aligned}$$

Moreover, by the proof of Theorem 2.2:

$$\left[\begin{array}{c}\hat{\theta}_n-\theta^0\\ \hat{\lambda}_n\end{array}\right]=-\left[\begin{array}{c}\mathcal{H}_n\\ \mathcal{P}_n\end{array}\right]\frac{1}{n}\sum_{t=1}^{n}m_{n,t}^{*}\times(1+o_p(1))\,. \tag{A.18}$$

Combine (A.17) and (A.18) to deduce:

$$\overline{g}_n^{*(\pi)}(\hat{\theta}_n)-E[g_t]=[I_h,-G_n\mathcal{H}_n-\Gamma_n\mathcal{P}_n]\frac{1}{n}\sum_{t=1}^{n}\left[\begin{array}{c}g_{n,t}^{*}-E[g_t]\\ m_{n,t}^{*}\end{array}\right]\times(1+o_p(1))\,.$$

In conjunction with the supposition $n^{1/2}\mathfrak{V}_n^{-1/2}\{E[g_{n,t}^{*}]-E[g_t]\}\to 0$, and by the construction of $\mathfrak{V}_n$, it now follows:

$$\begin{aligned}
&n^{1/2}\mathfrak{V}_n^{-1/2}\left(\overline{g}_n^{*(\pi)}(\hat{\theta}_n)-E[g_t]\right)\\
&\quad=\mathfrak{V}_n^{-1/2}\times[I_h,-G_n\mathcal{H}_n-\Gamma_n\mathcal{P}_n]\left[\begin{array}{cc}\Upsilon_n&\Gamma_n\\ \Gamma_n'&\Sigma_n\end{array}\right]^{1/2}\\
&\qquad\times\left[\begin{array}{cc}\Upsilon_n&\Gamma_n\\ \Gamma_n'&\Sigma_n\end{array}\right]^{-1/2}\frac{1}{n^{1/2}}\sum_{t=1}^{n}\left[\begin{array}{c}g_{n,t}^{*}-E[g_{n,t}^{*}]\\ m_{n,t}^{*}\end{array}\right]\times(1+o_p(1))\\
&\quad=\left[\begin{array}{cc}\Upsilon_n&\Gamma_n\\ \Gamma_n'&\Sigma_n\end{array}\right]^{-1/2}\frac{1}{n^{1/2}}\sum_{t=1}^{n}\left[\begin{array}{c}g_{n,t}^{*}-E[g_{n,t}^{*}]\\ m_{n,t}^{*}\end{array}\right]\times(1+o_p(1))\,.
\end{aligned}$$

Recall $E[m_{n,t}^{*}]=0$ by the martingale difference property. Therefore, by measurability and the geometrically $\beta$-mixing property, a generalization of CLT Lemma A.5 in Hill (2015a), cf. Lemma B.6 in Hill and Aguilar (2013), extends to $[g_{n,t}^{*}-E[g_{n,t}^{*}],m_{n,t}^{*\prime}]'$, hence $n^{1/2}\mathfrak{V}_n^{-1/2}(\overline{g}_n^{*(\pi)}(\hat{\theta}_n)-E[g_{n,t}^{*}])\xrightarrow{d}N(0,I_h)$.

Finally, $n^{1/2}\mathfrak{V}_n^{-1/2}\{E[g_{n,t}^{*}]-E[g_t]\}\to 0$ holds in the special case $\max\{\kappa_1^{(g)},\kappa_2^{(g)}\}\geq 2$ and $k_{i,n}^{(g)}\to\infty$ at a slowly varying rate. See Corollary 1.3 in Hill (2015b). ■

**Proof of Theorem 7.1.** In order to reduce notation we drop the risk level $\alpha$, and we write $k_n=k_n^{(y)}$.

**Claim (a).** We prove the claim for the bias-corrected estimator $\widehat{ES}_n^{(bc)(\pi)}$. Following arguments in Hill (2015b), by using $m_n/k_n\to\infty$ it can be shown $(n^{1/2}/\mathfrak{S}_n^{1/2})\{\widehat{ES}_n^{(bc*)(\pi)}-\widehat{ES}_{n,\alpha}^{(bc)(\pi)}\}\xrightarrow{p}0$.. Define

$$\begin{aligned}
&\mathcal{I}_{n,t}\equiv\left(\frac{n}{k_n^{(y)}}\right)^{1/2}\left(I\left(y_t\leq -l_n\right)-E\left[\left(y_t\leq -l_n\right)\right]\right)\quad\text{and}\quad\mathcal{B}_n^{*}\equiv\frac{1}{\kappa_1-1}\frac{k_n^{(y)}}{n}l_n^{(y)}\\
&\hat{g}_{n,t}^{*}\equiv y_tI\left(y_{(k_n^{(y)})}^{(-)}\leq y_t\leq y_{[\alpha n]}\right)\quad\text{and}\quad g_{n,t}^{*}\equiv y_tI\left(-l_n^{(y)}\leq y_t\leq q_\alpha\right).
\end{aligned}$$

We first show the limit distribution of $(n^{1/2}/\mathfrak{S}_n^{1/2})\{\sum_{t=1}^{n}\hat{\pi}_{n,t}^{*}\hat{g}_{n,t}^{*}+\hat{\mathcal{B}}_n-E[g_t]\}$ is identical

to the distribution limit of

$$\frac{n^{1/2}}{\mathfrak{S}_n^{1/2}}\left(\sum_{t=1}^{n}\hat{\pi}_{n,t}^{*}g_{n,t}^{*}+\mathcal{B}_n^{*}-E[g_t]\right) \tag{A.19}$$
$$=\frac{n^{1/2}}{\mathfrak{S}_n^{1/2}}\left(\sum_{t=1}^{n}\hat{\pi}_{n,t}^{*}g_{n,t}^{*}-E\left[g_{n,t}^{*}\right]+\frac{1}{\kappa_1-1}\left(\frac{k_n^{(y)}}{n}\right)^{1/2}l_n^{(y)}\frac{1}{n}\sum_{t=1}^{n}\mathcal{I}_{n,t}\right).$$

The property $m_n/k_n^{(y)}\to\infty$ can be shown to ensure $\hat{\kappa}_{1,m_n}$ does not affect asymptotics be replicating arguments in Hill (2015b, proof of Theorem 2.2), hence $\hat{\mathcal{B}}_n$ can be replaced with $\mathcal{B}_n^{*}$ for asymptotic arguments. Similarly, $I(y_{(k_n^{(y)})}^{(-)}\leq y_t\leq y_{[\alpha n]})$ can be replaced with $I(-l_n^{(y)}\leq y_t\leq q_\alpha)$, cf. Hill (2015a,b, 2013) and Hill and Aguilar (2013). Moreover, $(k_n^{1/2}/n)l_n^{(y)}=K(k_n/n)^{1-1/\kappa_1}/k_n^{1/2}\to 0$ given $\kappa_1>1$, and by arguments presented in Hill (2015a,b, 2013): $k_n^{1/2}(y_{(k_n)}^{(-)}/l_n^{(y)}+1)=\kappa_1^{-1}n^{-1/2}\sum_{t=1}^{n}\mathcal{I}_{n,t}+o_p(1)$. Finally, by arguments in Peng (2001, p. 259-264) it can be shown $(n/\mathfrak{S}_n)^{1/2}\{E\left[g_{n,t}^{*}\right]+(\kappa_1-1)^{-1}(k_n/n)l_n^{(y)}-E[g_t]\}\to 0$. The preceding properties together prove (A.19).

Next, use the fact that $g_{n,t}^{*}$ is not a function of $\theta$ to deduce from the proof of Theorem 6.1:

$$\sum_{t=1}^{n}\hat{\pi}_{n,t}^{*}\left(g_{n,t}^{*}+E\left[g_{n,t}^{*}\right]\right)=[1,-\Gamma_n\mathcal{P}_n]\frac{1}{n}\sum_{t=1}^{n}\begin{bmatrix} g_{n,t}^{*}-E\left[g_{n,t}^{*}\right]\\ m_{n,t}^{*}\end{bmatrix}\times(1+o_p(1)). \tag{A.20}$$

Combine (A.19) and (A.20) to obtain by asymptotic equivalence:

$$\frac{n^{1/2}}{\mathfrak{S}_n^{1/2}}\left(\sum_{t=1}^{n}\hat{\pi}_{n,t}^{*}\hat{g}_{n,t}^{*}+\hat{\mathcal{B}}_n-E[g_t]\right)$$
$$=\frac{n^{1/2}}{\mathfrak{S}_n^{1/2}}\left[1,-\Gamma_n\mathcal{P}_n,\frac{1}{\kappa_1-1}\left(\frac{k_n}{n}\right)^{1/2}l_n^{(y)}\right]\frac{1}{n}\sum_{t=1}^{n}\begin{bmatrix} g_{n,t}^{*}-E\left[g_{n,t}^{*}\right]\\ m_{n,t}^{*}\\ \mathcal{I}_{n,t}\end{bmatrix}\times(1+o_p(1)).$$

Therefore, by the definitions of $\mathcal{D}_n$, $\mathcal{W}_{n,t}$, $\mathfrak{W}_n$, and $\mathfrak{S}_n$, and by a generalization of CLT Lemma A.5 in Hill (2015a), cf. Lemma B.6 in Hill and Aguilar (2013):

$$\frac{n^{1/2}}{\mathfrak{S}_n^{1/2}}\left(\sum_{t=1}^{n}\hat{\pi}_{n,t}^{*}\hat{g}_{n,t}^{*}+\hat{\mathcal{B}}_n-E[g_t]\right)=\left(\frac{1}{\mathfrak{S}_n^{1/2}}\mathcal{D}_n'\mathfrak{W}_n^{1/2}\right)\mathfrak{W}_n^{-1/2}\frac{1}{n^{1/2}}\sum_{t=1}^{n}\mathcal{W}_{n,t}\times(1+o_p(1))\xrightarrow{d}N(0,1).$$

**Claims (b) and (c).** Write $\mathfrak{P}_n\equiv P(|\widehat{ES}_{n,\alpha}^{(bc*)(\pi)}-\widetilde{ES}_{n,\alpha}^{(\pi)}|<|\widehat{ES}_{n,\alpha}^{*(\pi)}-\widetilde{ES}_{n,\alpha}^{(\pi)}|)$. Since $k_n^{(y)}=o((\ln(n))^a)$ for some $a>0$, then $(n/\mathfrak{S}_n)^{1/2}|\mathcal{B}_n|\to\infty$ if $\kappa_1<2$ and $(n/\mathfrak{S}_n)^{1/2}|\mathcal{B}_n|\to 0$ if $\kappa_1\geq 2$ by using the order of $\mathfrak{S}_n$ derived in Step 1, and arguments in Hill (2015b, Section 1). Both claims are therefore proved in Step 2 if we show $\mathfrak{P}_n\to 1$ when $\kappa_1<2$, and $\mathfrak{P}_n\to 0$ when $\kappa_1\geq 2$.

**Step 1.** We first determine the order of $\mathfrak{S}_n$. By Lemma A.1 in Hill (2015b) $1/n\sum_{s,t=1}^n E[\mathcal{Y}^*_{n,s}\mathcal{Y}^*_{n,t}]$ $= E[\mathcal{Y}^{*2}_{n,t}] \times O(r_n)$ and $1/n\sum_{s,t=1}^n E[\mathcal{I}_{n,s}\mathcal{I}_{n,t}] = E[\mathcal{I}^2_{n,t}] \times O(\tilde{r}_n) = O(\tilde{r}_n)$, where $\{r_n, \tilde{r}_n\}$ are sequences of positive numbers, $r_n = O(\ln(n))$, $r_n = O(1)$ if $\kappa_1 > 2$, and $\tilde{r}_n = O(1)$. Therefore:

$$\begin{aligned}
\mathfrak{S}_n &\sim K\left(\Upsilon_n - \Gamma_n\mathcal{P}_n\Gamma_n + K\left(\frac{k_n^{(y)}}{n}\right)^{1-2/\kappa_1}\mathfrak{I}_n\right) \\
&= K\left(\frac{1}{n}\sum_{s,t=1}^n E\left[\mathcal{Y}^*_{n,s}\mathcal{Y}^*_{n,t}\right] - \Gamma_n\mathcal{P}_n\Gamma'_n + K\left(\frac{k_n^{(y)}}{n}\right)^{1-2/\kappa_1}\frac{1}{n}\sum_{s,t=1}^n E\left[\mathcal{I}_{n,s}\mathcal{I}_{n,t}\right]\right) \\
&\sim K\left(E\left[\mathcal{Y}^{*2}_{n,t}\right]\left(r_n - \tilde{r}_n\frac{\left\|E\left[\mathcal{Y}^*_{n,t}m^{*\prime}_{n,t}\right]\right\|^2}{E\left[\mathcal{Y}^{*2}_{n,t}\right]\|\Sigma_n\|}\right) + K\left(\frac{k_n^{(y)}}{n}\right)^{1-2/\kappa_1}\right) \\
&\sim K\left(E\left[\mathcal{Y}^{*2}_{n,t}\right](r_n - K) + K\left(\frac{k_n^{(y)}}{n}\right)^{1-2/\kappa_1}\right).
\end{aligned}$$

Now use Karamata theory, and $k_n^{(y)} = o(m_n)$, to deduce $\mathfrak{S}_n \sim K$ if $\kappa_1 > 2$, $\mathfrak{S}_n = O((\ln(n))^2)$ if $\kappa_1 = 2$, and if $\kappa_1 < 2$ then

$$\mathfrak{S}_n \sim K\left(\left(\frac{n}{m_n}\right)^{2/\kappa_1-1}O\left(\ln(n)\right) + K\left(\frac{n}{k_n^{(y)}}\right)^{2/\kappa_1-1}\right) = K\left(\frac{n}{k_n^{(y)}}\right)^{2/\kappa_1-1}(1+o(1)).$$

**Step 2:** Observe:

$$\begin{aligned}
\mathfrak{P}_n &= P\left(\left|\widehat{ES}^{(bc*)(\pi)}_{n,\alpha} - \widetilde{ES}_{n,\alpha}\right| < \left|\widehat{ES}^{*(\pi)}_{n,\alpha} - \widetilde{ES}_{n,\alpha}\right|\right) \\
&= P\left(\left|\left(\frac{n}{\mathfrak{S}_n}\right)^{1/2}\left\{\frac{1}{\alpha}\sum_{t=1}^n \hat{\pi}^*_{n,t}y_tI\left(y_t < y^{(-)}_{(k_n^{(y)})}\right) - \hat{\mathcal{B}}_n\right\}\right| \right. \\
&\qquad \left. < \left|\left(\frac{\mathfrak{V}_n}{\mathfrak{S}_n}\right)^{1/2}\left(\frac{n}{\mathfrak{V}_n}\right)^{1/2}\left\{\frac{1}{\alpha}\sum_{t=1}^n \hat{\pi}^*_{n,t}y_tI\left(y_t < y^{(-)}_{(k_n^{(y)})}\right) - \mathcal{B}_n\right\} + \left(\frac{n}{\mathfrak{S}_n}\right)^{1/2}\mathcal{B}_n\right|\right) \\
&= P\left(|\mathcal{Z}_{1,n}| < \left|\left(\frac{\mathfrak{V}_n}{\mathfrak{S}_n}\right)^{1/2}\mathcal{Z}_{2,n} + \left(\frac{n}{\mathfrak{S}_n}\right)^{1/2}\mathcal{B}_n\right|\right),
\end{aligned}$$

say, where $\mathfrak{V}_n$ is the scale for $\widehat{ES}^{*(\pi)}_{n,\alpha}$. In view of Claim (a), and Theorem 6.1, each $\mathcal{Z}_{i,n} \xrightarrow{d} N(0,1)$. If $\kappa_1 < 2$ then $(n/\mathfrak{S}_n)^{1/2}|\mathcal{B}_n| \to \infty$ hence $\mathfrak{P}_n \to 1$. If $\kappa_1 > 2$ then $(n/\mathfrak{S}_n)^{1/2}|\mathcal{B}_n| \to 0$, and $|\mathcal{Z}_{1,n} - \mathcal{Z}_{2,n}| \xrightarrow{p} 0$ and $\mathfrak{V}_n/\mathfrak{S}_n \to 1$ follow by noting $(k_n^{(y)}/n)^{1/2}l_n^{(y)} = K(k_n^{(y)}/n)^{1/2-1/\kappa_1} \to 0$, hence $\mathfrak{S}_n = \mathfrak{V}_n + o(1)$. Then for some standard normal random variable $\mathcal{Z}$, $\mathfrak{P}_n = P(|\mathcal{Z} + o_p(1)| < |\mathcal{Z} + o_p(1)|) \to 0$. The case $\kappa_1 = 2$ resulting in $\mathfrak{P}_n \to 0$ is similar. ■

# B Appendix: Tail-Trimmed Equations for ARMA-GARCH

We now show how to construct robust estimating equations for an ARMA(1,1)-GARCH(1,1) model. An extension to ARMA$(p,q)$-GARCH$(r,s)$ is identical. The model is

$$y_t = w^0 + a^0 y_{t-1} + b^0 u_{t-1} + u_t \text{ and } u_t = \sigma_t \epsilon_t,\ \left|a^0\right| < 1,\ \epsilon_t \text{ is iid},\ E[\epsilon_t] = 0,\ E[\epsilon_t^2] = 1 \qquad \text{(B.1)}$$

$$\sigma_t^2 = \omega^0 + \alpha^0 u_{t-1}^2 + \beta^0 \sigma_{t-1}^2, \text{ where } \omega^0 > 0,\ \alpha^0, \beta^0 \geq 0,\ \alpha^0 + \beta^0 > 0,\ E[\ln(\alpha^0 + \beta^0 \epsilon_t^2)] < \infty.$$

Assume $a^0 \neq -b^0$ to rule out common roots. Collect ARMA parameters $\psi \equiv [w, a, b]'$ and GARCH parameters $\delta \equiv [\omega, \alpha, \beta]$ and write $\theta \equiv [\psi', \delta']$.

Define $u_t(\psi) \equiv y_t - w - a y_{t-1} - b u_{t-1}(\psi)$, $\sigma_t^2(\theta) = \omega + \alpha u_{t-1}^2 + \beta \sigma_{t-1}^2(\theta)$, $\epsilon_t(\theta) = u_t(\psi)/\sigma_t(\theta)$ and

$$\begin{aligned} \upsilon_{\psi,t}(\psi) &\equiv \frac{\partial}{\partial \psi} u_t(\psi) \text{ hence } \upsilon_{\psi,t}(\psi) = -\left[1, y_{t-1}, u_{t-1}(\psi)\right]' - b \upsilon_{\psi,t-1}(\psi) \\ \varpi_{\psi,t}(\theta) &\equiv 2\frac{1}{\sigma_t(\theta)}\left[1, y_{t-1}, \upsilon_{\psi,t-1}(\psi)'\right]' \text{ and } s_{c,t}(\theta) \equiv \frac{\partial}{\partial c} \ln \sigma_t^2(\theta). \end{aligned}$$

The QML score equations are $m_t(\theta) \equiv [m_{\psi,t}(\theta)', m_{\delta,t}(\theta)']$, where

$$m_{\psi,t}(\theta) \equiv \left(\epsilon_t^2(\theta) - 1\right) s_{\psi,t}(\theta) + \epsilon_t(\theta) \varpi_{\psi,t}(\theta) \text{ and } m_{\delta,t}(\theta) \equiv \left(\epsilon_t^2(\theta) - 1\right) s_{\delta,t}(\theta).$$

The second set $m_{\delta,t}(\theta)$ are just the usual equations for a GARCH model. The first set $m_{\psi,t}(\theta)$ reflects ARMA parameters, in particular the additive term $\epsilon_t(\theta)\varpi_{\psi,t}(\theta)$ imbeds an iterative relationship through the moving average component. In the AR-GARCH case we simply have

$$m_{\psi,t}(\theta) \equiv \left(\epsilon_t^2(\theta) - 1\right) s_{\psi,t}(\theta) + 2\epsilon_t(\theta)\frac{1}{\sigma_t(\theta)}\left[1, y_{t-1}\right]'.$$

It is also useful to look at the Jacobian of $m_t(\theta)$ under the initial assumption that it exists, since we can target trimming in the equations $m_t(\theta)$ to ensure Jacobian robustness. We have

$$\frac{\partial}{\partial \theta} E\left[m_t\right] = -\begin{bmatrix} E\left[s_{\psi,t} s_{\psi,t}'\right] + \frac{1}{2} E\left[\varpi_{\psi,t}\varpi_{\psi,t}'\right], & E\left[s_{\delta,t} s_{\psi,t}'\right] \\ E\left[s_{\psi,t} s_{\delta,t}'\right], & E\left[s_{\delta,t} s_{\delta,t}'\right] \end{bmatrix}.$$

Now look at each component of $m_t(\theta)$ and the Jacobian products $s_{\psi,t}s_{\psi,t}'$, $\varpi_{\psi,t}\varpi_{\psi,t}'$, etc., to see how $m_t(\theta)$ should be trimmed. First, $\epsilon_t$ has a second moment, hence $\epsilon_t(\theta)\varpi_{\psi,t}(\theta)$ does not need to be trimmed by $\epsilon_t(\theta)$. Second, Francq and Zakoïan (2004, eq.'s (4.44), (4.49) and p.629) verify $s_{\delta,t}$ is square integrable, and $|(\partial/\partial\psi_i)\ln\sigma_t^2| \leq K|\upsilon_{i,\psi,t}|$ hence $s_{\psi,t}$ is square integrable only if $u_t$ (and therefore $y_t$) is. Third, by iterating on $(\partial/\partial\psi)u_{t-1}$ it can be shown that $\upsilon_{\psi,t} = -[1, y_{t-1}, u_{t-1}]' - b\upsilon_{\psi,t-1}$ has a second moment only if the ARMA error $u_t$ does (Francq and Zakoïan, 2004, p. 630). Similarly, since $\sigma_t^2 \geq \omega^0 > 0$ *a.s.*, we need only trim the components $y_{t-1}$ and $\upsilon_{\psi,t-1}(\psi)'$ in $\varpi_{\psi,t}(\theta) \equiv 2\sigma_t^{-1}(\theta)[1, y_{t-1}, \upsilon_{\psi,t-1}(\psi)']'$.

In terms of first order robustness, we therefore need only trim $\epsilon_t^2(\theta)$ by $\epsilon_t(\theta)$; trim each element of $s_{\delta,t}$ and $s_{\psi,t}$ by all elements of $s_{\psi,t}$; and iteratively trim each stochastic element of $\upsilon_{\psi,t}(\psi)$ by both $y_{t-1}$ and $u_{t-1}(\psi)$. Define trimming indicators:

$$\hat{I}_{n,t}^{(y)} \equiv I\left(|y_t| \leq y^{(a)}_{(k_n^{(y)})}\right) \text{ and } \hat{I}_{n,t}^{(u)}(\theta) \equiv I\left(|u_t(\theta)| \leq u^{(a)}_{(k_n^{(u)})}(\theta)\right)$$

$$\hat{I}_{n,t}^{(s_\psi)}(\theta) \equiv \prod_{i=1}^{3} I\left(|s_{i,\psi,t}(\theta)| \leq s^{(a)}_{i,\psi,(k_{i,n}^{(s_\psi)})}(\theta)\right)$$

and trimmed variables

$$\hat{\epsilon}_{n,t}^*(\theta) \equiv \epsilon_t(\theta)\hat{I}_{n,t}^{(\epsilon)}(\theta),\ \hat{u}_{n,t}^*(\theta) \equiv u_t(\theta)\hat{I}_{n,t}^{(u)}(\theta)\hat{I}_{n,t}^{(y)}(\theta),\ \hat{y}_{n,t}^* \equiv y_t\hat{I}_{n,t}^{(u)}(\theta)\hat{I}_{n,t}^{(y)}(\theta,\ \hat{s}_{n,\cdot,t}^*(\theta) = s_{\cdot,t}(\theta)\hat{I}_{n,t}^{(s_\psi)}(\theta)$$

$$\hat{\upsilon}_{n,\psi,1}^*(\psi) \equiv -\left[1,0,0\right]' \text{ and } \hat{\upsilon}_{n,\psi,t}^*(\psi) = -\left[1,\hat{y}_{n,t-1}^*,\hat{u}_{n,t-1}^*(\psi)\right]' - b\hat{\upsilon}_{n,\psi,t-1}^*(\psi)$$

$$\hat{\varpi}_{\psi,n,t}^*(\theta) \equiv 2\sigma_t^{-1}(\theta)\left[1,\hat{y}_{n,t-1}^*,\hat{\upsilon}_{n,\psi,t-1}^*(\psi)'\right]'.$$

The trimmed equations are therefore

$$\hat{m}_{n,\psi,t}^*(\theta) \equiv \left(\hat{\epsilon}_{n,t}^{*2}(\theta) - \frac{1}{n}\sum_{t=1}^{n}\hat{\epsilon}_{n,t}^{*2}(\theta)\right)\hat{s}_{n,\psi,t}^*(\theta) + \epsilon_t(\theta)\hat{\varpi}_{\psi,n,t}^*(\theta)$$

$$\hat{m}_{n,\delta,t}(\theta) \equiv \left(\hat{\epsilon}_{n,t}^{*2}(\theta) - \frac{1}{n}\sum_{t=1}^{n}\hat{\epsilon}_{n,t}^{*2}(\theta)\right)\hat{s}_{n,\delta,t}^*(\theta).$$

First order asymptotics reveals the Jacobian satisfies $(\partial/\partial)E[\hat{m}_{n,t}^*(\theta)] = \mathcal{J}_n^*(1 + o(1))$ where $\mathcal{J}_n^*$ $= (\partial/\partial)E[m_{n,t}^*(\theta)](1 + o(1))$. It also reveals $\mathcal{J}_n^*$ has the product $\varpi_t(\theta)\hat{\varpi}_{\psi,n,t}^*(\theta)'$ which contains $y_{t-1}\hat{u}_{n,t-1}^*(\psi)$ and $\hat{y}_{n,t-1}^*u_{t-1}(\psi)$: both are suitably trimmed since each $(\hat{u}_{n,t-1}^*(\psi),\hat{y}_{n,t-1}^*)$ are trimmed with the compound indicator $\hat{I}_{n,t}^{(u)}(\theta)\hat{I}_{n,t}^{(y)}(\theta)$. The approximate Jacobian $\mathcal{J}_n^*$ also has $s_{\psi,t}(\theta)\hat{s}_{n,\delta,t}^*(\theta)$, whichis robust to heavy tails precisely by trimming $s_{\delta,t}(\theta)$ by sample extremes of $s_{\psi,t}(\theta)$. And so on: $\hat{m}_{n,t}^*(\theta) \equiv [\hat{m}_{n,\psi,t}^*(\theta)',\hat{m}_{n,\delta,t}^*(\theta)']'$ is suitably trimmed, in particular $||(\partial/\partial\theta)E[\hat{m}_{n,t}^*]|| < \infty$.

Francq and Zakoïan (2004, Assumption A10, Theorem 3.2) require $y_t$ itself to have a finite fourth moment in order to obtain an asymptotically normal QML estimator of model (B.1). Under trimming, we do not require any additional moment conditions on $y_t$.

# References


Aguilar, M., and J. B. Hill (2015): "Robust score and portmanteau tests of volatility spillover," *Journal of Econometrics*, 184, 37 – 61.

Anatolyev, S. (2005): "GMM, GEL, Serial Correlation, and Asymptotic Bias," *Econometrica*, 73, 983–1002.

Andrews, D. W. K. (1999): "Estimation When a Parameter is on a Boundary," *Econometrica*, 67, 1341–1383.

——— (2001): "Testing When a Parameter Is on the Boundary of the Maintained Hypothesis," *Econometrica*, 69, 683–734.

Antoine, B., H. Bonnal, and E. Renault (2007): "On the Efficient Use of the Informational Content of Estimating Equations: Implied Probabilities and Euclidean Empirical Likelihood," *Journal of Econometrics*, 138, 461–487.

Back, K., and D. P. Brown (1993): "Implied Probabilities in GMM Estimators," *Econometrica*, 61, 971–975.

Basrak, B., R. A. Davis, and T. Mikosch (2002): "Regular Variation of GARCH Processes," *Stochastic Processes and Their Applications*, 99, 95–115.

Berkes, I., and L. Horvath (2004): "The Efficiency of the Estimators of the Parameters in Garch Processes," *The Annals of Statistics*, 32, 633–655.

Bollerslev, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327.

Bonnal, H., and E. Renault (2004): "Minimum Chi-Square Estimation with Conditional Moment Restrictions," *Working Paper, CIREQ, Universite de Montreal*.

Bougerol, P., and N. Picard (1992): "Stationarity of Garch Processes and of Some Nonnegative Time Series," *Journal of Econometrics*, 52, 115–127.

Breiman, L. (1965): "On Some Limit Theorems Similar to the Arc-Sin Law," *Theory of Probability and its Applications*, 10, 323–331.

Brown, B. W., and W. K. Newey (1998): "Efficient Semiparametric Estimation of Expectations," *Econometrica*, 66, 453–464.

Cantoni, E., and E. Ronchetti (2001): "Robust Inference for Generalized Linear Models," *Journal of the American Statistical Society*, 96, 1022–1030.

Carrasco, M., and X. Chen (2002): "Mixing and Moment Properties of Various GARCH and Stochastic Volatility Models," *Econometric Theory*, 18, 17–39.

Chan, N. H., and S. Ling (2006): "Empirical Likelihood for GARCH Models," *Econometric Theory*, 22, 403–428.

Chen, S. (2008): "Nonparametric Estimation of Expected Shortfall," *Journal of Financial Econometrics*, 1, 87–107.

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

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

Doukhan, P., P. Massart, and E. Rio (1995): "Invariance Principles for Absolutely Regular Empirical Processes," *Annales de l'Institut Henri Poincar*, 31, 393–427.

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

Engle, R. F., and J. Mezrich (1996): "GARCH for Groups," *Risk*, 9, 36–40.

Francq, C., and J.-M. Zakoïan (2000): "Estimating Weak Garch Representations," *Econometric Theory*, 16, 692–728.

——— (2004): "Maximum Likelihood Estimation of Pure GARCH and ARMA-GARCH Processes," *Bernoulli*, 10, 605–637.

Garcia, R., E. Renault, and G. Tsafack (2007): "Proper Conditioning for Coherent VaR in Portfolio Management," *Management Science*, 53, 483–494.

Godambe, V. P. (1985): "The Foundation of Finite Sample Estimationin Stochastic Processes," *Biometrika*, 72, 419–428.

Gonzalez-Rivera, G., and F. C. Drost (1999): "Efficiency comparisons of Maximum-Likelihood-Based Estimators in GARCH Models," *Journal of Econometrics*, 93, 93–111.

Guggenberger, P. (2008): "Finite Sample Evidence Suggesting a Heavy Tail Problem of the Generalized Empirical Likelihood Estimator," *Econometric Reviews*, 27, 526–541.

Guggenberger, P., and R. J. Smith (2008): "Generalized Empirical Likelihood Tests in Time Series Models with Potential Identification Failure," *Journal of Econometrics*, 142, 134–161.

Haeusler, E., and J. L. Teugels (1985): "On Asymptotic Normality of Hill's Estimator for the Exponent of Regular Variation," *The Annals of Statistics*, 13, 743–756.

Hall, P. (1990): "Asymptotic Properties of the Bootstrap for Heavy-Tailed Distributions," *Annals of Statistics*, 18, 1342–1360.

Hall, P., and J. L. Horowitz (1996): "Bootstrap Critical Values for Tests Based on Generalized-Method-of-Moments Estimators," *Econometrica*, 64, 891–916.

Hall, P., and Q. Yao (2003): "Inference in ARCH and GARCH Models with Heavy-Tailed Errors," *Econometrica*, 71, 285–317.

Hansen, L. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hansen, L., J. Heaton, and A. Yaron (1996): "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business and Economic Statistics*, 14, 262–280.

Hill, B. M. (1975): "A Simple General Approach to Inference about the Tail of a Distribution," *Annals of Statistics*, 3(5), 1163–1174.

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

——— (2011): "Tail and Non-Tail Memory with Applications to Extreme Value and Robust Statistics," *Econometric Theory*, 27, 844–884.

——— (2012): "Heavy-Tail and Plug-In Robust Consistent Conditional Moment Tests of Functional Form," in *Festschrift in Honor of Hal White*, ed. by X. Chen, and N. Swanson, pp. 241–274. Springer: New York.

——— (2013): "Least Tail-Trimmed Squares for Infinite Variance Autoregressions," *Journal of Time Series Analysis*, 34, 168–186.

——— (2015a): "Robust Estimation and Inference for Heavy Tailed GARCH," *Bernoulli*, 21, 1629–1669.

——— (2015b): "Robust Expected Shorfall Estimation for Infinite Variance Time Series," *Journal of Financial Econometrics*, 13, 1–44.

——— (2015c): "Tail Index Estimation for a Filtered Dependent Time Series," *Statistica Sinica*, 25, 609–630.

Hill, J. B., and M. Aguilar (2013): "Moment Condition Tests for Heavy Tailed Time Series," *Journal of Econometrics*, 172, 255–274.

Hill, J. B., and A. Prokhorov (2014): "Supplemental Material for "GEL Estimation for Heavy-Tailed GARCH Models with Robust Empirical Likelihood Inference"," *Working Paper, Sydney University, Discipline of Business Analytics*.

Hill, J. B., and E. Renault (2010): "Generalized Method of Moments with Tail Trimming," Discussion paper, Dept. of Economics, University of North Carolina - Chapel Hill.

Hill, J. B., and E. Renault (2012): "Variance Targeting for Heavy Tailed Time Series," Discussion paper, Dept. of Economics, University of North Carolina - Chapel Hill.

Ibragimov, R. (2009): "Portfolio Diversification and Value at Risk under Thick-Tailedness," *Quantitative Finance*, 9, 565–580.

Imbens, G. W. (1997): "One-step Estimators for Over-Identified Generalized Method of Moments Models," *The Review of Economic Studies*, 64, 359–383.

Imbens, G. W., R. H. Spady, and P. Johnson (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333–357.

Inoue, A., and M. Shintani (2006): "Bootstrapping GMM Estimators for Time Series," *Journal of Econometrics*, 133, 531–555.

Jensen, S. T., and A. Rahbek (2004a): "Asymptotic Inference for Nonstationary GARCH," *Econometric Theory*, 20, 1203–1226.

——— (2004b): "Asymptotic Normality of the QMLE Estimator of ARCH in the Nonstationary Case," *Econometrica*, 72, 641–646.

Kitamura, Y. (1997): "Empirical Likelihood Methods with Weakly Dependent Processes," *The Annals of Statistics*, 25, 2084–2102.

Kitamura, Y., and M. Stutzer (1997): "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861–874.

Li, D. X., and H. J. Turtle (2000): "Semiparametric ARCH Models: An Estimating Function Approach," *Journal of Business & Economic Statistics*, 18, 174–186.

Linton, O., J. Pan, and H. Wang (2010): "Estimation for a Nonstationary Semi-Strong GARCH(1,1) Model with Heavy-Tailed Errors," *Econometric Theory*, 26, 1–28.

Linton, O., and Z. Xiao (2013): "Estimation of and Inference about the Expected Shortfall for Time Series with Infinite Variance," *Econometric Theory*, 29, 771–807.

Lumsdaine, R. L. (1995): "Finite-Sample Properties of the Maximum Likelihood Estimator in GARCH(1,1) and IGARCH(1,1) Models: A Monte Carlo Investigation," *Journal of Business and Economic Statistics*, 13, 1–10.

Mancini, L., E. Ronchetti, and F. Trojani (2005): "Optimal Conditionally Unbiased Bounded-Influence Inference in Dynamic Location and Scale," *Journal of the American Statistic Association*, 100, 628–641.

Mehra, K. L., and M. S. Rao (1975): "On Functions of Order Statistics for Mixing Processes," *Annals of Statistics*, 3, 874–883.

Meitz, M., and P. Saikkonen (2011): "Parameter Estimation in Nonlinear ARGARCH Models," *Econometric Theory*, 27, 1236–1278.

Nelson, D. B. (1990): "Stationarity and Persistence in the GARCH(1,1) Model," *Econometric Theory*, 6, 318–334.

Newey, W. K., and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood estimators," *Econometrica*, 72, 219–255.

Owen, A. (1990): "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18, 90–120.

——— (1991): "Empirical Likelihood for Linear Models," *The Annals of Statistics*, 19, 1725–1747.

Parente, P., and R. Smith (2011): "GEL Methods for Non-Smooth Moment Indicators," *Econometric Theory*, 27, 47–73.

Peng, L. (2001): "Estimating the Mean of a Heavy Tailed Distribution," *Statistics and Probability Letters*, 52, 255–264.

Peng, L. (2004): "Empirical-Likelihood-Based Confidence Interval for the MEan with a Heavy-Tailed Distribution," *Annals of Statistics*, 32, 1192–1214.

Peng, L., and Q. Yao (2003): "Least Absolute Deviations Estimation for ARCH and GARCH Models," *Biometrika*, 90, 967–975.

Qin, J., and J. Lawless (1994): "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics*, 22, 300–325.

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

Ronchetti, E., and F. Trojani (2001): "Robust Inference with GMM Estimators," *Journal of Econometrics*, 101, 37–69.

Rothenberg, T. J. . (1984): "Approximating the Distributions of Econometric Estimators and Test Statistics," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. D. Intriligator, vol. 2. North Holland, New York.

Sakata, S., and H. White (1998): "High Breakdown Point Conditional Dispersion Estimation with Application to S&P 500 Daily Returns Volatility," *Econometrica*, 66, 529–567.

Scaillet, O. (2004): "Nonparametric Estimation and Sensitivity Analysis of Expected Shortfall," *Mathematical Finance*, 14, 115–129.

Skoglund, J. (2010): "A Simple Efficient GMM Estimator of GARCH Models," *Working Paper: Dept. of Economic Statistics, Stockholm School of Economics*.

Smith, R. J. (1997): "Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation," *The Economic Journal*, 107, 503–519.

Smith, R. J. (2011): "GEL Criteria for Moment Condition Models," *Econometric Theory*, 27, 1192–1235.

Wagner, N., and T. A. Marsh (2005): "Measuring Tail Thickness under GARCH and an Application to Extreme Exchange Rate Changes," *Journal of Empirical Finance*, 12, 165–185.

Worms, J., and R. Worms (2011): "Empirical Likelihood Based Confidence Regions for First Order Parameters of Heavy-Tailed Distributions," *Journal of Statistical Planning and Inference*, 141, 2769–2786.

Zhu, K., and S. Ling (2011): "Global Self-Weighted and Local Quasi-Maximum Exponential Likelihood Estimators for ARMA-GARCH/IGARCH Models," *Annals of Statistics*, 39, 2131–2163.

**TABLE 1:** Base Case[a] : Estimation Results for $\beta^0 = .6$

$\epsilon_t \sim \bar{P}_{2.5}$ and $\kappa_y = 1.5$[e]

| | $n = 100$ | | | | $n = 250$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Bias | RMS[b] | KS[c] | 95% CR[d] | Bias | RMS | KS | 95% CR |
| TT-EL | .0059 | .1696 | 1.332 | .245, .854 | -.0005 | .1483 | 1.186 | .475, .745 |
| TT-CUE | .0022 | .1690 | 1.034 | .189, .881 | .0006 | .1399 | 1.245 | .363, .791 |
| TT-ET | -.0056 | .1758 | 1.145 | .215, .849 | .0012 | .1412 | 1.192 | .399, .761 |
| EL | -.0019 | .1695 | 1.435 | .234, .858 | .0132 | .1374 | 2.464 | .312, .824 |
| CUE | -.0057 | .1789 | 1.277 | .173, .881 | -.0079 | .1555 | 1.867 | .259, .869 |
| ET | -.0030 | .1801 | 1.840 | .206, .857 | .0075 | .1382 | 1.414 | .302, .826 |
| WLQML[f] | .0490 | .3523 | 2.231 | .221, .841 | -.0382 | .2652 | 1.182 | .286, .795 |
| Log-LAD | -.0691 | .3771 | 3.078 | .198, .799 | .0026 | .2587 | 1.454 | .412, .747 |
| QMTTL | .0032 | .2307 | 1.342 | .179, .868 | .0007 | .1676 | 1.082 | .323, .796 |
| QML | -.0462 | .1236 | 3.97 | .212, .846 | -.0324 | .0761 | 2.189 | .212, .847 |

$\epsilon_t \sim N(0,1)$ and $\kappa_y = 4.1$

| | $n = 100$ | | | | $n = 250$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Bias | RMS | KS | 95% CR | Bias | RMS | KS | 95% CR |
| TT-EL | -.0068 | .1096 | .9453 | .340, .783 | .0058 | .0792 | 1.312 | .458, .745 |
| TT-CUE | -.0038 | .1012 | 1.192 | .369, .771 | .0022 | .0803 | .6871 | .421, .769 |
| TT-ET | -.0042 | .1086 | .9532 | .389, .796 | .0035 | .0799 | 1.132 | .453, .750 |
| EL | -.0002 | .1052 | .8061 | .392, .791 | .0071 | .0799 | 1.564 | .456, .748 |
| CUE | -.0012 | .1094 | .6799 | .377, .801 | .0092 | .0802 | 1.267 | .414, .808 |
| ET | -.0024 | .1104 | .9488 | .381, .788 | .0033 | .0823 | 1.512 | .456, .747 |
| WLQML | -.0814 | .3891 | 3.113 | .314, .812 | .0146 | .2596 | 2.102 | .412, .800 |
| Log-LAD | -.0639 | .2987 | 2.573 | .354, .802 | -.0599 | .2354 | 2.865 | .427, .786 |
| QMTTL | -.0325 | .1034 | 1.892 | .400, .846 | -.0123 | .0723 | 1.298 | .435, .765 |
| QML | -.1022 | .1497 | 3.599 | .180, .769 | -.0921 | .1332 | 2.893 | .243, .751 |

a. Base-case trimming fractiles are $k_n^{(\epsilon)} = [.05n/\ln(n)]$ and $k_n^{(y)} = [.2\ln(n)]$.
b. The square root of the empirical mean squared error.
c. The Kolmogorov-Smirnov statistic divided by the 5% critical value: KS $> 1$ indicates rejection of normality at the 5% level.
d. Simulation average 95% confidence region for $\theta_3^0 = .6$ computed by the empirical likelihood method.
e. Tail index of $y_t$ is $\kappa_y$.[17]
f. GEL and GELITT estimators are computed using weights $x_t(\theta) = [s_t'(\theta), s_{t-1}'(\theta)]'$. TT denotes "tail-trimmed", e.g. TT-EL is GELITT with the EL criterion.
f. WLQML is Weighted Laplace QML; QMTTL is Quasi-Maximum Tail-Trimmed Likelihood.

[17]The GARCH process $\{y_t\}$ satisfies $P(|y_t| > a) = da^{-\kappa_y}(1 + o(1))$ and $E|\alpha^0\epsilon_t^2 + \beta^0|^{\kappa_y/2} = 1$. We draw $R = 10,000$ iid $\epsilon_t$ from $P_{2.5}$ or $N(0,1)$ and report $\arg\min_{\kappa\in\mathcal{K}} |1/R\sum_{t=1}^{R} |\alpha^0\epsilon_t^2 + \beta^0|^{\kappa/2} - 1|$ where $\mathcal{K} = \{.001, .002, ..., 10\}$.

**TABLE 2 :** Base Case[a] : t-tests[b] at 5% level for $\beta^0$

$\epsilon_t \sim \bar{P}_{2.5}$ and $\kappa_y = 1.5$

| | $n = 100$ | | | | $n = 250$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $H_0$ | $H_1^1$ | $H_1^2$ | $H_1^3$ | $H_0$ | $H_1^1$ | $H_1^2$ | $H_1^3$ |
| TT-EL[c] | .041[d] | .592 | .869 | .951 | .045 | .818 | .970 | 1.00 |
| TT-CUE | .042 | .568 | .815 | .925 | .042 | .840 | .982 | 1.00 |
| TT-ET | .039 | .617 | .852 | .926 | .053 | .829 | .976 | 1.00 |
| EL | .030 | .638 | .874 | .928 | .038 | .810 | .959 | .927 |
| CUE | .038 | .443 | .704 | .856 | .035 | .775 | .948 | .990 |
| ET | .038 | .609 | .832 | .911 | .051 | .807 | .954 | .980 |
| WLQML[e] | .001 | .004 | .006 | .368 | .002 | .101 | .238 | .486 |
| Log-LAD | .029 | .103 | .275 | .813 | .028 | .870 | 1.00 | 1.00 |
| QMTTL | .043 | .496 | .718 | .817 | .046 | 1.00 | 1.00 | 1.00 |
| QML | .059 | .878 | 1.00 | 1.00 | .093 | 1.00 | 1.00 | 1.00 |

$\epsilon_t \sim N(0,1)$ and $\kappa_y = 4.1$

| | $n = 100$ | | | | $n = 250$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $H_0$ | $H_1^1$ | $H_1^2$ | $H_1^3$ | $H_0$ | $H_1^1$ | $H_1^2$ | $H_1^3$ |
| TT-EL | .047 | .903 | 1.00 | 1.00 | .053 | 1.00 | 1.00 | 1.00 |
| TT-CUE | .047 | .830 | .970 | 1.00 | .047 | 1.00 | 1.00 | 1.00 |
| TT-ET | .048 | .934 | 1.00 | 1.00 | .048 | 1.00 | 1.00 | 1.00 |
| EL | .056 | .944 | 1.00 | 1.00 | .061 | 1.00 | 1.00 | 1.00 |
| CUE | .055 | .923 | 1.00 | 1.00 | .053 | 1.00 | 1.00 | 1.00 |
| ET | .059 | .899 | 1.00 | 1.00 | .046 | 1.00 | 1.00 | 1.00 |
| WLQML | .004 | .086 | .151 | .428 | .008 | .085 | .222 | .579 |
| Log-LAD | .028 | .067 | .118 | .329 | .034 | .410 | .639 | .980 |
| QMTTL | .053 | .980 | 1.00 | 1.00 | .052 | 1.00 | 1.00 | 1.00 |
| QML | .063 | .188 | .449 | .625 | .067 | .320 | .476 | .688 |

a. Base-case trimming fractiles are $k_n^{(\epsilon)} = [.05n/\ln(n)]$ and $k_n^{(y)} = [.2\ln(n)]$.

b. The true $\beta^0 = .6$. The hypotheses are $H_0$: $\beta = .6$; $H_1^1$: $\beta = .5$; $H_1^2$: $\beta = .35$; and $H_1^3$: $\beta = 0$.

c. GEL and GELITT estimators are computed using weights $x_t(\theta) = [s_t'(\theta), s_{t-1}'(\theta)]'$. TT denotes "tail-trimmed", e.g. TT-EL is GELITT with the EL criterion.

d. Rejection frequencies at the 5% level.

e. WLQML is Weighted Laplace QML. QMTTL is Quasi-Maximum Tail-Trimmed Likelihood.

**TABLE 3:** IGARCH etc.[a] : TT-CUE Results for $\beta^0$

| | $n = 100$ | | | | | $n = 250$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\epsilon_t \sim \bar{P}_{2.5}$ and $\kappa_y = 1.5$ | | | | | | | | | |
| $\alpha^0, \beta^0$ | Bias | RMS[b] | KS[c] | 95% CR[d] | | Bias | RMS | KS | 95% CR |
| .30,.60 | -.006 | .178 | 1.17 | .109, .907 | | .002 | .129 | 1.04 | .253, ..890 |
| .40,.60 | .009 | .179 | 1.13 | .093, .905 | | -.001 | .136 | 1.09 | .132, .867 |
| .30,.70 | -.008 | .158 | 1.21 | .092, .943 | | -.006 | .114 | .986 | .355, .912 |
| .45,.60 | .005 | .155 | 1.10 | .107, .910 | | .006 | .138 | 1.05 | .167, .860 |
| .35,.70 | -.009 | .147 | 1.18 | .172, .945 | | -.007 | .118 | 1.04 | .271, .898 |
| $\epsilon_t \sim N(0,1)$ and $\kappa_y = 4.1$ | | | | | | | | | |
| $\alpha^0, \beta^0$ | Bias | RMS | KS | 95% CR | | Bias | RMS | KS | 95% CR |
| .30,.60 | -.001 | .097 | .740 | .343, .841 | | .006 | .088 | .851 | .348, .835 |
| .40,.60 | -.004 | .105 | .985 | .322, .821 | | -.003 | .075 | .720 | .415, .779 |
| .30,.70 | .005 | .098 | .993 | .420, .931 | | .005 | .077 | .994 | .483, .862 |
| .45,.60 | .007 | .098 | 1.12 | .354, .808 | | -.006 | .071 | 1.05 | .398, .778 |
| .35,.70 | .008 | .103 | 1.15 | .370, .863 | | -.006 | .079 | .987 | .451, .820 |

a. GARCH and IGARCH models are considered. Trimming fractiles are $k_n^{(\epsilon)} = [.05n/\ln(n)]$ and $k_n^{(y)} = [3\ln(n)]$.

b. The square root of the empirical mean squared error.

c. The Kolmogorov-Smirnov statistic divided by the 5% critical value: KS $> 1$ indicates rejection of normality at the 5% level.

d. Simulation average 95% confidence region for $\beta^0$ computed by the empirical likelihood method.

**TABLE 4 :** IGARCH etc.[a] : TT-CUE t-tests[b] at 5% level for $\beta^0$

| | $n = 100$ | | | | | $n = 250$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\epsilon_t \sim \bar{P}_{2.5}$ and $\kappa_y = 1.5$ | | | | | | | | | |
| $\alpha^0, \beta^0$ | $H_0$ | $H_1^1$ | $H_1^2$ | $H_1^3$ | | $H_0$ | $H_1^1$ | $H_1^2$ | $H_1^3$ |
| .30,.60 | .044[c] | .521 | .752 | .865 | | .045 | .822 | .954 | 1.00 |
| .40,.60 | .042 | .549 | .787 | .906 | | .045 | .729 | .922 | .993 |
| .30,.70 | .053 | .760 | .931 | .978 | | .045 | .962 | 1.00 | 1.00 |
| .45,.60 | .045 | .649 | .893 | .947 | | .044 | .741 | .940 | 1.00 |
| .35,.70 | .053 | .840 | .952 | .985 | | .047 | .929 | 1.00 | 1.00 |
| $\epsilon_t \sim N(0,1)$ and $\kappa_y = 4.1$ | | | | | | | | | |
| $\alpha^0, \beta^0$ | $H_0$ | $H_1^1$ | $H_1^2$ | $H_1^3$ | | $H_0$ | $H_1^1$ | $H_1^2$ | $H_1^3$ |
| .30,.60 | .041 | .931 | 1.00 | 1.00 | | .043 | 1.00 | 1.00 | 1.00 |
| .40,.60 | .039 | .832 | 1.00 | 1.00 | | .042 | 1.00 | 1.00 | 1.00 |
| .30,.70 | .042 | .958 | 1.00 | 1.00 | | .053 | 1.00 | 1.00 | 1.00 |
| .45,.60 | .042 | .882 | 1.00 | 1.00 | | .043 | 1.00 | 1.00 | 1.00 |
| .35,.70 | .044 | .856 | 1.00 | 1.00 | | .048 | 1.00 | 1.00 | 1.00 |

a. GARCH and IGARCH models are considered. Trimming fractiles are $k_n^{(\epsilon)} = [.05n/\ln(n)]$ and $k_n^{(y)} = [3\ln(n)]$.
b. The hypotheses are $H_0$: $\beta = \beta^0$; $H_1^1$: $\beta = \beta^0 - .1$; $H_1^2$: $\beta = \beta^0 - .25$; and $H_1^3$: $\beta = 0$.
c Rejection frequencies at the 5% level.

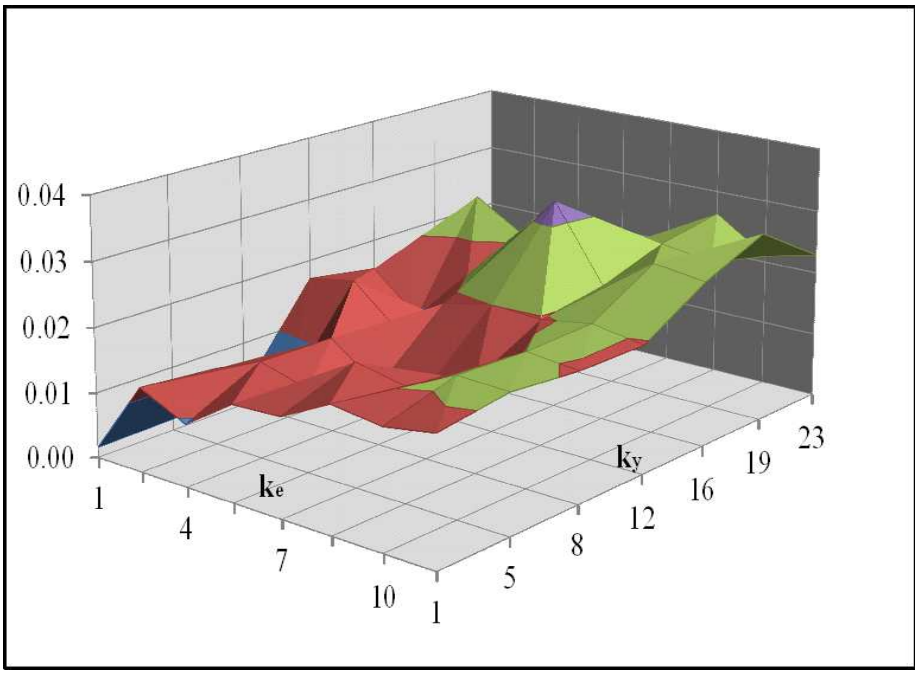



Figure 1: Simulation bias for tail-trimmed CUE. The plot is over grid of trimming fractiles $\{k_\epsilon, k_y\}$. The model is $y_t = \epsilon_t \sigma_t$ and $\sigma_t^2 = 1 + .3y_{t-1}^2 + .6\sigma_{t-1}^2$, where $\epsilon_t$ has power law tails with index $\kappa = 2.5$, and the sample size $n = 100$.

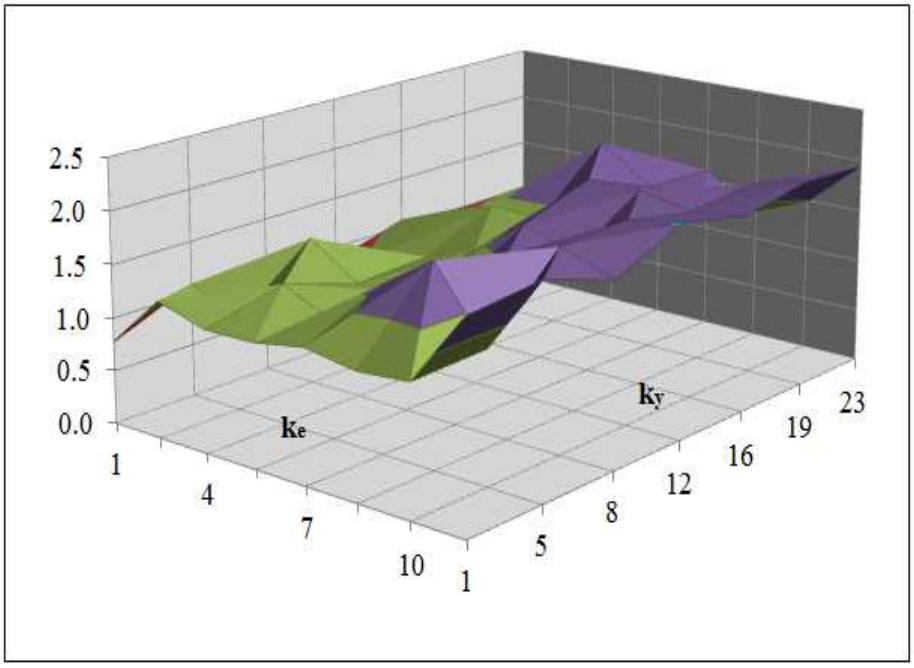
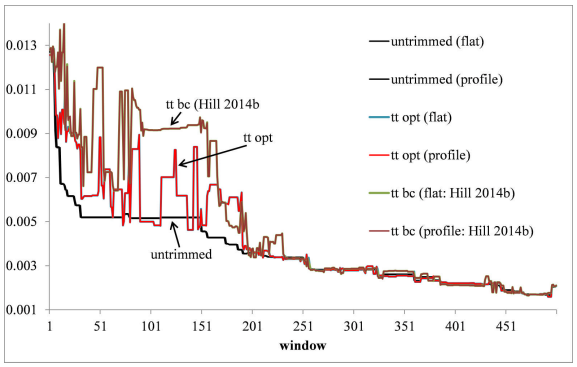
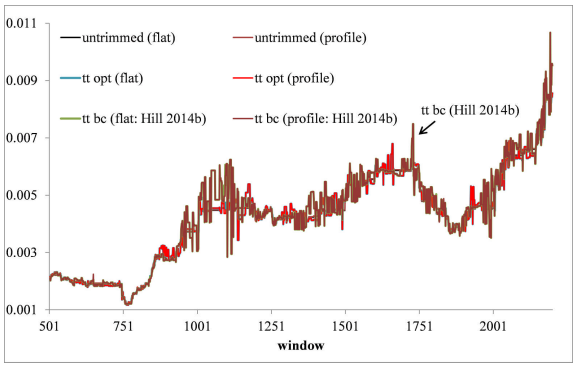



Figure 2: Kolmogorov-Smirnov statistic scaled by its 5% critical value for tail-trimmed CUE. The plot is over grid of trimming fractiles $\{k_\epsilon, k_y\}$. The model is $y_t = \epsilon_t \sigma_t$ and $\sigma_t^2 = 1 + .3y_{t-1}^2$ $+ .6\sigma_{t-1}^2$, where $\epsilon_t$ has power law tails with index $\kappa = 2.5$, and the sample size $n = 100$.

(a) Ruble: Windows 1-500: Years 1999-2000




(b) Ruble: Windows 501-2200: Years 2001-2008




(c) HSI

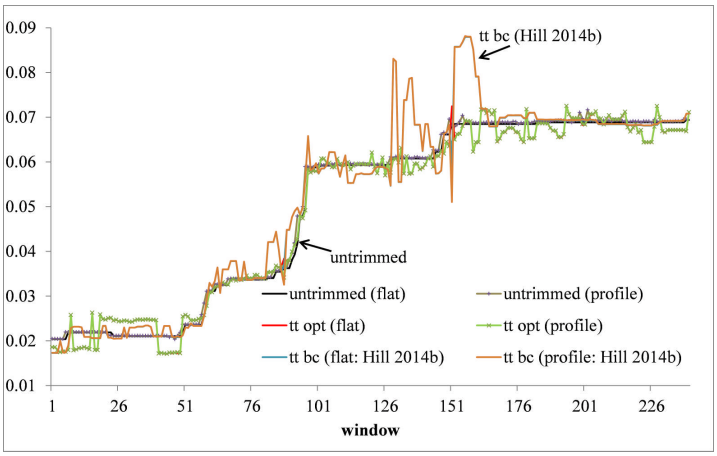
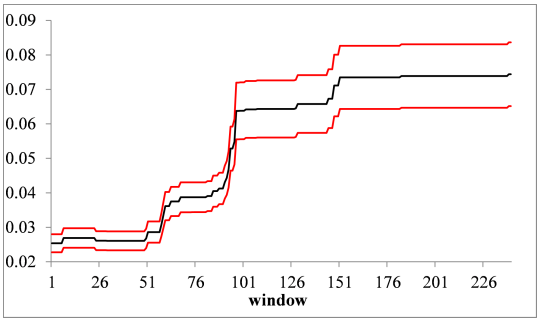
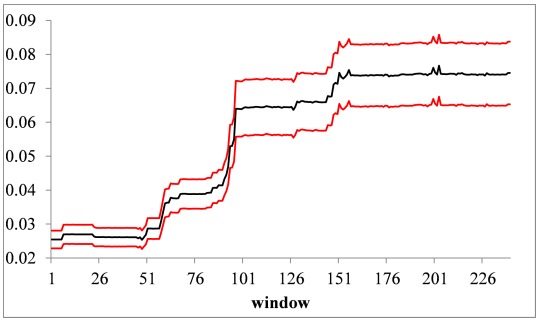



Figure 3: Rolling window expected shortfall: comparison of trimmed and untrimmed estimators. *untrimmed* are untrimmed expected shortfall estimates; *tt opt* are tail-trimmed estimation with optimal bias correction; and *tt bc (Hill 2014b)* are bias corrected tail-trimmed estimates with the fractile $m_n(\phi)$ range used in Hill (2015b). In each case *flat* or *profile* weighting are used. Hill (2015b) only computes *tt bc (flat)*. We break the Ruble rolling windows into two groups to highlight the crisis year 1999 (the initial 248 trading days), the most volatile period in the sample. The Ruble panel (a) is 1999-2000 and panel (b) is 2001-2008. The HSI panel (c) is 1996-1998.

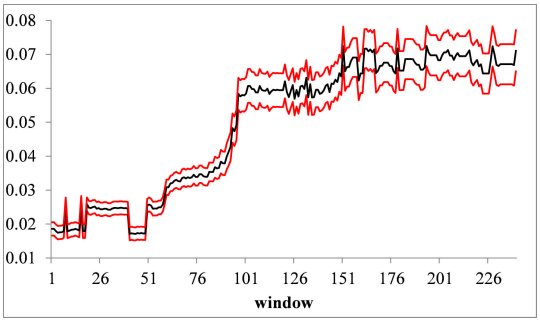
(a) untrimmed, flat weighted




(b) untrimmed, profile weighted




(c) tail-trimmed with optimal bias correction, flat weighted




(d) tail-trimmed with optimal bias correction, profile weighted

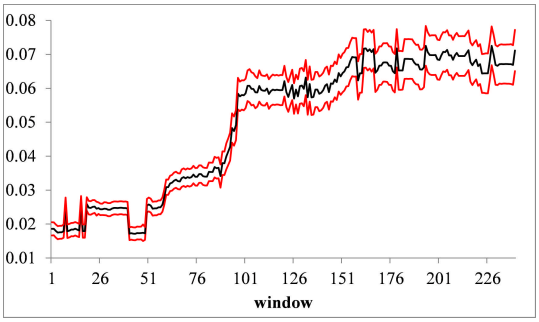
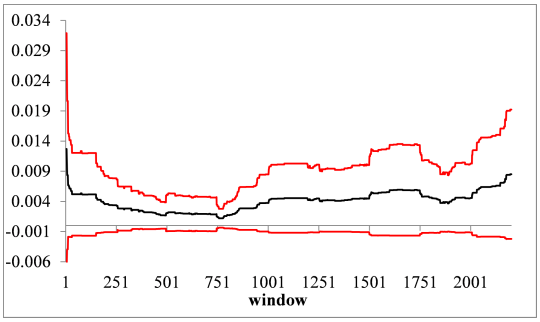
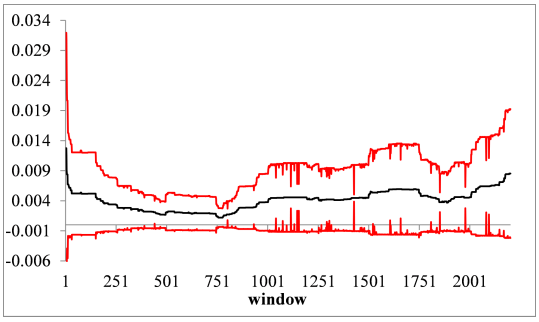



Figure 4: Rolling window expected shortfall estimates for HSI daily returns.

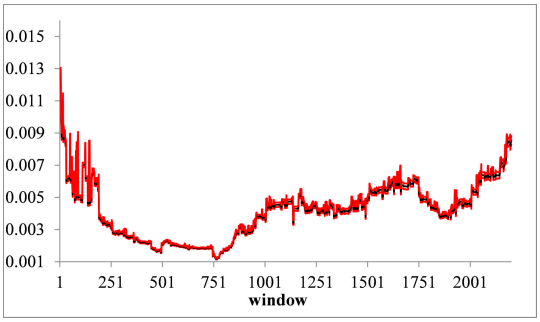
(a) untrimmed, flat weighted




(b) untrimmed, profile weighted




(c) tail-trimmed with optimal bias correction, flat weighted




(d) tail-trimmed with optimal bias correction, profile weighted

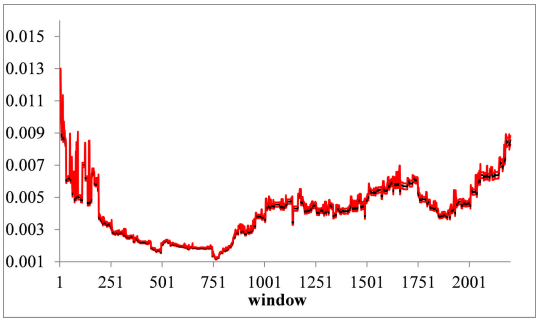



Figure 5: Rolling window expected shortfall estimates for Ruble daily returns.