Extreme Conditional Tail Moment Estimation under Serial Dependence

Yannick Hoga*

May 18, 2018

*Faculty of Economics and Business Administration, University of Duisburg-Essen, Universitätsstraße 12, D-45117 Essen, Germany, tel. +49 201 1834365, yannick.hoga@vwl.uni-due.de. The author wishes to thank the Co-Editor Fabio Trojani and two anonymous referees for their insightful suggestions, which helped to significantly improve the paper. Furthermore, the author is grateful to Christoph Hanck for his detailed comments on an earlier version of this article. Full responsibility is taken for all remaining errors. Support of DFG (HA 6766/2-2) is gratefully acknowledged.

Abstract

A wide range of risk measures can be written as functions of conditional tail moments and Value-at-Risk, for instance the Expected Shortfall. In this paper we derive joint central limit theory for semi-parametric estimates of conditional tail moments, including in particular Expected Shortfall, at arbitrarily small risk levels. We also derive confidence corridors for Value-at-Risk at different levels far out in the tail, which allows for simultaneous inference. We work under a semi-parametric Pareto-type assumption on the distributional tail of the observations and only require an extremal-near epoch dependence assumption on the serial dependence. In simulations, our semi-parametric Expected Shortfall estimate is often shown to be more accurate in terms of mean absolute deviation than extant non- and semi-parametric estimates. An empirical application to the extreme swings in VW log-returns during the failed takeover attempt by Porsche illustrates the proposed methods.

Keywords: Confidence Corridor, Expected Shortfall, E-NED, Pareto-type Tails, Value-at-Risk

JEL classification: C12 (Hypothesis Testing), C13 (Estimation), C14 (Semiparametric and Non-parametric Methods)

1 Motivation

The need to quantify risk, defined broadly, has lead to a burgeoning literature on risk measures. Two of the most popular risk measures in the financial industry are the Value-at-Risk (VaR) at level $p \in (0,1)$, defined as the upper p-quantile of the distribution of losses X, $\text{VaR}_p = F^{\leftarrow}(1-p)$, and the Expected Shortfall (ES) at level p, defined as the expected loss given an exceedance of VaR_p, $\text{ES}_p = \text{E}\left[X \mid X > \text{VaR}_p\right]$. ES is defined if $\text{E}\,|X| < \infty$ and is sometimes also called conditional tail expectation or tail-VaR. In contrast to ES, VaR is not a coherent risk measure in the sense of Artzner *et al.* (1999) and is uninformative as to the expected loss beyond the VaR. Yet, VaR is easy to estimate and to backtest (e.g., Daníelsson, 2011).

A unifying perspective on VaR, ES and a wide range of other popular risk measures was presented by El Methni *et al.* (2014). They introduced the conditional tail moment (CTM), i.e., the a-th moment ($a > 0$) of the loss given a VaR_p-exceedance, $\text{CTM}_a(p) = \text{E}\left[X^a \mid X > \text{VaR}_p\right]$. For $a = 1$, the conditional tail moment is simply the ES. For an appropriate choice of $a < 1$ the conditional tail moment may still be used for extremely heavy-tailed time series with $\text{E}\,|X| = \infty$, when ES can no longer be used. For instance, there is evidence that economic losses in the aftermath of natural disasters have infinite means (Ibragimov *et al.*, 2009; Ibragimov and Walden, 2011). El Methni *et al.* (2014) showed that many risk measures are functions of VaR and CTMs. Hence, by virtue of the continuous mapping theorem, weak limit theory for estimators of these risk measures can be grounded on joint asymptotics of VaR and CTM estimates.

Denote the ordered observations of a time series $X_1, \ldots, X_n$ by $X_{(1)} \geq \ldots \geq X_{(n)}$. While – in the spirit of El Methni *et al.* (2014) – we develop limit theory for many risk measures, we shall frequently focus on our estimator of ES or, equivalently, $\mathrm{CTM}_1(p)$. ES estimation for time series is a topic of recent interest, yet the literature almost exclusively focuses on the case where $\mathrm{E}\,|X_i|^2 < \infty$; see, e.g., Scaillet (2004), Chen (2008). However, evidence for infinite variance models is wide-spread. For instance, IGARCH models have a tail index equal to 2 and hairline infinite variance (Ling, 2007, Thm. 2.1 (iii)). We refer to Engle and Bollerslev (1986) and the references therein for evidence of the plausibility of IGARCH models for exchange rates and interest rates. Infinite variance phenomena can be found more generally in, e.g., insurance and internet traffic applications (Resnick, 2007, Examples 4.1 & 4.2), and emerging market stock returns and exchange rates (Hill, 2013, 2015a).

To the best of our knowledge, only Linton and Xiao (2013) and Hill (2015a) avoid a finite variance assumption for ES estimation of time series. Linton and Xiao (2013) essentially study a simple non-parametric estimate of ES,

$$\widehat{\mathrm{ES}}_p = \frac{1}{pn} \sum_{i=1}^{n} X_i I_{\{X_i \geq X_{(\lfloor pn \rfloor)}\}}, \tag{1}$$

where I_A denotes the indicator function for a set A, and $\lfloor \cdot \rfloor$ rounds to the nearest smallest integer. Linton and Xiao (2013) assume regularly varying tails:

$$P\left\{|X_i| > x\right\} = x^{-1/\gamma} L(x), \qquad \text{where } L(\cdot) \text{ is slowly varying, i.e., } \lim_{y \to \infty} \frac{L(xy)}{L(x)} = 1 \quad \forall\, x > 0. \tag{2}$$

In the case of the Pareto Type I distribution, $L(\cdot)$ is identically a constant, which is why distributions with (2) may be said to be of Pareto-type. Concretely, Linton and Xiao (2013) impose $\gamma \in (1/2, 1)$. Since moments of order greater than or equal to $1/\gamma$ do not exist but smaller ones do (de Haan and Ferreira, 2006, Ex. 1.16), this rules out infinite-mean models by $\gamma < 1$ (in which case ES does not exist anyway) and finite variance models by $\gamma > 1/2$. For geometrically strong-mixing $\{X_i\}$, they derive the stable limit of $n^{1-\gamma}(\widehat{\mathrm{ES}}_p - \mathrm{ES}_p)$, which however depends on the unknown γ. For feasible inference, they consider a subsampling procedure.

Hill (2015a), who also works with geometrically strong-mixing random variables (r.v.s), considers a tail-trimmed estimate

$$\widehat{\mathrm{ES}}_p^{(*)} = \frac{1}{pn} \sum_{i=1}^{n} X_i I_{\{X_{(k_n)} \geq X_i \geq X_{(\lfloor pn \rfloor)}\}}, \tag{3}$$

where the integer trimming sequence $1 \leq k_n < n$ tends to infinity with $k_n = o(n)$. This improves the convergence rate to $\sqrt{n}/g(n)$ for some slowly varying function $g(n) \to \infty$ if $\gamma \in [1/2, 1)$. His results also extend to $\gamma < 1/2$, where he obtains the standard $\sqrt{n}$-rate. In both cases, Hill (2015a) delivers

standard Gaussian limit theory, although – in contrast to Linton and Xiao (2013) – he requires a second-order refinement of (2). To deal with possibly non-vanishing bias terms that may arise due to trimming, Hill (2015a) exploits regular variation and proposes an ES estimator $\widehat{\mathrm{ES}}_p^{(2)} = \widehat{\mathrm{ES}}_p^{(*)} + \widehat{\mathcal{R}}_n^{(2)}$ with some optimal bias correction $\widehat{\mathcal{R}}_n^{(2)}$.

Despite working under a *semi*-parametric Pareto-tail assumption as in (2), Linton and Xiao (2013) and Hill (2015a) (essentially) only consider *non*-parametric estimators of ES, viz., $\widehat{\mathrm{ES}}_p$ and $\widehat{\mathrm{ES}}_p^{(2)}$. Only Hill (2015c) exploits assumption (2) for purposes of bias correction via $\widehat{\mathcal{R}}_n^{(2)}$ in the ES estimate $\widehat{\mathrm{ES}}_p^{(2)}$. In this paper we take a different tack and use (2) as a motivation for a truly *semi*-parametric estimate of ES, and indeed more generally of CTMs. In a regression environment with covariates and independent, identically distributed (i.i.d.) observations, similar estimators have been studied by El Methni *et al.* (2014).

Our first main contribution is to derive the joint weak Gaussian limit of our VaR and CTM estimators under a general notion of dependence, covering and significantly extending the geometrically strong-mixing framework of Linton and Xiao (2013) and Hill (2015a). Thus, not only do we cover estimators of ES (as Linton and Xiao, 2013, and Hill, 2015a, do), but also – among others – those of VaR, conditional tail variance (Valdez, 2005) and conditional tail skewness (Hong and Elshahat, 2010); see El Methni *et al.* (2014). In our extreme value setting, we necessarily require that $p = p_n \to 0$ as $n \to \infty$, thus disadvantaging our estimator in a direct comparison of the convergence rates obtained by Linton and Xiao (2013) and Hill (2015a) for $\widehat{\mathrm{ES}}_p$ and $\widehat{\mathrm{ES}}_p^{(2)}$; see also Remark 6 below. Nonetheless, we obtain a convergence rate that can improve the $n^{1-\gamma}$-rate for $\widehat{\mathrm{ES}}_p$. While the $\sqrt{n}/g(n)$-rate of $\widehat{\mathrm{ES}}_p^{(2)}$ cannot be beaten, we show in simulations that our estimator often has a lower mean absolute deviation (MAD). This is true for a wide range of values $p \in [0.001, 0.05]$, where – quite expectedly, as we focus on $p = p_n \to 0$ – the relative advantage becomes larger, the smaller p.

Our second main contribution is to derive confidence corridors for VaR at different levels. This is important because '[i]n financial risk management, the portfolio manager may be interested in different percentiles [...] of the potential loss and draw some simultaneous inference. This type of information provides the basis for dynamically managing the portfolio to control the overall risk at different levels' (Wang and Zhao, 2016, p. 90). Working with VaR – albeit conditioned on past returns – Wang and Zhao (2016) derive a functional central limit theorem for VaR estimates indexed by the level $p \in [\delta, 1-\delta]$ for some $\delta > 0$. While Wang and Zhao (2016, Rem. 2) conjecture that an extension to the interval $p \in (0, 1)$ may be possible, their current results exclude the tails of the distributions, which are of particular interest in risk management. We fill this gap in the present extreme value setting, where the tail is the natural focus.

For estimators that, like ours, are motivated by extreme value theory (EVT) the choice of the number of upper order statistics (typically denoted by k_n) to use is always tricky; see, e.g, Resnick (2007, Sec. 4.4) and Mancini and Trojani (2011, Sec. 1.4). As a third contribution, we adapt a recently proposed method for choosing k_n by Daníelsson *et al.* (2016) to the particular task of ES estimation. This approach works quite well in simulations in the sense that it yields estimates of $\mathrm{CTM}_1(p_n)$ that are often preferable to other popular competitors in terms of MAD.

We remark that we focus on *unconditional*, instead of *conditional* (upon past observations) risk measures. Both types of measures have their distinct uses. While conditional risk measures (issued from, e.g., GARCH-type models) are calculated on a daily basis by risk managers in banks to adapt to ever evolving market conditions, unconditional risk measures are used for longer-term risk assessments. For instance, unconditional risk measures can be used to judge the severity of historic stock market crashes (Novak and Beirlant, 2006). Gupta and Liang (2005) use (unconditional) VaR estimators from EVT to examine whether hedge funds are adequately capitalized to avoid bankruptcy. To assess the long-term viability of financial institutions, regulators indeed require (unconditional) VaR estimates. Under the Solvency II Directive, insurers are required to report VaR at level 99.5% and as a measure for default risk banks even have to calculate VaR at level 99.9%, as set out by the Basel Committee on Banking Supervision (2016). At these extreme levels, calculation of risk measures naturally calls for EVT-based procedures like the ones investigated in this paper. For more discussion on the relative merits of conditional and unconditional risk measures, we refer to Hoga (2017b, p. 25) and the references therein.

We also remark that the methods for unconditional risk measure estimation – developed in this paper – can be fruitfully used in calculating conditional risk measures as well. For conditional VaR and ES estimation, McNeil and Frey (2000) were the first to propose using (unconditional) EVT-based methods after a preliminary step of filtering out time-varying volatility. Their approach has been found to work well in a comparative study by Kuester *et al.* (2006) and has since been refined by Mancini and Trojani (2011), who robustify both the model estimation stage (required to extract volatility estimates) and the VaR estimation of the residual process with volatility filtered out. Note that in standard GARCH models, asymptotic normality of Gaussian quasi-maximum likelihood estimation (QMLE) requires existing fourth moments of the innovations. Thus, since innovations may possess heavier tails, estimators that are robust to heavy-tailed innovations may be required for model estimation; see, e.g., Hill (2015b). We illustrate how our estimators may be used to calculate conditional risk measures in an empirical application. Deriving the asymptotic properties of these estimators is under active current investigation.

The rest of the paper proceeds as follows. Section 2 states the main theoretical results and is structured as follows. Subsection 2.1 introduces the estimator as well as the dependence concept we work with. The next Subsection 2.2 derives joint central limit theory for CTMs and VaR. In Subsection 2.3 we obtain confidence corridors for VaR at different levels, allowing for simultaneous inference. In the simulations in Section 3, the finite-sample performance of our ES estimator is compared with several non- and semi-parametric competitors in terms of MAD. Section 4 applies the results to the time series of VW log-returns to judge the severity of the losses during the attempted takeover by Porsche, that ultimately failed. The final Section 5 concludes. Proofs are relegated to the Appendix.

2 Main results

2.1 Preliminaries

Let $\{X_i\}$ be a strictly stationary sequence of non-negative r.v.s. As is customary in extreme value theory, we study the right tail. In practice, non-negativity may be achieved via a simple transformation, e.g., $X_i I_{\{X_i \geq 0\}}$ or $-X_i I_{\{-X_i \geq 0\}}$ if interest centers on the right or left tail, respectively. Define the survivor function $\overline{F}(\cdot) = 1 - F(\cdot)$, where F denotes the distribution function of X_1. We assume regularly varying tails $\overline{F}(\cdot) \in RV_{-1/\gamma}$, i.e.,

$$\lim_{x \to \infty} \frac{\overline{F}(\lambda x)}{\overline{F}(x)} = \lambda^{-1/\gamma} \qquad \forall\ \lambda > 0, \tag{4}$$

where $\gamma > 0$ is called the *extreme value index* and $\alpha = 1/\gamma$ the *tail index*. Note that (4) is equivalent to

$$\overline{F}(x) = x^{-1/\gamma} L(x), \qquad \text{where } L(\cdot) \text{ is slowly varying.} \tag{5}$$

This in turn is equivalent to (de Haan and Ferreira, 2006, p. 25)

$$U(x) = x^{\gamma} L_U(x), \qquad \text{where } U(x) = F^{\leftarrow}(1 - 1/x) \quad \text{and} \quad L_U(\cdot) \text{ is slowly varying.} \tag{6}$$

Since (4) is an asymptotic relation, we require an *intermediate sequence* $k_n \to \infty$ with $k_n = o(n)$ and $1 \leq k_n < n$ for statistical purposes. This sequence k_n determines the number of upper order statistics used for estimating γ and is restricted by the following assumption.

Assumption 1. *There exists a function $A(\cdot)$ with $\lim_{x\to\infty} A(x) = 0$ such that for some $\rho < 0$*

$$\lim_{x \to \infty} \frac{\frac{\overline{F}(\lambda x)}{\overline{F}(x)} - \lambda^{-1/\gamma}}{A(x)} = \lambda^{-1/\gamma} \frac{\lambda^{\rho/\gamma} - 1}{\gamma \rho} \qquad \forall\ \lambda > 0. \tag{7}$$

Additionally, $\sqrt{k_n}A\left(U(n/k_n)\right) \longrightarrow 0$*, as* $n \to \infty$*.*

Remark 1. Assumption 1 controls the speed of convergence in (4) and is consequently referred to as a second-order condition in EVT. Equivalently, it may also be written in terms of the quantile function $U(\cdot)$ from (6) (see de Haan and Ferreira, 2006, Thm. 2.3.9). In this form, it is widely-used in tail index (e.g., Einmahl *et al.*, 2016; Hoga, 2017a) and extreme quantile estimation (e.g., Chan *et al.*, 2007; Hoga, 2017b). Examples of d.f.s satisfying Assumption 1 are abundant. For instance, d.f.s expanding as

$$\overline{F}(x) = c_1 x^{-1/\gamma} + c_2 x^{-1/\gamma+\rho/\gamma}\left(1+o(1)\right), \qquad x \to \infty, \qquad (c_1 > 0,\ c_2 \neq 0,\ \gamma > 0,\ \rho < 0) \tag{8}$$

fulfill Assumption 1 with the indicated γ and ρ, and $k_n = o(n^{-2\rho/(1-2\rho)})$ (de Haan and Ferreira, 2006, pp. 76–77). The more negative ρ, the closer the tail is to actual Pareto decay ($\rho = -\infty$). In the Pareto case, $k_n = o(n)$ can be chosen quite large, which is desirable because more observations can be used in estimation; cf. Remark 6. The expansion in (8) is satisfied by, e.g., the Student $t(\nu)$-distribution with $\gamma = 1/\nu$ and $\rho = -2$, where $\nu > 0$ denotes the degrees of freedom.

Define $x_p = F^{\leftarrow}(1-p)$ as the $(1-p)$-quantile for short. Most of the literature, including Linton and Xiao (2013) and Hill (2015a), focuses on the case where $p \in (0,1)$ is fixed. EVT however allows for $p = p_n \to 0$ as $n \to \infty$. Approximations derived from EVT often provide better approximations when p is small – the case of particular interest in risk management –, as they take the semi-parametric tail (4) into account. The following two motivations show how regular variation of the tail is taken into account.

First, we use the regular-variation assumption (4) to estimate x_{p_n} in $\mathrm{CTM}_a(p_n) = \mathrm{E}\left[X^a \mid X > x_{p_n}\right]$ as follows. Note that p_n can be very small, such that x_{p_n} may lie outside the range of observations $X_1, \ldots, X_n$. Then, the idea is to base estimation of x_{p_n} on a less extreme (in-sample) quantile $x_{k_n/n}$ and use (4) to extrapolate from that estimate. Concretely, set $x = x_{k_n/n}$, $\lambda = x_{p_n}/x_{k_n/n}$ and use (4) as an approximation to obtain

$$\left(\frac{x_{p_n}}{x_{k_n/n}}\right)^{-1/\gamma} \approx \frac{1-F(x_{p_n})}{1-F(x_{k_n/n})} \approx \frac{np_n}{k_n}. \tag{9}$$

Replacing population quantities (γ and $x_{k_n/n}$) with empirical quantities ($\widehat{\gamma}$ and $X_{(k_n+1)}$), this motivates the so-called Weissman (1978) estimator $\widehat{x}_{p_n} = d_n^{\widehat{\gamma}} X_{(k_n+1)}$, where $d_n = k_n/(np_n)$. It has been used in, e.g., Drees (2003), Chan *et al.* (2007), and Hoga and Wied (2017). Of course, there is a wide range of

estimators $\widehat{\gamma}$. We use the Hill (1975) estimator

$$\widehat{\gamma} = \widehat{\gamma}_{k_n} = \frac{1}{k_n} \sum_{i=1}^{k_n} \log \left(X_{(i)} / X_{(k_n+1)} \right)$$

in the following, which is arguably the most popular one (see, e.g., Hsing, 1991; Hill, 2010, and the references therein).

For the second approximation we exploit (4) once again. Together with Theorem 4.1 of Pan *et al.* (2013), which was obtained from Karamata's theorem, this assumption implies $\mathrm{CTM}_a(p_n) \sim \frac{x^a_{p_n}}{1-a\gamma}$ as $n \to \infty$. Asymptotic equivalence, $a_n \sim b_n$, is defined as $\lim_{n\to\infty} a_n/b_n = 1$. Thus, the following estimate suggests itself:

$$\widehat{\mathrm{CTM}}_a(p_n) := \frac{\widehat{x}^a_{p_n}}{1 - a\widehat{\gamma}}. \tag{10}$$

This estimator accounts for the regular variation both in estimating x_{p_n} (through (9)) as well as in calculating the expected loss above x_{p_n} (through $\mathrm{CTM}_a(p_n) \sim \frac{x^a_{p_n}}{1-a\gamma}$).

Next, we introduce a sufficiently general dependence concept. The asymptotic behavior of $\widehat{\mathrm{CTM}}_a(p_n)$ crucially relies on that of $\widehat{\gamma}$ (see the proof of Theorem 1). To the best of our knowledge, the most general conditions under which extreme value index estimators have been studied are those in Hill (2010). He develops central limit theory for the Hill (1975) estimator under L_2-extremal-near epoch dependence (L_2-E-NED). Similar to the mixing conditions of Hsing (1991), dependence is restricted only in the extremes. However, the NED property is often more easily verified (e.g., for ARMA–GARCH models) and offers more generality, whereas mixing conditions are typically harder to verify and some simple time series models fail to be mixing Andrews (1984).

For the following introduction to E-NED, it will be illustrative to keep an ARMA(p,q)–GARCH$(\overline{p},\overline{q})$ model $\{X_i\}$ in mind. It is generated by the ARMA(p,q) structure

$$X_i = \mu + \sum_{t=1}^{p} \phi_t X_{i-t} + \sum_{t=1}^{q} \theta_t \epsilon_{i-t} + \epsilon_i, \tag{11}$$

which is driven by a GARCH$(\overline{p},\overline{q})$ process $\{\epsilon_i\}$, i.e.,

$$\epsilon_i = \sigma_i U_i, \quad \text{where} \quad \sigma_i^2 = \omega + \sum_{t=1}^{\overline{p}} \alpha_t \epsilon_{i-t}^2 + \sum_{t=1}^{\overline{q}} \beta_t \sigma_{i-t}^2. \tag{12}$$

In the following, dependence is restricted separately in the errors $\{\epsilon_i\}$ and the actual observed process $\{X_i\}$.

Consider a process $\{\epsilon_i\}$ (the GARCH process in the above example) and a possibly vector-valued functional of it, $\left\{E_{n,i}\right\}_{n\in\mathbb{N};i=1,\ldots,n}$. The array nature of $E_{n,i}$ allows for tail functionals, such

as $E_{n,i} = I_{\{\epsilon_i > a_{n,i}\}}$ for some triangular array $a_{n,i} \to \infty$ as $n \to \infty$. The $E_{n,i}$ induce σ-fields $\mathcal{F}_{n,s}^t = \sigma\left(E_{n,i} \,:\, s \le i \le t\right)$ (where $E_{n,i} = 0$ for $i \notin \{1,\dots,n\}$), which can be used to restrict dependence in $\{\epsilon_i\}$ using the mixing coefficients

$$\varepsilon_{n,q_n} := \sup_{A \in \mathcal{F}_{n,-\infty}^i,\ B \in \mathcal{F}_{n,i+q_n}^\infty :\ i \in \mathbb{Z}} \left|P(A \cap B) - P(A)P(B)\right|,$$
$$\omega_{n,q_n} := \sup_{A \in \mathcal{F}_{n,-\infty}^i,\ B \in \mathcal{F}_{n,i+q_n}^\infty :\ i \in \mathbb{Z}} \left|P(B|A) - P(B)\right|.$$

Here, $\{q_n\} \subset \mathbb{N}$ is a sequence of integer displacements with $1 \le q_n < n$ and $q_n \to \infty$. We then say that $\{\epsilon_i\}$ is *F-strong (uniform) mixing with size* $\lambda > 0$ if

$$(n/k_n) q_n^\lambda \varepsilon_{n,q_n} \underset{(n\to\infty)}{\longrightarrow} 0 \qquad \left((n/k_n) q_n^\lambda \omega_{n,q_n} \underset{(n\to\infty)}{\longrightarrow} 0 \right).$$

Given $\{\epsilon_i\}$ thus restricted, it remains to restrict dependence in the observed series $\{X_i\}$ (the ARMA–GARCH process in the above example). Hill (2010) shows that the asymptotics of the Hill (1975) estimator can be grounded on tail arrays $\left\{I_{\{X_i > b_n e^u\}}\right\}$, where $b_n = U(1 - k_n/n)$. Hence, dependence in $\{X_i\}$ only needs to be restricted via $\left\{I_{\{X_i > b_n e^u\}}\right\}$. This is achieved by assuming that, for some $p > 0$, $\{X_i\}$ is *L_p-E-NED on* $\left\{\mathcal{F}_{n,1}^i\right\}$ *with size* $\lambda > 0$, i.e.,

$$\left\| I_{\{X_i > b_n e^u\}} - P\left\{ X_i > b_n e^u \;\middle|\; \mathcal{F}_{n,i-q_n}^{i+q_n} \right\} \right\|_p \le f_{n,i}(u) \cdot \psi_{q_n},$$

where $f_{n,i} : [0,\infty) \to [0,\infty)$ is Lebesgue measurable, $\sup_{i=1,\dots,n} \sup_{u \ge 0} f_{n,i}(u) = \mathcal{O}\left((k_n/n)^{1/p}\right)$, and $\psi_{q_n} = o(q_n^{-\lambda})$. For more on this dependence concept, we refer to Hill (2009, 2010, 2011).

Assumption 2. *$\{X_i\}$ is L_2-E-NED on $\left\{\mathcal{F}_{n,1}^i\right\}$ with size $\lambda = 1/2$. The constants $f_{n,i}(u)$ are integrable on $[0,\infty)$ with $\sup_{i=1,\dots,n} \int_0^\infty f_{n,i}(u)\mathrm{d}u = \mathcal{O}(\sqrt{k_n/n})$. The base $\{\epsilon_i\}$ is either F-uniform mixing with size $r/[2(r-1)]$, $r \ge 2$, or F-strong mixing with size $r/(r-2)$, $r > 2$.*

The final assumption we require is

Assumption 3. *The covariance matrix of*

$$\begin{pmatrix} \frac{1}{\sqrt{k_n}} \sum_{i=1}^n \left[\log(X_i/b_n)_+ - \mathrm{E}\log(X_i/b_n)_+\right] \\ \frac{1}{\sqrt{k_n}} \sum_{i=1}^n \left[I_{\left\{X_i > b_n e^{u/\sqrt{k_n}}\right\}} - P\left\{ X_i > b_n e^{u/\sqrt{k_n}} \right\} \right] \end{pmatrix}$$

is positive definite uniformly in $n \in \mathbb{N}$ for all $u \in \mathbb{R}$. Here, $x_+ := \max(x, 0)$.

Assumptions 2 and 3 are identical to Assumptions A.2 and D in Hill (2010), whereas Assumption 1 is stronger than the corresponding Assumption B in Hill (2010). Assumption 3 is used to show consistency of estimates of the asymptotic variance of the Hill (1975) estimator in Hill (2010, Thm. 3).

This estimator, $\widehat{\sigma}^2_{k_n}$, appears in Theorem 1, because the asymptotics of $\widehat{x}_{p_n}$ are grounded on those of $\widehat{\gamma}$; see the proof of Theorem 1. The strengthening of Assumption B of Hill (2010) in Assumption 1 is required to derive limit theory for $\widehat{x}_{p_n}$, similarly as in the proof of Thm. 4.3.9 in de Haan and Ferreira (2006).

2.2 Limit theory for extreme conditional tail moments

Theorem 1. *Let $a_1, \ldots, a_J$ be positive and $a_{J+1} = 1$. Assume that*

$$np_n = o(k_n) \qquad \text{and} \qquad \log(np_n) = o(\sqrt{k_n}). \tag{13}$$

Suppose that Assumption 1 is met for $0 < \gamma < \max\{a_1, \ldots, a_{J+1}\}$. Suppose further that Assumptions 2 and 3 are met. Then,

$$\frac{1}{\widehat{\sigma}_{k_n}} \frac{\sqrt{k_n}}{\log d_n} \left[\left(\frac{\widehat{\mathrm{CTM}}_{a_j}(p_n)}{\mathrm{CTM}_{a_j}(p_n)} - 1 \right)_{j=1,\ldots,J}, \left(\frac{\widehat{x}_{p_n}}{x_{p_n}} - 1 \right) \right]' \tag{14}$$

converges in distribution to a zero-mean Gaussian limit with covariance matrix $\boldsymbol{\Sigma} = (a_i a_j)_{i,j \in \{1,\ldots,J+1\}}$. In (14)*,*

$$\widehat{\sigma}^2_{k_n} := \frac{1}{k_n} \sum_{i,j=1}^{n} w\left(\frac{s-t}{\gamma_n}\right) \left[\log\left(\max\left\{ \frac{X_i}{X_{(k_n+1)}}, 1 \right\} \right) - \frac{k_n}{n}\widehat{\gamma} \right] \left[\log\left(\max\left\{ \frac{X_j}{X_{(k_n+1)}}, 1 \right\} \right) - \frac{k_n}{n}\widehat{\gamma} \right]$$

is a kernel-variance estimator with Bartlett kernel $w(\cdot)$, bandwidth $\gamma_n \to \infty$ with $\gamma_n = o(n)$, and $k_n/\sqrt{n} \to \infty$.

Remark 2. Condition (13) restricts the decay of $p_n \to 0$. There, $p_n = o(k_n/n)$ describes the upper bound, required for the EVT approach to make sense, whereas $\log\big((n/k_n)p_n\big) = o\big(\log(np_n)\big) = o(\sqrt{k_n})$ prohibits p_n from decaying to zero too fast and thus describes the boundary, where extrapolation becomes infeasible.

Remark 3. The estimator $\widehat{\sigma}^2_{k_n}$ is due to Hill (2010, Sec. 4). Other possible choices for the kernel $w(\cdot)$ include the Parzen, quadratic spectral and Tukey-Hanning kernel.

Remark 4. It is interesting to contrast Theorem 1 with the fixed-p result in Linton and Xiao (2013). There, replacing the estimate $X_{(\lfloor pn \rfloor)}$ with the true quantile x_p in (1) does not change the limit of $n^{1-\gamma}(\widehat{\mathrm{ES}}_p - \mathrm{ES}_p)$ and the joint distribution of the VaR and the ES estimate is asymptotically independent (Linton and Xiao, 2013, pp. 778–779). In our case where $p = p_n \to 0$, the ES estimate *is* essentially the VaR estimate by (10) and the limit distributions of both estimates are perfectly linearly dependent by (A.3) in the Appendix.

Remark 5. The result of Theorem 1 is sufficient to deliver weak limit theory not only for estimates of VaR and ES, but also for a wide range of other risk measures, e.g., the conditional tail variance, conditional tail skewness, conditional VaR. For terminology and more detail, we refer to El Methni *et al.* (2014).

Remark 6. It may be instructive to compare the rate of convergence in Theorem 1 for our ES estimator $\widehat{\mathrm{CTM}}_1(p_n)$ with the rates of $\widehat{\mathrm{ES}}_p$ and $\widehat{\mathrm{ES}}_p^{(2)}$. As pointed out in Remark 4, for $\gamma \in (1/2, 1)$ Linton and Xiao (2013) obtained a rate of $n^{1-\gamma}$ for $\widehat{\mathrm{ES}}_p$. Up to slowly varying terms, Hill (2015a) improves this rate to $\sqrt{n}$ for $\widehat{\mathrm{ES}}_p^{(2)}$ and general $\gamma < 1$. Recalling from remarks above equation (10) that $\mathrm{CTM}_1(p_n) \sim U(1/p_n)/(1-\gamma)$, Theorem 1 implies

$$\frac{1-\gamma}{\widehat{\sigma}_{k_n}} \frac{\sqrt{k_n}}{(\log d_n) U(1/p_n)} \left(\widehat{\mathrm{CTM}}_1(p_n) - \mathrm{CTM}_1(p_n) \right) \underset{(n\to\infty)}{\xrightarrow{\mathcal{D}}} N(0,1). \tag{15}$$

In order to specify the rate in (15) more precisely, we assume (8). To maximize the rate, we choose k_n as large as allowed by Assumption 1, i.e., $k_n = n^{1-\delta}/g(n) = o(n^{-2\rho/(1-2\rho)})$ for $\delta = 1/(1-2\rho)$; see Remark 1. Here, $g(\cdot)$ denotes an arbitrary slowly varying function with $g(n) \underset{(n\to\infty)}{\longrightarrow} \infty$ as slowly as desired; e.g., $g(n) = \log n$ or $g(n) = \log(\log n)$. Since $p_n \to 0$ in our framework such that $U(1/p_n) \to \infty$ in the denominator of the left-hand side of (15), $\mathrm{CTM}_1(p_n)$ is at a disadvantage compared with $\widehat{\mathrm{ES}}_p$ and $\widehat{\mathrm{ES}}_p^{(2)}$, where $p \in (0,1)$ is fixed. So to make the comparison fairer, we choose the slowest possible rate for p_n allowed by $np_n = o(k_n)$ from (13). Concretely, we set $p_n = k_n/(n \cdot g(n)) = 1/(n^\delta g^2(n))$. So the rate in (15) is given by

$$\frac{\sqrt{k_n}}{(\log d_n) U(1/p_n)} \overset{(6)}{=} \sqrt{n} \frac{n^{-\delta/2} p_n^\gamma}{\sqrt{g(n)} (\log k_n/(np_n)) L_U(1/p_n)} = \sqrt{n} \frac{n^{-\delta(1/2+\gamma)}}{\sqrt{g(n)} (\log g(n)) L_U(1/p_n) g(n)^{2\gamma}}.$$

Hence, up to terms of slow variation, the rate is given by $n^{-\delta(1/2+\gamma)+1/2}$.

Two intuitive observations can be made. First, the larger γ (i.e., the heavier the tail), the slower the rate of convergence. This is to be expected, because the Hill estimate $\widehat{\gamma}$ (upon which our asymptotic results rest) has larger variance for larger γ – everything else being equal. For instance, for the $t(\nu)$-distribution with $\gamma = 1/\nu$ and $\rho = -2$, one may choose $k_n = o(n^{-2\rho/(1-2\rho)}) = o(n^{4/5})$ irrespective of the degrees of freedom ν; see again Remark 1. Then, for i.i.d. observations with d.f.s satisfying Assumption 1, de Haan and Ferreira (2006, Thm. 3.2.5) implies $\sqrt{k_n}(\widehat{\gamma}_{k_n} - \gamma) \xrightarrow{\mathcal{D}} N(0, \gamma^2)$, as $n \to \infty$, which shows that the asymptotic variance is larger the heavier the tail, i.e., the larger γ.

Second, the more negative ρ, the smaller $\delta = 1/(1-2\rho) > 0$ and hence the faster the rate. This result is also expected, since a more negative ρ implies a better fit to true Pareto behavior and hence more upper order statistics can be used for tail estimation by Remark 1. So the more closely the actual tail resembles the Pareto shape, the better the estimators derived from Theorem 1 can be expected

to work relative to non-parametric estimates.

So under the caveat that $\widehat{\mathrm{CTM}}_1(p_n)$ is at a disadvantage, a direct comparison of the convergence rates reveals the following. While the $\sqrt{n}$-rate (up to terms of slow variation) of $\widehat{\mathrm{ES}}_p^{(2)}$ cannot be obtained, the $n^{1-\gamma}$-rate of $\widehat{\mathrm{ES}}_p$ can be improved upon. For instance, for the t_ν-distribution (where, again, $\gamma = 1/\nu$ and $\rho = -2$) we obtain a rate – up to terms of slow variation – of $n^{-\delta(1/2+\gamma)+1/2} = n^{1/5(2-\gamma)}$, which is faster (slower) than $n^{1-\delta}$ for $\gamma > 3/4$ ($\gamma < 3/4$), i.e., for heavier (lighter) tails.

Remark 7. The above observation that ES estimation with $\widehat{\mathrm{CTM}}_1(p_n)$ is more difficult the heavier the tail, also holds for the non-parametric estimate $\widehat{\mathrm{ES}}_p$. This not only holds for the infinite variance case, where the $n^{1-\gamma}$-rate decays the heavier the tail (i.e., the larger γ), but also for the finite variance case with a $\sqrt{n}$-rate; cf. Table 1. We refer to Yamai and Yoshiba (2002) for some supporting simulation evidence and some nice intuition. We also refer to Csörgő *et al.* (1991) for some necessary and sufficient conditions for a central limit theorem to even hold for $\widehat{\mathrm{ES}}_p$ under an i.i.d. assumption. We are not aware of similar results for our ES estimator $\widehat{\mathrm{CTM}}_1(p_n)$. However, there exist well-known necessary and sufficient conditions for asymptotic normality of the Hill (1975) estimator, upon which our asymptotic results are based; see Geluk *et al.* (1997) and the references therein.

2.3 Simultaneous inference on VaR

Working with VaR conditioned on past returns, Wang and Zhao (2016) and Francq and Zakoïan (2016) argue that it is desirable in risk management to be able to draw simultaneous inference on VaR at multiple risk levels. Theorem 2 below shows that in our (unconditional) extreme value context this is particularly easy. Heuristically, if the assumptions of Theorem 1 are met for some sequence $p_n \to 0$, then this also holds for the sequence $p_n(t) := p_n t$ for $t \in \left[\underline{t}, \overline{t}\right]$ $(0 < \underline{t} < \overline{t} < \infty)$, which suggests that $\widehat{x}_{p_n}(t) := X_{(k_n+1)}[k/(np_n(t))]^{\widehat{\gamma}}$ and $\widehat{x}_{p_n}$ should behave very similarly. Note that $\widehat{x}_{p_n} = \widehat{x}_{p_n}(1)$.

Theorem 2. *Under the conditions of Theorem 1 we have that, for* $0 < \underline{t} < \overline{t} < \infty$,

$$\sup_{t\in[\underline{t},\overline{t}]} \left| \frac{1}{\widehat{\sigma}_{k_n}} \frac{\sqrt{k_n}}{\log d_n(t)} \log\left(\frac{\widehat{x}_{p_n}(t)}{x_{p_n}(t)}\right) \right| \underset{(n\to\infty)}{\mathcal{D}} |Z|,$$

where $Z \sim \mathcal{N}(0,1)$, $d_n(t) = k_n/(np_n(t))$ *and* $x_{p_n}(t) = F^{\leftarrow}(1 - p_n(t))$.

The uniform convergence in $t \in [\underline{t}, \overline{t}]$ of Theorem 2 suggests the following $(1-\beta)$-confidence corridor for VaR with levels between $p_n(\underline{t})$ and $p_n(\overline{t})$:

$$\widehat{x}_{p_n}(t) \exp\left\{ -\Phi\left(1 - \frac{\beta}{2}\right) \frac{\log(d_n(t))}{\sqrt{k_n}} \right\} \leq x_{p_n}(t) \leq \widehat{x}_{p_n}(t) \exp\left\{ \Phi\left(1 - \frac{\beta}{2}\right) \frac{\log(d_n(t))}{\sqrt{k_n}} \right\}. \tag{16}$$

It is surprising that the width of the confidence corridor for $x_{p_n}(t)$ does not depend on the values of $\underline{t}$ and $\bar{t}$. Indeed, the confidence corridor is simply obtained by calculating pointwise confidence intervals for $\widehat{x}_{p_n}(t)$. This can be explained by the Pareto-approximation that pins down the tail behavior very precisely by extrapolation. Clearly, in finite samples one may not choose $\bar{t}$ too large, because then the quality of the Pareto-approximation will suffer, rendering confidence corridors (16) imprecise. Also, in actual applications one may not choose $\underline{t}$ too small, as this would push the boundaries of extrapolation too far. So in practice a judicious choice of $\underline{t}$ and $\bar{t}$ (and p_n) is required. In an application in Section 4, some guidance on this issue is given. A similar, yet non-uniform, version of Theorem 2 is given under a more restrictive β-mixing condition by Drees (2003, Thm. 2.2).

Remark 8. Gomes and Pestana (2007, Sec. 3.4) find in simulations that the finite-sample distribution of $\log\left(\widehat{x}_{p_n}/x_{p_n}\right)$ is in better agreement with the asymptotic distribution than $(\widehat{x}_{p_n}/x_{p_n}-1)$. This may be due to $\log(\widehat{x}_{p_n}) = \widehat{\gamma}\log(d_n) + \log(X_{(k_n+1)})$ being a linear function of $\widehat{\gamma}$, upon which the asymptotic results rest; see the proof of Theorem 1.

Remark 9. A close inspection of the proofs of Theorems 1 and 2 reveals that the methodology of this section may also be applied to conditional tail moments. For instance, for our ES estimator we obtain

$$\sup_{t\in[\underline{t},\bar{t}]}\left|\frac{1}{\widehat{\sigma}_{k_n}}\frac{\sqrt{k_n}}{\log d_n(t)}\log\left(\frac{\widehat{\mathrm{CTM}_1}(p_n(t))}{\mathrm{CTM}_1(p_n(t))}\right)\right| \underset{(n\to\infty)}{\overset{\mathcal{D}}{\longrightarrow}} |Z|,$$

where $Z\sim\mathcal{N}(0,1)$. A confidence corridor can then be constructed similarly as in (16).

3 Simulations

This section compares the MAD of our ES estimator $\widehat{\mathrm{CTM}_1}(p_n)$ with several competitors. Specifically, we investigate the optimally bias-corrected estimator $\widehat{\mathrm{ES}}^{(2)}_{p_n}$ of Hill (2015a), the untrimmed $\widehat{\mathrm{ES}}_{p_n}$, the estimator $\widehat{\mathrm{CTM}_1}^{\mathrm{Pick}}(p_n)$ based on the Pickands (1975) estimator

$$\widehat{\gamma}^{\mathrm{Pick}} = \widehat{\gamma}^{\mathrm{Pick}}_{k_n} = \frac{1}{\log 2}\log\left(\frac{X_{(\lfloor k_n/4\rfloor)}-X_{(\lfloor k_n/2\rfloor)}}{X_{(\lfloor k_n/2\rfloor)}-X_{(k_n)}}\right)$$

instead of the Hill (1975) estimator, and classical peaks over threshold (POT) estimation. In short, POT fits a Generalized Pareto distribution to the excesses above some high threshold. The fitted distribution is then used to calculate ES. For more detail, we refer to McNeil and Frey (2000, Sec. 4.1). We also compare the MAD with the kernel-smoothed estimator $\widetilde{ES}_p$ of Scaillet (2004). However, its performance is hardly distinguishable from $\widehat{\mathrm{ES}}_{p_n}$, so that the results are omitted. This is in agreement with Chen (2008). All results in the following are based on 10,000 replications.

	$\mathrm{CTM}_1(p_n)$	$\widehat{\mathrm{ES}}_p$ and $\widetilde{ES}_p$		$\widehat{\mathrm{ES}}_p^{(2)}$		POT
Extrapolation	Yes	No		No		Yes
Sim. inf.	Yes	No		No		No
Serial dep.	L_2-E-NED	geom. α-mixing		geom. α-mixing		i.i.d.
Distr. tail	$\gamma \in (0,1)$	No	$\gamma \in (1/2,1)$	$\gamma \in (0,1/2)$	$\gamma \in [1/2,1)$	$\gamma \in (0,1)$
Conv. rate	$\sqrt{k_n}/\log d_n$	$\sqrt{n}$	$n^{1-\gamma}$	$\sqrt{n}$	$\sqrt{n}/L(n)$	$\sqrt{k_n}$
r-th moment	$r=1$	$r>2$	$r=1$	$r=2$	$r=1$	$r=1$

Table 1: Conditions on serial dependence, the distributional tail, and existing moments $\mathrm{E}\,|X_1|^r < \infty$ under which different ES estimators have a non-degenerate distribution with stated convergence rate. The rows 'Extrapolation' and 'Sim. inf' indicate whether extrapolation beyond the range of the data and simultaneous inference for different levels is possible for the respective estimator. In row 'Distr. tail', $\gamma \in (a,b)$ indicates that $\overline{F}(x) = x^{-1/\gamma}L(x)$ must hold for some $\gamma \in (a,b)$. In row 'Conv. rate', $L(\cdot)$ denotes a slowly varying function.

An overview of the ES estimators under consideration is given in Table 1, which includes sufficient conditions for a non-degenerate limit of these estimators. For $\mathrm{CTM}_1(p_n)$, the results are due to Theorems 1 and 2 in this paper. In case no distributional assumption is imposed, Chen (2008) proves the asymptotic normality of $\widehat{\mathrm{ES}}_p$ and $\widetilde{ES}_p$. Linton and Xiao (2013) derive the asymptotic limits of these two estimators, if $\overline{F}(x) = x^{-1/\gamma}L(x)$ is imposed for $\gamma \in (1/2,1)$. For $\widehat{\mathrm{ES}}_p^{(2)}$, the corresponding reference is Hill (2015c). Finally, for POT we refer to Smith (1987, Thm. 3.2). Note that once extrapolation is required for dependent data, $\mathrm{CTM}_1(p_n)$ is the only estimator with a known asymptotic limiting distribution.

In applying POT, we follow Mancini and Trojani (2011), Chavez-Demoulin *et al.* (2014) and others by choosing the 90%-quantile as a threshold. The estimators $\widehat{\mathrm{CTM}_1}(p_n)$, $\widehat{\mathrm{CTM}_1}^{\mathrm{Pick}}(p_n)$ and $\widehat{\mathrm{ES}}_{p_n}^{(2)}$ depend on a sequence k_n that is only specified asymptotically. Hence, some guidance for the choice of k_n in finite samples is required. For $\widehat{\mathrm{ES}}_{p_n}^{(*)}$ in $\widehat{\mathrm{ES}}_{p_n}^{(2)} = \widehat{\mathrm{ES}}_{p_n}^{(*)} + \widehat{\mathcal{R}}_n^{(2)}$, Hill (2015a, Sec. 3) proposes to choose the trimming sequence $k_n = \min\left\{1, \lfloor 0.25 \cdot n^{2/3}/(\log n)^{2\cdot 10^{-10}} \rfloor\right\}$ as a fixed function of n. However, for the bias correction term $\widehat{\mathcal{R}}_n^{(2)}$, which is a function of the Hill (1975) estimator, he uses a data-dependent choice of the intermediate sequence k_n. We follow Hill's (2015a) recipe in the simulations for $\widehat{\mathrm{ES}}_{p_n}^{(2)}$.

For the choice of $k = k_n$ in $\widehat{\mathrm{CTM}_1}(p_n)$ we again take a different tack and modify a data-adaptive algorithm recently proposed by Daníelsson *et al.* (2016). Their method is based on the following considerations. Replacing p_n by j/n in (9), the Pareto-type tail suggests – similarly as before – the following estimate of the $(1-j/n)$-quantile: $\widehat{x}_{j/n}(k) = (k/j)^{\widehat{\gamma}_k} X_{(k+1)}$. The quality of the Pareto-approximation for this particular choice of k may now be judged by $\sup_{j=1,\ldots,k_{\max}} |X_{(j+1)} - \widehat{x}_{j/n}(k)|$,

i.e., a comparison of empirical quantiles and quantiles estimated using the Pareto-approximation. Here, $k_{\max}$ indicates the range over which the fit is assessed. These considerations motivate the choice

$$k^*_{\text{VaR}} = \underset{k=k_{\min},\ldots,k_{\max}}{\arg\min} \left[\sup_{j=1,\ldots,k_{\max}} \left| X_{(j+1)} - \widehat{x}_{j/n}(k) \right| \right], \tag{17}$$

where $k_{\min}$ is the smallest choice of k one is willing to entertain (see also below). While the choice k^*_{VaR} is well-suited conceptually for quantile estimation and $\widehat{\text{CTM}}_1(p_n)$ is essentially a scaled quantile estimate, it may occasionally happen that $\widehat{\gamma}_{k^*_{\text{VaR}}} \geq 1$, rendering ES estimates $\widehat{\text{CTM}}_1(p_n)$ to be of different sign than quantile estimates.

To avoid such a nonsensical result, we adapt the general idea behind the choice of k^*_{VaR} to our particular task of ES estimation. Instead of assessing the fit of the Pareto-motivated quantile estimates to (non-parametric) empirical quantiles, we now assess the fit of Pareto-motivated ES estimates, $\widehat{\text{CTM}}_1(j/n) = \widehat{x}_{j/n}(k)/(1-\widehat{\gamma}_k)$, to the non-parametric estimates $\widehat{\text{ES}}_{j/n}$ from (1). Then, by analogy, we choose

$$k^*_{\text{ES}} = \underset{k=k_{\min},\ldots,k_{\max}}{\arg\min} \left[\sup_{j=1,\ldots,k_{\max}} \left| \widehat{\text{ES}}_{j/n} - \widehat{\text{CTM}}_1(j/n) \right| \right]. \tag{18}$$

With this particular choice, an estimate $\widehat{\gamma}_{k^*_{\text{ES}}} \geq 1$ was always avoided in our simulations. Since the largest level we use is $p_n = 0.05$, the requirement $np_n/k_n = o(1)$ from (13) suggests $k_{\min} = \lfloor p_n \cdot n \rfloor = \lfloor 0.05 \cdot n \rfloor$. Furthermore, we use $k_{\max} = \lfloor n^{0.9} \rfloor$. We apply the above method for $\widehat{\text{CTM}}^{\text{Pick}}_1(p_n)$ as well, where $\widehat{\text{CTM}}_1(p_n)$ is replaced with $\widehat{\text{CTM}}^{\text{Pick}}_1(p_n)$ in (18). Following Hill (2010), we use the bandwidth $\gamma_n = (k^*_{\text{ES}})^{0.25}$ for $\widehat{\sigma}^2_{k_n}$.

We carry out the comparison for a range of i.i.d. and dependent data. For independent data we use three classes of distributions. First, the Burr distribution, $\text{Burr}(\beta, \lambda, \tau)$, with survivor function

$$\overline{F}(x) = \left(\frac{\beta}{\beta + x^\tau} \right)^\lambda, \qquad x > 0,\ \tau > 0,\ \beta > 0,\ \lambda > 0.$$

This is a popular class of distributions in insurance, because it offers more flexibility than the Pareto Type II distribution (e.g., Burnecki *et al.*, 2011). Its tail index is given by $\alpha = \tau\lambda$ and the slowly varying function $A(\cdot)$ in Assumption 1 can be chosen as a constant multiple of $x^{-\tau}$. Hence, the larger $\tau > 0$, the faster the convergence to true Pareto behavior in (7). In insurance applications one often finds for the tail index that $\alpha \in (1, 2)$ (see, e.g., Resnick, 2007), which motivates our choices of $\tau = 1.5$ and $\lambda = 1$, and $\tau = 6$ and $\lambda = 0.25$, both resulting in $\alpha = 1.5$. For the latter choice where τ is larger (and hence the Pareto approximation more accurate), we expect improved performance of our estimator relative to $\widehat{\text{ES}}_{p_n}$ and $\widehat{\text{ES}}^{(2)}_{p_n}$, where the latter only partially takes into account the Pareto-type tail for bias correction. Note that since $\alpha = 1.5 < 2$, the considered Burr distributions possess infinite

variance.

As a second class, we use the Pareto Type I distribution, Pa(α), with survivor function $\overline{F}(x) = x^{-\alpha}$, $x \geq 1$. For this distribution, the Pareto-type tail assumption (4) even holds without the limit. We use Pa(3) and Pa(1.5), where the latter distribution – unlike the former – does not possess a finite variance, since the tail index is $\alpha = 1.5 < 2$.

Finally, we use Student $t(\nu)$-distributed data, with $\nu = 10$ and $\nu = \infty$, corresponding to the standard normal distribution. We use these two light-tailed distributions (with tail index $\alpha = \nu > 2$) to assess the performance of our estimator in sufficiently challenging cases. Note that the Pareto-type tail assumption (4) is not satisfied for $t(\infty)$. For all the i.i.d. data, we calculate the true ES from $\mathrm{CTM}_1(p) = (1/p)\int_{1-p}^{1} \mathrm{VaR}_\alpha \ \mathrm{d}\alpha$, where VaR_α can simply be obtained from the quantile function of the distributions.

As models for dependent data, we use the following:

GARCH:	$\xi_i = \sigma_i \varepsilon_i,$	where $\sigma_i^2 = 10^{-6} + 0.3X_i^2 + 0.6\sigma_{i-1}^2, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} (0,1)$
AR–GARCH:	$X_i = 0.7X_{i-1} + \xi_i,$	with ξ_i as in GARCH,
AR:	$X_i = 0.7X_{i-1} + \varepsilon_i,$	with i.i.d. ε_i.

Depending on whether the innovations ε_i in these models are standard normally- or $t(\nu)$-distributed, the corresponding models will be termed GARCH-N(0,1), GARCH-t(2.5), AR–GARCH-N(0,1), AR–GARCH-t(2.5), AR-N(0,1) and AR-t(10). Of course, for the GARCH and AR–GARCH models the $t(\nu)$-distributed innovations have to be standardized to have zero mean and unit variance.

While the stationary distribution of AR-N(0,1) is again normal with exponentially decaying tail, the stationary distributions of the models AR-t(10), GARCH-N(0,1) and GARCH-t(2.5) are known to have regularly varying tails in the sense of (4); see Fasen *et al.* (2010). The AR-t(10) inherits its tail index $\alpha = 10$ from the innovations. The tail index of the two GARCH models is determined by the unique positive solution $\alpha > 0$ of

$$\mathrm{E}\left[0.3 \cdot \varepsilon_1^2 + 0.6\right]^{\alpha/2} = 1.$$

For the GARCH-N(0,1) (GARCH-t(2.5)) model, this solution is $\alpha = 4.09$ ($\alpha = 2.18$). To the best of our knowledge, no results on the regular variation of AR(1)–GARCH(1, 1) processes exist. Yet, as both AR(1)–ARCH(1) and GARCH(1, 1) processes have regularly varying tails (see Fasen *et al.*, 2010, and the references therein), the same property is likely to hold for AR(1)–GARCH(1, 1) models as well. We remark that verifying the second-order Assumption 1 is notoriously difficult for time series models, so it is frequently treated as a given (Sun and Zhou, 2014; Hill, 2015c).

For the dependent data, the true $\mathrm{ES}_p = \mathrm{E}[X \mid X > \mathrm{VaR}_p]$ is computed by simulating trajectories $X_1, \ldots, X_N$ of length $N = 100,000$ and computing $\frac{1}{pN} \sum_{i=1}^{N} X_i I_{\{X_i \geq X_{(\lfloor pN \rfloor)}\}}$, as in (1). This is repeated $B = 10,000$ times and the average over all estimates is taken as the true value (Hill, 2015a, p. 17).

The ratios of the MAD of the estimators $\widehat{\mathrm{ES}}_{p_n}$, $\widehat{\mathrm{ES}}^{(2)}_{p_n}$, POT, and $\widehat{\mathrm{CTM}}^{\mathrm{Pick}}_1(p_n)$ over the MAD of $\widehat{\mathrm{CTM}}_1(p_n)$ are displayed in Figure 1 (for independent data) and Figure 2 (for dependent data). Note that in both figures the y-axis limits across rows are different to zoom in on the relevant parts of the respective plots. We consider $p_n = 0.001, 0.002, \ldots 0.05$ and length $n = 2000$. The following conclusions can be drawn from Figure 1:

1. From (a) and (b) we conclude the following. As mentioned above, the $\mathrm{Burr}(1, 0.25, 6)$ distribution more closely resembles a Pareto Type I distribution than the $\mathrm{Burr}(1, 1, 1.5)$, since $\tau = 6$ is larger in the former case. Thus, the better the underlying distribution fits a true Pareto tail shape, the larger the relative advantage of $\widehat{\mathrm{CTM}}_1(p_n)$ – which exploits the Pareto shape – in terms of MAD.

2. From (c) and (d) we see that, while theoretically ES estimation is made more difficult for $\widehat{\mathrm{CTM}}_1(p_n)$ the heavier the tail (Remark 6), for the other estimators it becomes comparatively more difficult, because MAD ratios are much higher for the heavier tailed Pa(1.5) distribution. Also, we find that the performance of $\widehat{\mathrm{CTM}}_1(p_n)$ improves even more vis-à-vis the non-parametric $\widehat{\mathrm{ES}}_{p_n}$, the more extreme the quantile level considered. This is to be expected, because $\widehat{\mathrm{CTM}}_1(p_n)$ focuses on small levels by construction.

   Note that the estimator $\widehat{\mathrm{ES}}^{(2)}_{p_n}$ appears to behave erratically. For Pa(1.5) it works better than $\widehat{\mathrm{ES}}_{p_n}$, yet for Pa(3) it is less precise, particularly for extreme quantiles.

3. The light-tailed distributions in (e) and (f) present challenging cases for our estimation method and indeed all methods relying on the Pareto-type tail assumption in (4), i.e., $\widehat{\mathrm{ES}}^{(2)}_{p_n}$, $\widehat{\mathrm{CTM}}_1(p_n)$ and $\widehat{\mathrm{CTM}}^{\mathrm{Pick}}_1(p_n)$. The best estimators in both cases are the simple non-parametric $\widehat{\mathrm{ES}}_{p_n}$ and POT, which also works for non-fat tailed distributions (Embrechts *et al.*, 1997, Sec. 6.5; McNeil and Frey, 2000). However, the comparative advantage becomes smaller for the $t(10)$-distribution, i.e., the heavier the tail. This is as expected, since for the $t(\nu)$-distribution (with $\gamma = 1/\nu$) $\widehat{\mathrm{ES}}_{p_n}$ has a faster convergence rate than $\widehat{\mathrm{CTM}}_1(p_n)$ for $\gamma < 3/4$, and here we consider $\gamma = 0$ in (e) and $\gamma = 1/10$ in (f) (cf. Remark 6).

The results for the time series models in Figure 2 lend further evidence to the above conclusions 2 and 3. For instance, comparing panels (a) (where $\alpha = 4.09$) and (b) (where $\alpha = 2.18$) we find an

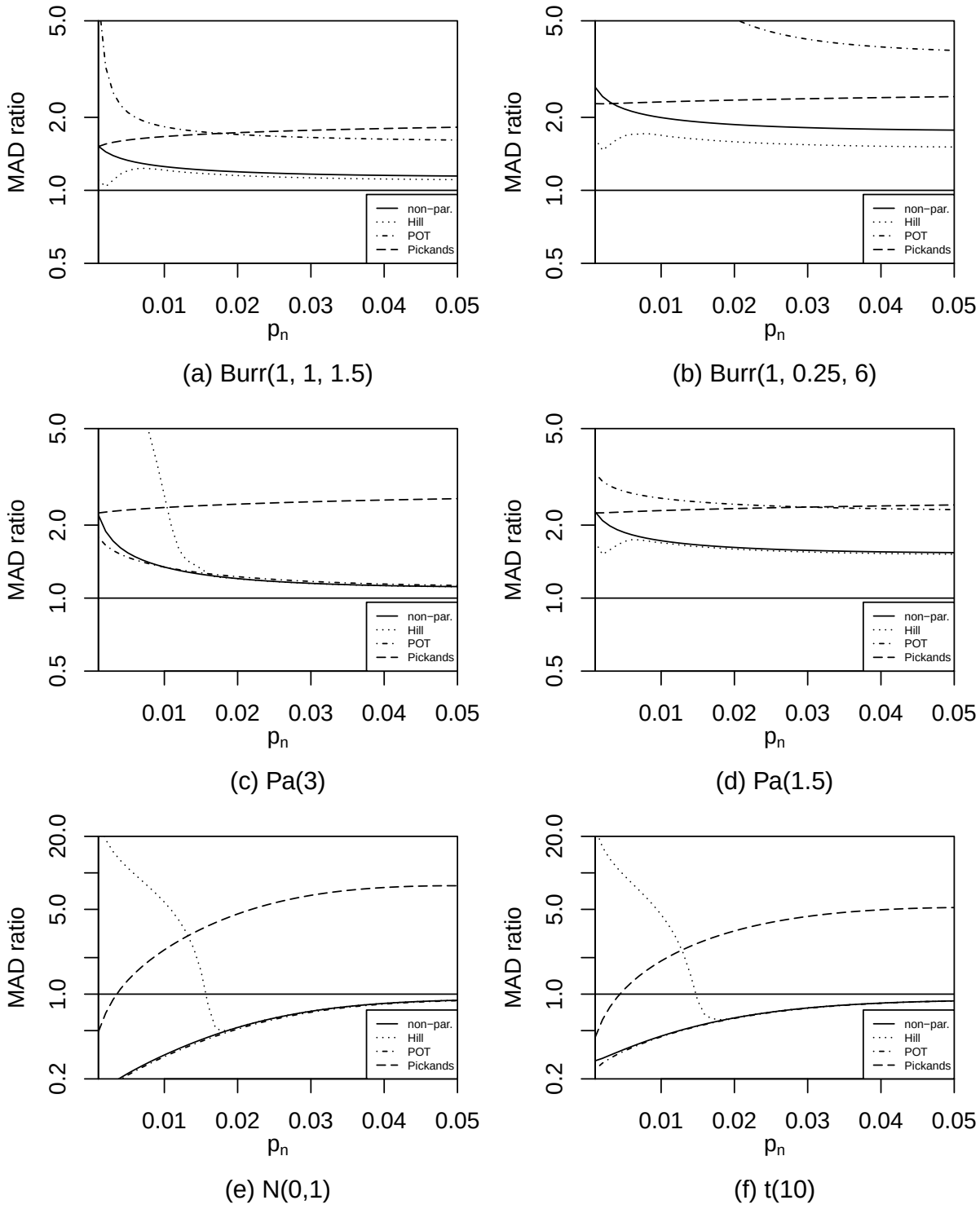

Figure 1: Ratio of MADs for different estimators over MAD for $\widehat{\mathrm{CTM}}_1(p_n)$ for i.i.d. data: the non-parametric estimator $\widehat{\mathrm{ES}}_{p_n}$ (solid), Hill's (2015a) estimator $\widehat{\mathrm{ES}}^{(2)}_{p_n}$ (dashed), POT-based estimator (dot-dashed), $\widehat{\mathrm{CTM}}_1^{\mathrm{Pick}}(p_n)$ based on Pickands's (1975) estimator (long-dashed).

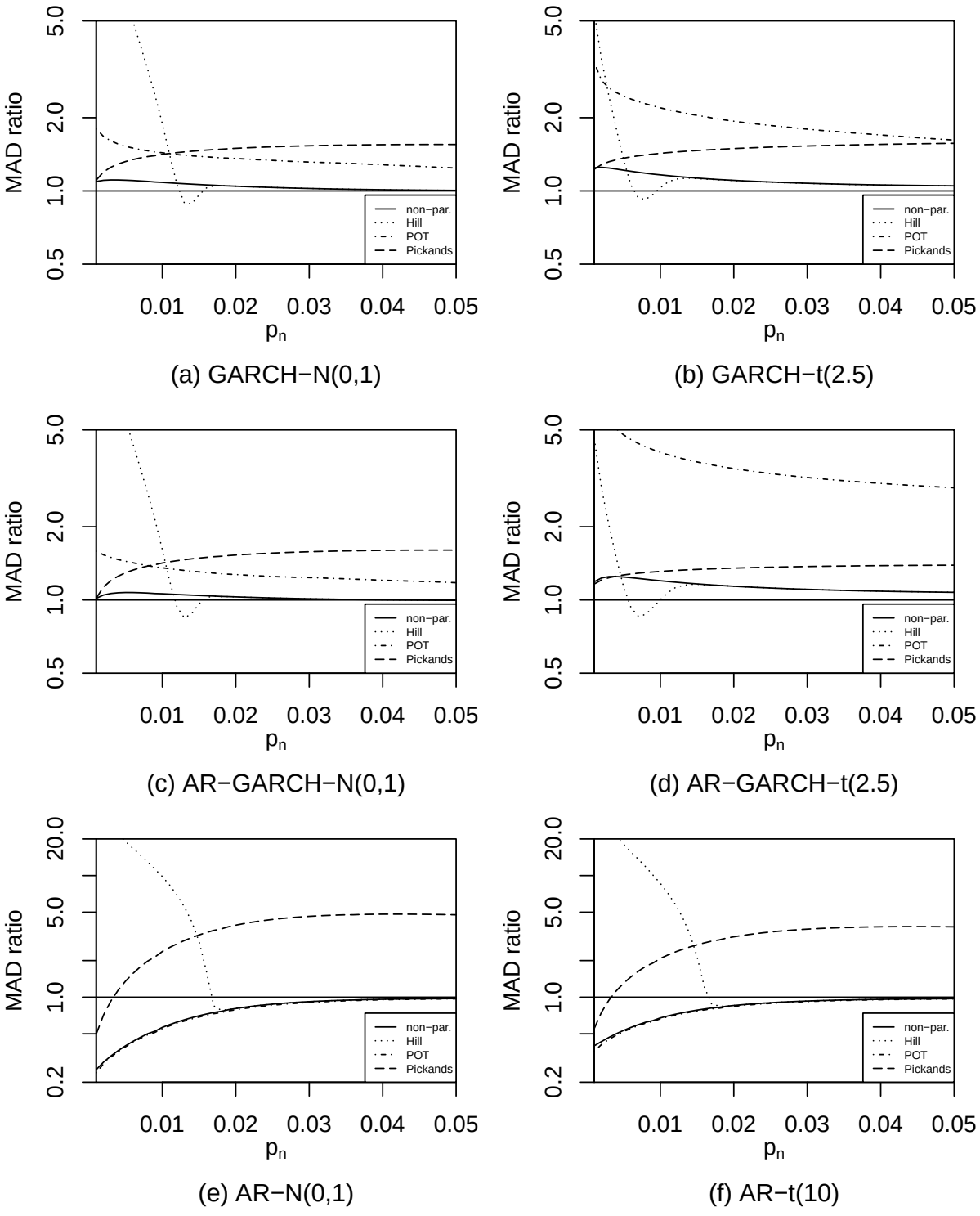

Figure 2: Ratio of MADs for different estimators over MAD for $\widehat{\mathrm{CTM}}_1(p_n)$ for dependent data: the non-parametric estimator $\widehat{\mathrm{ES}}_{p_n}$ (solid), Hill's (2015a) estimator $\widehat{\mathrm{ES}}_{p_n}^{(2)}$ (dashed), POT-based estimator (dot-dashed), $\widehat{\mathrm{CTM}}_1^{\mathrm{Pick}}(p_n)$ based on Pickands's (1975) estimator (long-dashed).

improved performance of $\widehat{\mathrm{CTM}}_1(p_n)$ for the heavier tailed model and the improvements are more marked the more extreme the quantile level considered. E.g., the non-parametric estimator $\widehat{\mathrm{ES}}_{p_n}$ has up to 10% higher MAD in (a) and up to 25% higher MAD in (b). Regarding the third conclusion, we again find our estimator outperformed by $\widehat{\mathrm{ES}}_p$ and POT for the models where the stationary distribution is either light-tailed (i.e., does not possess a Pareto-type tail) or hardly distinguishable from a light-tailed distribution (Figure 2 (e) & (f)). It is interesting to note that POT performs quite well in (e) and (f) despite being valid only for independent data. We remark that the first conclusion is hard to verify because, as mentioned above, the second-order behavior in Assumption 1 is largely unexplored for time series models.

Comparing only $\widehat{\mathrm{ES}}_{p_n}$ and $\widehat{\mathrm{CTM}}_1(p_n)$ in Figures 1 and 2, we find that each estimator offers a different robustness-efficiency trade-off. The estimates $\widehat{\mathrm{ES}}_{p_n}$ are clearly more robust, while – at least for dependent data – the loss in efficiency seems mild. Having said that, $\widehat{\mathrm{CTM}}_1(p_n)$ is more efficient uniformly across p_n when the tails are rather heavy, as is frequently the case for financial data.

Figures 1 and 2 suggest that particularly for small levels $p_n \leq 0.01$ the relative performance of $\widehat{\mathrm{CTM}}_1(p_n)$ is very good. Hence, we investigate coverage of our confidence corridors for $p_n = 0.01$ and $t \in [0.1, 1]$, such that all upper quantiles in the range between 0.1% and 1% are covered. Following the suggestion of Daníelsson *et al.* (2016) for the choice of k_n in (17) (and using a bandwidth of $\gamma_n = (k^*_{\mathrm{VaR}})^{0.25}$), we have calculated coverage probabilities of the 90%-confidence corridor for $x_{p_n}(t)$ in (16) for the i.i.d. data. Table 2 displays the results. For the heavy-tailed models with Pareto-type tail, we find the uniform coverage to be quite convincing. Unsurprisingly, once the Pareto assumption is no longer satisfied – as for the N(0,1)-distribution – or the tails are very light – as for the $t(10)$-distribution –, the extrapolation to the very small levels of 0.1% is no longer accurate and coverage breaks down.

Finally, we investigate for each distribution how much coverage changes, when only considering the least extreme 99%-quantile $x_{p_n=0.01}$, i.e., when considering *pointwise* instead of *uniform* coverage.

	Burr$(1,1,1.5)$	Burr$(1,0.25,6)$	Pa(3)	Pa(1.5)	N(0,1)	$t(10)$
Coverage $x_{p_n}(t)$	85.7	90.9	91.3	90.4	0	7.3
Coverage x_{p_n}	89.1	92.4	92.3	92.3	97.6	96.3
k^*_{VaR}	216	310	315	318	107	107
k^*_{ES}	194	310	321	330	105	107

Table 2: Coverage (in %) of true $x_{p_n}(t)$ and x_{p_n} for $p_n = 0.01$ and $t \in [0.1, 1]$. Nominal level set to 90%. Value of k^*_{VaR} is the average value of (17) used in the estimation of $x_{p_n}(t)$ and x_{p_n}. Value of k^*_{ES} is the average value of (18).

Again, Table 2 shows the empirical coverage probabilities. Interestingly, for the distributions where uniform coverage was accurate, the pointwise coverage is not markedly different. This suggests that much of the estimation uncertainty lies in estimating the least extreme quantile ($x_{0.01}$ in this case) and the extrapolation to smaller levels does not significantly affect coverage. We thus conclude that the Pareto tail pins down the actual tail behavior very well for these models. For the remaining two models, the results change dramatically. Since now extrapolation is only required for the least extreme quantile, we find coverage to be much improved, although somewhat too high.

We also present the average values of k^*_{VaR} used in the estimation of the quantiles in Table 2. Additionally, average values of k^*_{ES} are also shown. These value were used to compute $\text{CTM}_1(p_n)$ in Figure 1. Both choices show qualitatively similar behavior. Roughly, the closer the tails are to true Pareto behavior, the more upper order statistics are used for estimation. For instance, for the Burr$(1, 0.25, 6)$-distribution we use $k^*_{\text{ES}} = 310$ on average, while we only use $k^*_{\text{ES}} = 194$ for Burr$(1, 1, 1.5)$.

Overall, we conclude that – where appropriate – our ES estimator $\text{CTM}_1(p_n)$ can improve estimation precision vis-à-vis other commonly used semi-parametric and non-parametric estimators. These relative improvements tend to be larger, the better the Pareto-approximation, and/or the more extreme the quantile to be estimated, and/or the heavier the tail of the data. In these cases, our uniform confidence intervals also appear to work well. If the Pareto-approximation is not satisfied or the tails are very light, neither $\widehat{\text{CTM}}_1(p_n)$ nor the uniform confidence intervals work well. This highlights the importance of empirically checking the Pareto-type tail assumption (4), as is done in the following Section 4.

We have only included semi- and non-parametric estimators of ES in our simulations. In case belief in a parametric model is strong, parametric estimators of ES may offer a good robustness-efficiency trade-off. For instance, in the above GARCH-N(0,1) model, one may estimate ES from a sample $\xi_1, \ldots, \xi_n$ by fitting a GARCH-N(0,1) model in the first step (using Gaussian maximum likelihood), and in a second step calculate ES implied by the fitted model. To carry out the second step, one may simply adopt a 'brute-force' approach, similarly as for the calculation of the true values of ES. Due to computational constraints, we choose to simulate $B = 1,000$ trajectories of length $N = 2,000$. Proceeding in this way, we find MAD ratios for the parametric estimator to be 0.66 for $p_n = 0.001$ and 0.78 for $p_n = 0.05$; cf. Figure 2 (a). We refer to Kang and Babbs (2012) for an empirical application of the brute-force method.

There may be at least two drawbacks to the brute-force approach. First, it may be computationally prohibitive in large scale applications, where often quick and accurate estimates of tail risk for a large

number of assets are sought. This may be the case when having to set margin or capital requirements for a very large number of assets or client portfolios. For instance, calculating the true value of ES for a (true or fitted) GARCH-N(0,1) model based on $B = 10,000$ trajectories of length $N = 100,000$, takes about one hour on a standard desktop computer. The computational burden can increase even further if a bootstrap procedure is required to obtain a measure of estimation uncertainty. The computational cost of our proposed method, which includes a measure of estimation uncertainty, is negligible in contrast.

Second, the approach of relying on a parametric model can be dangerous in our extreme value setting. Drees (2008) has shown in simulations for extreme VaR estimation that even a slight misspecification of the model, that is not detectable by statistical tests, can lead to distorted estimates. Thus, even if there is strong evidence for a certain parametric model, parametric estimates of extreme ES (or indeed any other extreme risk measure) should be supplemented by semi-parametric extreme value estimates in our view.

4 An application to extreme returns of VW shares

In this section we illustrate the use of Theorems 1 and 2 by calculating ES estimates and VaR corridors. We do so for the $n = 3474$ log-losses of the German auto maker VW's ordinary shares from March 27, 1995 to October 24, 2008 downloaded from *finance.yahoo.com*. (If P_i denotes the adjusted closing prices, the log-losses are defined as $X_i = -\log(P_i/P_{i-1})$. A similar analysis could of course be carried out for the log-returns $-X_i$.) This period was chosen to precede the tumultuous week of trading in VW shares from October 27, 2008 to October 31, 2008. Preceding this week, the sports car maker Porsche built up a huge position in VW shares in a takeover attempt that ultimately failed. Porsche announced on Sunday – October 26, 2008 – that it had indirect control of 74.1% of VW. Since the German state of Lower Saxony owned another 20.2% of VW, this left short-sellers scrambling to buy the remaining shares to close their positions. The shares closed at €210.85 on Friday, October 24, more than doubling on the next trading day – Monday, October 27 – to €520, and again almost doubling to €945 on Tuesday. During a few minutes of trading on Tuesday, VW was the world's most valuable company by market capitalization. Wednesday then saw the shares almost halve in value, closing at €517.

The magnitude of the log-returns from Monday, Tuesday and Wednesday of 0.904, 0.597 and -0.603, respectively, is very large indeed if compared with previous historical returns, which are displayed in Figure 3. In fact, a log-loss of 0.603 has not been observed before. Thus, one must assess the magnitude of a previously unseen event, which provides a natural application of the extreme value

methods proposed in this paper. Since our estimator $\widehat{\mathrm{CTM}}_1(p_n)$ is – to the best of our knowledge – the only one known to be asymptotically normal under serial dependence (which is obviously present in the VW log-returns), while also allowing for extrapolation, it is natural to consider it as it is the only theoretically sound choice; cf. Table 1.

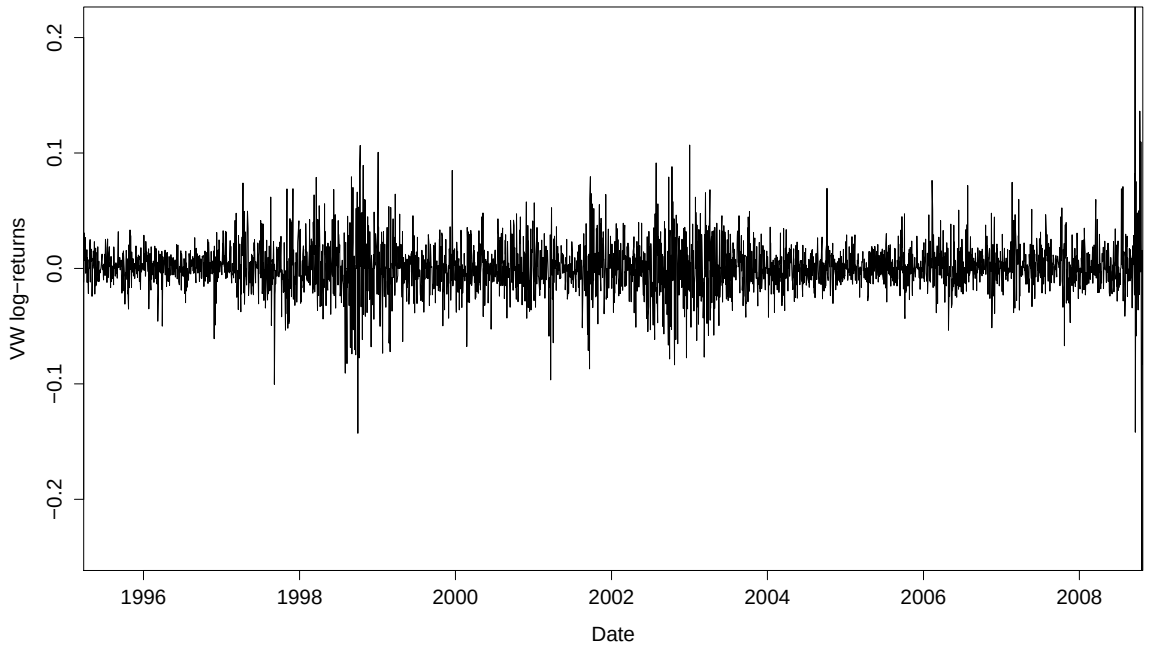

Figure 3: VW log-returns from March 27, 1995 to October 24, 2008

To get a better sense of the significance of the log-loss of 0.603 we apply the methodology developed in this paper. Before doing so, we check that Theorems 1 and 2 may reasonably be applied. To this end, we fit a standard AR(1)–GARCH(1, 1) model with skewed-t distributed innovations to the time series. Visual inspection and standard Ljung-Box tests of the (raw and squared) standardized residuals reveal that they may reasonably be considered i.i.d. and thus an adequate fit of our model. Under quite general conditions, AR(1)–GARCH(1, 1) models are stationary and L_2-E-NED (Hill, 2011, Sec. 4).

To the best of our knowledge, the Pareto-type tail assumption (4) has only been verified for the smaller class of AR(1)–ARCH(1) models by Borkovec and Klüppelberg (2001), so it seems worthwhile to check it empirically. To do so, we use the *Pareto quantile plot* of Beirlant *et al.* (1996). The idea is to use (6), i.e., $U(x) = x^\gamma L_U(x)$. Since $\log L_U(x)/\log x \to 0$ as $x \to \infty$ (de Haan and Ferreira, 2006, Prop. B.1.9.1), we obtain $\log U(x) \sim \gamma \log x$. Thus, for small j, the plot of

$$\left(-\log\left(\frac{j}{n+1}\right),\ \log X_{(j)} \approx \log U((n+1)/j)) \right), \qquad j = 1, \ldots, n,$$

should be roughly linear with positive slope $\gamma > 0$, if (6) holds with positive extreme value index. Since some log-losses are negative, rendering $\log X_{(j)}$ to be undefined, we only use the positive log-losses for the Pareto quantile plot in panel (a) in Figure 4. A roughly linear behavior with positive slope can be discerned from $-\log(j/(n+1)) = 2$ onwards, but it is not quite satisfactory, as the *Hill plot* of $k_n \mapsto \widehat{\gamma}_{k_n}$ in panel (b) is highly unstable. A better approximation to linearity in the Pareto quantile plot and more stable Hill estimates can often be obtained by a slight shift of the data. Here, a positive shift of 0.05 sufficed, as the plots in (c) and (d) for the shifted data reveal. The positive slope of the roughly linear portion in the Pareto quantile plot and the strictly positive and very stable Hill estimates for k_n up to 1000 strongly suggest a Pareto-type tail with positive tail index for the VW log-losses. From the stable portion of the Hill plot in panel (d) we read off an estimate of the extreme value index of $\widehat{\gamma} = 0.2$. The 95%-confidence intervals for γ for different values of k_n are indicated by the shaded area in panel (d). They were computed using Thm. 2 of Hill (2010) and $\widehat{\sigma}_{k_n}$; see also Equation (A.1) in the Appendix. The null hypothesis $\gamma = 1$, which would invalidate our analysis for ES, is clearly rejected for all k_n. Since there is strong evidence for $\gamma < 0.5$, we also conclude that the log-losses possess a finite variance. All in all, we are confident that Theorems 1 and 2 can be applied.

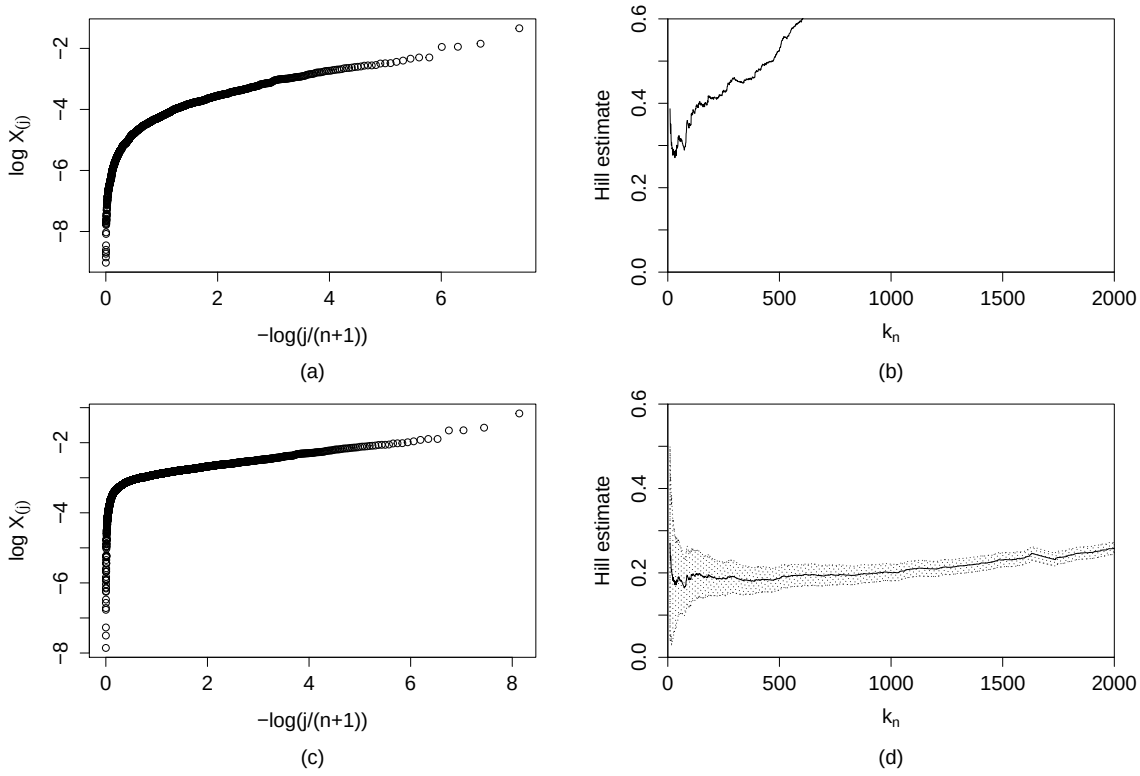

Figure 4: Pareto quantile plot and Hill plot for raw log-losses (in (a) and (b)) and for log-losses shifted by 0.05 (in (c) and (d)). The shaded area around the Hill estimates in panel (d) signifies 95%-confidence intervals.

Figure 5 displays the results, i.e., the VaR and ES estimates for levels between $p_n = 0.05$ and 0.0001. In view of the much more stable Hill estimates (upon which our VaR and ES estimators are based) for the shifted data in Figure 4, we carry out VaR and ES calculations for the shifted data and then subtract 0.05 from the results to arrive at estimates for the original series of log-losses. To compute VaR and ES estimates we use $k^*_{\text{VaR}} = 365$ and $k^*_{\text{ES}} = 902$ respectively, which have been calculated with $k_{\min}$ and $k_{\max}$ chosen as in the simulations. Incidentally, from the Hill plot in panel (d) of Figure 4 the use of k_n around a similar value of around 1000 seems sensible, because smaller values of k_n lead to roughly the same estimate (yet a slower convergence rate of $\widehat{\gamma}$) and for larger values the Hill plot is slightly upward trending, suggesting a possible bias. The choice of $p_n = 0.05$ is compatible with the theory requirement $np_n = o(k_n)$, since $np_n = 3474 \cdot 0.05 = 173.7$ is small relative to $k_n = k^*_{\text{VaR}} = 365$ and $k_n = k^*_{\text{ES}} = 902$.

In more detail, Figure 5 displays VaR estimates $\widehat{x}_{p_n}$ (solid) and ES estimates $\widehat{\text{CTM}}_1(p_n)$ (dotted), $\widehat{\text{CTM}}_1^{\text{Pick}}(p_n)$ (dotted) and POT (dot-dashed). As is customary in extreme value theory, the risk level p_n is not plotted directly, but rather the m-year return level; see, e.g., Coles (2001, Sec. 4.4.2). Since there are approximately 250 trading days in a year, a probability of $p_n = 1/250$ corresponds to a return period of 1 year. Thus, the return level with return period of 1 year is, on average, only exceeded once a year. Similarly, the 2-year return period corresponds to $p_n = 1/500$, and so forth. As is also customary, we plot the return period on a log-scale to zoom in on the very large return periods that are of particular interest in risk management. The estimated and empirical quantiles (calculated simply as $X_{(\lfloor np_n \rfloor + 1)}$) are in reasonable agreement, strengthening further the belief that our methods are appropriate.

We have also estimated VaR via the brute-force method described in Section 3 (with $B = 10,000$ and $N = 100,000$). However, beyond return periods of one year, the brute-force VaR estimates lie above even the ES estimates of the other methods. Thus, there is no agreement in the tails between the empirical quantiles and those obtained by brute-force, suggesting that this method is not adequate here. This empirical result is in line with the above mentioned simulations by Drees (2008), where a slight mis-specification of the model is magnified in the tails. The discrepancy between the VaR estimates reflects the fact that, while the AR(1)–GARCH(1,1) models the volatility dynamics quite well, it does not model extreme quantiles well. Due to this, we do not consider the brute-force approach further in this section.

Most empirical estimates lie within the 95%-confidence corridor for VaR at different levels (grey area in Figure 5) calculated from Theorem 2. It has the interpretation that the null hypothesis that the true $x_{p_n}(t)$ lies in this gray area (for $t = [0.002, 1]$ and $p_n = 0.05$) cannot be rejected at the

5% level. In this sense, it provides an informative description of the tail region.

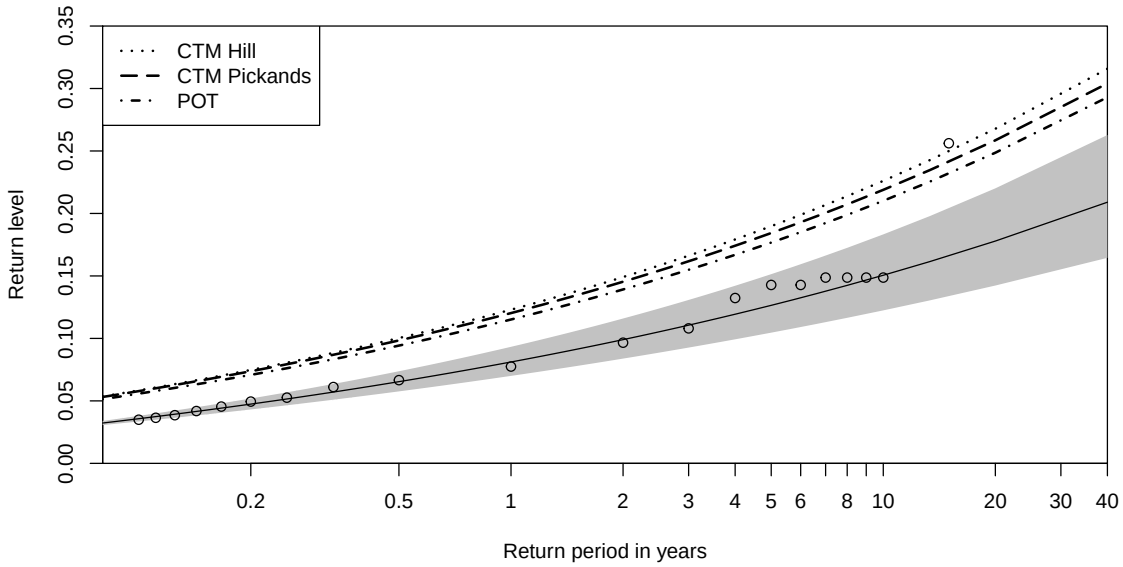

Figure 5: Return level plot for VW log-losses (solid line). Grey area indicates 95%-confidence corridor for return levels. ES estimates shown as the dotted ($\widehat{\mathrm{CTM}}_1(p_n)$), dashed ($\widehat{\mathrm{CTM}}_1^{\mathrm{Pick}}(p_n)$) and dot-dashed (POT) line.

The dotted line in Figure 5 indicates ES estimates based on $\widehat{\mathrm{CTM}}_1(p_n)$. As the expected loss given a VaR exceedance, the ES estimates provide further insight on tail risk. Since we applied a shift to the data, which affects the Hill (1975) estimator (upon which $\widehat{\mathrm{CTM}}_1(p_n)$ is based), Figure 5 also includes additional ES estimates as a robustness check. The estimates $\widehat{\mathrm{CTM}}_1^{\mathrm{Pick}}(p_n)$ (dashed line), based on the shift-invariant Pickands (1975) estimator, do not differ significantly from $\widehat{\mathrm{CTM}}_1(p_n)$. The same holds for the – also shift-invariant – POT estimates (dot-dashed line), even though POT can only validly be applied for independent data.

All in all, nothing in Figure 5 suggests that a log-loss of 0.603 was to be expected. Even ES estimates for a return period of 40 years do not come close to this value. Of course, further extrapolation of VaR and ES estimates in Figure 5 would be possible to see for which return period a return level of 0.603 is obtained. However, in view of the restriction on p_n imposed by (13) (see also Remark 2) and related applications of extreme value theory (Drees, 2003), we feel that extrapolation well beyond a level of $p_n = 0.0001 \approx 1/(2.87 \cdot n)$ is no longer justified.

The above analysis has revealed that the log-loss of 0.603 on October 29, 2008 was quite an unexpected event based on estimates of the d.f. $F(\cdot)$ of the losses. However, *conditionally* on the extraordinary returns of 0.904 and 0.597 prior to October 29, 2008, such a log-loss may be much more likely due to the well-known persistence of volatility. To frame the problem more formally, we define

the *conditional* d.f. of the log-losses given the past history as

$$F_n(x) := P\left\{X_{n+1} \leq x \mid X_n, X_{n-1}, \ldots\right\}.$$

Now with regard to this conditional d.f. a log-loss of 0.603 may have been more likely. We assess this in the remainder of this section.

A natural candidate for a model of $F_n(\cdot)$ is the (zero-mean) AR(1)–GARCH(1,1) model previously found to provide a good description of the volatility dynamics. Under this model, the (right-tail) conditional VaR, $\mathrm{CVaR}_{p,n}$, and conditional ES, $\mathrm{CES}_{p,n}$, can be written as (see, e.g., McNeil and Frey, 2000)

$$\mathrm{CVaR}_{p,n} := F_n^{\leftarrow}(1-p) = \phi_1 X_n + \sigma_n \,\mathrm{VaR}_p^U, \tag{19}$$

$$\mathrm{CES}_{p,n} := \mathrm{E}\left[X_{n+1} \;\middle|\; X_{n+1} > \mathrm{CVaR}_p, X_n, X_{n-1}, \ldots\right] = \phi_1 X_n + \sigma_n \,\mathrm{ES}_p^U, \tag{20}$$

where $\mathrm{VaR}_p^U := F_U^{\leftarrow}(1-p)$, $\mathrm{ES}_p^U := \mathrm{E}[U \mid U > \mathrm{VaR}_p^U]$ with $F_U(\cdot)$ denoting the d.f. of the innovations U_i, cf. Equations (11) and (12). The implication is that $\mathrm{CVaR}_{p,n}$ and $\mathrm{CES}_{p,n}$ can be estimated in a two-step procedure. First, we fit the model via QMLE to extract $\widehat{\phi}_1$, $\widehat{\sigma}_n$ and the residuals $\widehat{U}_1, \ldots, \widehat{U}_n$. In a second step, estimates of VaR_p^U and ES_p^U based on the residuals are obtained using the estimators presented in this paper. Denote these by $\widehat{\mathrm{VaR}}_p^U$ and $\widehat{\mathrm{ES}}_p^U$, respectively. Thus, to estimate (19) and (20) we use

$$\widehat{\mathrm{CVaR}}_{p,n} = \widehat{\phi}_1 X_n + \widehat{\sigma}_n \widehat{\mathrm{VaR}}_p^U \qquad \text{and} \qquad \widehat{\mathrm{CES}}_{p,n} = \widehat{\phi}_1 X_n + \widehat{\sigma}_n \widehat{\mathrm{ES}}_p^U. \tag{21}$$

Since we now take into account the two additional extreme returns, we present estimates $\widehat{\mathrm{CVaR}}_{p,n+2}$ and $\widehat{\mathrm{CES}}_{p,n+2}$ in Figure 6. Note that there was no need to include these two positive returns in our unconditional analysis, because our estimators only exploit the form of the left-tail of the returns. In the conditional analysis in contrast, the impact of the extreme positive gains on estimates of the conditional loss – through the volatility $\sigma_n^2 = \omega + \alpha_1(X_{n-1} - \phi_1 X_{n-2})^2 + \beta_1 \sigma_{n-1}^2$ – can be substantial; cf. (19) and (20). Note that to estimate VaR_p^U and ES_p^U in the second step of the two-step procedure, we have again applied a shift (by 1.5 to the right) to the data for more stable Hill (1975) estimates. The Hill plot for the residuals (which is omitted for brevity) indicates an extreme value index γ slightly below 0.2. This provides evidence for finite fourth moments of the innovations, which is required for asymptotically normal QMLE. Unlike in the unconditional analysis, the optimal choices for the number of upper order statistics are now in complete agreement with $k^*_{\mathrm{VaR}} = k^*_{\mathrm{ES}} = 204$.

Figure 6 demonstrates that – conditionally on the state of the market – a log-loss of 0.603 was not even a particularly extreme event. The least extreme estimates of $\mathrm{CVaR}_{p,n+2}$ (solid line) and

$\mathrm{CES}_{p,n+2}$ (dotted line) shown in Figure 6 are for level $p = 5\%$, and even these are well above 0.603. In fact, using completely non-parametric estimates of VaR_p^U and ES_p^U in the two-step procedure reveals that a log-loss of 0.603 roughly corresponds to $\mathrm{CVaR}_{p,n+2}$ at level $p = 9\%$. Thus, we conclude that if current volatility levels are taken into account, a log-loss of the observed magnitude could have been expected and consequently planned for. Of course, given the unstable state of the market, it is highly questionable if one could have found a counterparty willing to hedge the risk.

Similarly as in the unconditional analysis, the result of the conditional analysis is corroborated by employing additional shift-invariant estimators for $\widehat{\mathrm{ES}}_p^U$ in (21). Using a POT-based estimator again yields similar results. The estimates based on the Pickands (1975) estimator are all lower. However, these may be considered unreliable, as for longer return periods the CES estimates are sometimes barely above the non-parametrically estimated CVaR values, which should be a natural lower bound for the corresponding CES values. Based on the close agreement between POT estimates and $\widehat{\mathrm{CTM}}_1(p_n)$, our results appear robust.

5 Summary

Our first main contribution is to derive central limit theory for a wide range of popular risk measures – including VaR and ES – in time series. As in Linton and Xiao (2013) and Hill (2015a), we do so under

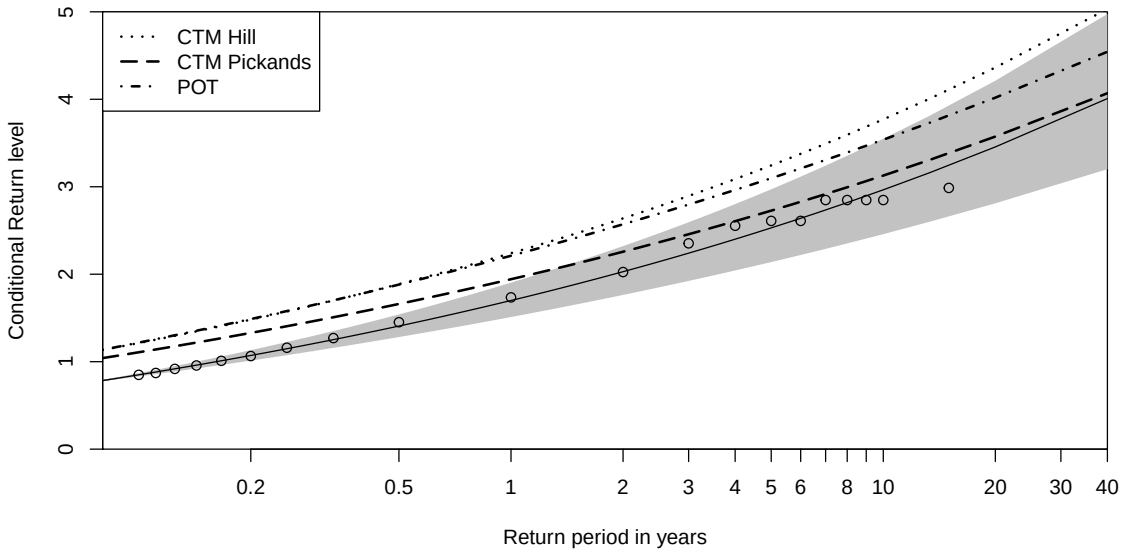

Figure 6: Plot of $\widehat{\mathrm{CVaR}}_{p,n+2}$ for VW log-losses (solid line). Grey area indicates 95%-confidence corridor, which are calculated ignoring the parameter estimation uncertainty of the AR(1)–GARCH(1,1) parameters. Estimates of $\mathrm{CES}_{p,n+2}$ shown as the dotted ($\widehat{\mathrm{ES}}_p^U = \widehat{\mathrm{CTM}}_1(p_n)$), dashed ($\widehat{\mathrm{ES}}_p^U = \widehat{\mathrm{CTM}}_1^{\mathrm{Pick}}(p_n)$) and dot-dashed (POT) line.

a Pareto-type tail assumption. Yet, we exploit the Pareto approximation to motivate an estimator of (among other risk measures) ES, whereas Linton and Xiao (2013) consider a non-parametric ES estimator and Hill (2015a) only uses the Pareto assumption for bias correction of his tail-trimmed ES estimator. Asymptotic theory is derived under an E-NED property, which is significantly more general than the geometrically α-mixing assumption of Linton and Xiao (2013) and Hill (2015a). It is shown in simulations that our estimator (which fully takes into account the regularly varying tail) often provides better estimates in terms of MAD than a wide range of competitors. Our second main contribution is to derive uniform confidence corridors for VaR and also the other risk measures covered by our analysis. Furthermore, we propose a method for choosing the sample fraction k_n used in the estimation of ES, which is used in the simulations. Finally, we illustrate our procedure with VW log-losses prior to the takeover attempt by Porsche. We find that the huge losses in the aftermath of this failed bid were statistically very unlikely. Yet, taking into account the extremely volatile state of the market, these losses were to be expected.

References

Andrews D. 1984. Non-strong mixing autoregressive processes. *Journal of Applied Probability* **21**: 930–934.

Artzner P, Delbaen F, Eber JM, Heath D. 1999. Coherent measures of risk. *Mathematical Finance* **9**: 203–228.

Basel Committee on Banking Supervision. 2016. *Minimum Capital Requirements for Market Risk.* Basel: Bank for International Settlements (http://www.bis.org/bcbs/publ/d352.pdf).

Beirlant J, Vynckier P, Teugels J. 1996. Tail index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association* **91**: 1659–1667.

Borkovec M, Klüppelberg C. 2001. The tail of the stationary distribution of an autoregressive process with ARCH(1) errors. *The Annals of Applied Probability* **11**: 1220–1241.

Burnecki K, Janczura J, Weron R. 2011. Building loss models. In Čížek P, Härdle W, Weron R (eds.) *Statistical Tools for Finance and Insurance*, Berlin: Springer, 2 edn., pages 363–370.

Chan N, Deng SJ, Peng L, Xia Z. 2007. Interval estimation of value-at-risk based on GARCH models with heavy-tailed innovations. *Journal of Econometrics* **137**: 556–576.

Chavez-Demoulin V, Embrechts P, Sardy S. 2014. Extreme-quantile tracking for financial time series. *Journal of Econometrics* **181**: 44–52.

Chen S. 2008. Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics* **6**: 87–107.

Coles S. 2001. *An Introduction to Statistical Modeling of Extreme Values*. London: Springer.

Csörgő S, Haeusler E, Mason D. 1991. The asymptotic distribution of extreme sums. *The Annals of Probability* **19**: 783–811.

Daníelsson J. 2011. *Financial Risk Forecasting*. Chichester: Wiley.

Daníelsson J, de Haan L, Ergun L, de Vries C. 2016. Tail index estimation: Quantile driven threshold selection.

de Haan L, Ferreira A. 2006. *Extreme Value Theory*. New York: Springer.

Drees H. 2003. Extreme quantile estimation for dependent data, with applications to finance. *Bernoulli* **9**: 617–657.

Drees H. 2008. Some aspects of extreme value statistics under serial dependence. *Extremes* **11**: 35–53.

Einmahl J, de Haan L, Zhou C. 2016. Statistics of heteroscedastic extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78**: 31–51.

El Methni J, Gardes L, Girard S. 2014. Non-parametric estimation of extreme risk measures from conditional heavy-tailed distributions. *Scandinavian Journal of Statistics* **41**: 988–1012.

Embrechts P, Klüppelberg C, Mikosch T. 1997. *Modelling Extremal Events*. Berlin: Springer.

Engle R, Bollerslev T. 1986. Modelling the persistence of conditional variances. *Econometric Reviews* **5**: 1–50.

Fasen V, Klüppelberg C, Schlather M. 2010. High-level dependence in time series models. *Extremes* **13**: 1–33.

Francq C, Zakoïan JM. 2016. Looking for efficient QML estimation of conditional VaRs at multiple risk levels. *Annals of Economics and Statistics* **123/124**: 9–28.

Geluk J, de Haan L, Resnick S, Stărică C. 1997. Second-order regular variation, convolution and the central limit theorem. *Stochastic Processes and their Applications* **69**: 139–159.

Gomes M, Pestana D. 2007. A sturdy reduced-bias extreme quantile (VaR) estimator. *Journal of the American Statistical Association* **102**: 280–292.

Gupta A, Liang B. 2005. Do hedge funds have enough capital? A value-at-risk approach. *Journal of Financial Economics* **77**: 219–253.

Hill B. 1975. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics* **3**: 1163–1174.

Hill J. 2009. On functional central limit theorems for dependent, heterogeneous arrays with applications to tail index and tail dependence estimation. *Journal of Statistical Planning and Inference* **139**: 2091–2110.

Hill J. 2010. On tail index estimation for dependent, heterogeneous data. *Econometric Theory* **26**: 1398–1436.

Hill J. 2011. Tail and nontail memory with applications to extreme value and robust statistics. *Econometric Theory* **27**: 844–884.

Hill J. 2013. Least tail-trimmed squares for infinite variance autoregressions. *Journal of Time Series Analysis* **34**: 168–186.

Hill J. 2015a. Expected shortfall estimation and Gaussian inference for infinite variance time series. *Journal of Financial Econometrics* **13**: 1–44.

Hill J. 2015b. Robust estimation and inference for heavy tailed GARCH. *Bernoulli* **21**: 1629–1669.

Hill J. 2015c. Tail index estimation for a filtered dependent time series. *Statistica Sinica* **25**: 609–629.

Hoga Y. 2017a. Change point tests for the tail index of β-mixing random variables. *Econometric Theory* **33**: 915–954.

Hoga Y. 2017b. Testing for changes in (extreme) VaR. *The Econometrics Journal* **20**: 23–51.

Hoga Y, Wied D. 2017. Sequential monitoring of the tail behavior of dependent data. *Journal of Statistical Planning and Inference* **182**: 29–49.

Hong J, Elshahat A. 2010. Conditional tail variance and conditional tail skewness. *Journal of Financial and Economic Practice* **10**: 147–156.

Hsing T. 1991. On tail index estimation using dependent data. *The Annals of Statistics* **19**: 1547–1569.

Ibragimov R, Jaffee D, Walden J. 2009. Non-diversification traps in markets for catastrophic risk. *Review of Financial Studies* **22**: 959–993.

Ibragimov R, Walden J. 2011. Value at risk and efficiency under dependence and heavy-tailedness: Models with common shocks. *Annals of Finance* **7**: 285–318.

Kang L, Babbs SH. 2012. Modeling overnight and daytime returns using a multivariate generalized autoregressive conditional heteroskedasticity copula model. *The Journal of Risk* **14**: 35–63.

Kuester K, Mittnik S, Paolella M. 2006. Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics* **4**: 53–89.

Ling S. 2007. Self-weighted and local quasi-maximum likelihood estimators for ARMA–GARCH/IGARCH models. *Journal of Econometrics* **140**: 849–873.

Linton O, Xiao Z. 2013. Estimation of and inference about the expected shortfall for time series with infinite variance. *Econometric Theory* **29**: 771–807.

Mancini L, Trojani F. 2011. Robust value at risk prediction. *Journal of Financial Econometrics* **9**: 281–313.

McNeil A, Frey R. 2000. Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance* **7**: 271–300.

Novak S, Beirlant J. 2006. The magnitude of a market crash can be predicted. *Journal of Banking & Finance* **30**: 453–462.

Pan X, Leng X, Hu T. 2013. The second-order version of Karamata's theorem with applications. *Statistics & Probability Letters* **83**: 1397–1403.

Pickands J. 1975. Statistical inference using extreme order statistics. *The Annals of Statistics* **3**: 119–131.

Resnick S. 2007. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. New York: Springer.

Scaillet O. 2004. Nonparametric estimation and sensitivity analysis of expected shortfall. *Mathematical Finance* **14**: 115–129.

Smith R. 1987. Estimating tails of probability distributions. *The Annals of Statistics* **15**: 1174–1207.

Sun P, Zhou C. 2014. Diagnosing the distribution of GARCH innovations. *Journal of Empirical Finance* **29**: 287–303.

Valdez E. 2005. Tail conditional variance for elliptically contoured distributions. *Belgian Actuarial Bulletin* **5**: 26–36.

Wang CS, Zhao Z. 2016. Conditional value-at-risk: Semiparametric estimation and inference. *Journal of Econometrics* **195**: 86–103.

Weissman I. 1978. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association* **73**: 812–815.

Yamai Y, Yoshiba T. 2002. Comparative analyses of expected shortfall and value-at-risk: Their estimation error, decomposition, and optimization. *Monetary and Economic Studies* **20**: 87–121.

Appendix

Proof of Theorem 1: From Hill (2010, Thm. 2) we get

$$\frac{\sqrt{k_n}}{\sigma_{k_n}}\left(\widehat{\gamma}-\gamma\right) \underset{(n\to\infty)}{\overset{\mathcal{D}}{\longrightarrow}} \mathcal{N}(0,1), \tag{A.1}$$

where $\sigma_{k_n}^2 = \mathrm{E}[\sqrt{k_n}(\widehat{\gamma}-\gamma)]^2$. Note that Hill's (2010) Assumption B (required in his Thm. 2) can be seen to be implied by Assumption 1. Concretely, write (7) in terms of the slowly varying function $L(\cdot)$ from (5) to obtain

$$\lim_{x\to\infty}\frac{\frac{L(\lambda x)}{L(x)}-1}{A(x)} = \frac{\lambda^{\rho/\gamma}-1}{\gamma\rho},$$

where $A(\cdot)$ is a function with bounded increase due to $A(\cdot)\in RV_{\rho/\gamma}$ for $\rho/\gamma<0$ (de Haan and Ferreira, 2006, Thm. B.3.1). Also note that $\liminf_{n\to\infty}\sigma_{k_n}>0$ by arguments in Hill (2010, Sec. 3.2).

Hence, from (A.1) and arguments in the proof of Theorem 4.3.9 in de Haan and Ferreira (2006), we get

$$\frac{1}{\sigma_{k_n}}\frac{\sqrt{k_n}}{\log d_n}\left(\frac{\widehat{x}_{p_n}}{x_{p_n}}-1\right) \underset{(n\to\infty)}{\overset{\mathcal{D}}{\longrightarrow}} \mathcal{N}(0,1). \tag{A.2}$$

Here, we have also used that

$$\sqrt{k_n}\left(\frac{X_{(k_n+1)}}{U(n/k_n)}-1\right) = O_P(1)$$

from Hill (2010, Lem. 3) and the fact that $\log(x)\sim x-1$, as $x\to 1$. Next we show that

$$\frac{\sqrt{k_n}}{\log d_n}\left(\frac{\widehat{\mathrm{CTM}_a(p_n)}}{\mathrm{CTM}_a(p_n)}-1\right) = \frac{\sqrt{k_n}}{\log d_n}\left(\frac{\widehat{x}^a_{p_n}}{x^a_{p_n}}-1\right)+o_P(1). \tag{A.3}$$

To do so, expand

$$\frac{\sqrt{k_n}}{\log d_n}\left(\frac{\widehat{\mathrm{CTM}}_a(p_n)}{\mathrm{CTM}_a(p_n)}-1\right)=\frac{\sqrt{k_n}}{\log d_n}\left(\frac{\widehat{x}_{p_n}^a}{x_{p_n}^a}\cdot\frac{1-a\gamma}{1-a\widehat{\gamma}}\cdot\frac{\frac{x_{p_n}^a}{1-a\gamma}}{\mathrm{CTM}_a(p_n)}-1\right). \tag{A.4}$$

By (A.1),

$$\frac{1-a\gamma}{1-a\widehat{\gamma}}=1+\mathcal{O}_P(1/\sqrt{k_n}). \tag{A.5}$$

From Pan *et al.* (2013, Thm. 4.2),

$$\lim_{n\to\infty}\frac{1}{A\left(U(1/p_n)\right)}\left(\frac{\mathrm{CTM}_a(p_n)}{x_{p_n}^a}-\frac{1}{1-a\gamma}\right)=\frac{a}{(1/\gamma-a)(1/\gamma-a-\rho)}.$$

Due to $np_n=o(k_n)$ from (13) and monotonicity of $U(\cdot)$, we have $U(n/k_n)=\mathcal{O}\left(U(1/p_n)\right)$. Thus, with Assumption 1,

$$A\left(U(1/p_n)\right)=\mathcal{O}\left(A\left(U(n/k_n)\right)\right)=o(1/\sqrt{k_n}).$$

With the foregoing, this implies

$$\frac{\mathrm{CTM}_a(p_n)}{\frac{x_{p_n}^a}{1-a\gamma}}-1=o\left(\frac{1}{\sqrt{k_n}}\right). \tag{A.6}$$

Combining (A.4)–(A.6), (A.3) follows.

In view of (A.3) and $\left|\widehat{\sigma}_{k_n}^2-\sigma_{k_n}^2\right|=o_P(1)$ (Hill, 2010, Thm. 3), it suffices to prove the claim of the theorem for the sequence of random vectors

$$\frac{1}{\sigma_{k_n}}\frac{\sqrt{k_n}}{\log d_n}\left[\left(\frac{\widehat{x}_{p_n}^{a_j}}{x_{p_n}^{a_j}}-1\right)_{j=1,\ldots,J},\left(\frac{\widehat{x}_{p_n}}{x_{p_n}}-1\right)\right]'.$$

Let $b_1,\ldots,b_{J+1}\in\mathbb{R}$. Then, using a Cramér-Wold device, it suffices to consider

$$\frac{1}{\sigma_{k_n}}\frac{\sqrt{k_n}}{\log d_n}\sum_{j=1}^{J+1}b_j\left(\frac{\widehat{x}_{p_n}^{a_j}}{x_{p_n}^{a_j}}-1\right).$$

(Recall $a_{J+1}=1$.) Invoking a Skorohod construction (e.g., de Haan and Ferreira, 2006, Thm. A.0.1) similarly as in de Haan and Ferreira (2006, Example A.0.3), we may assume that the convergence in (A.2) holds almost surely (a.s.) on a different probability space:

$$\frac{1}{\sigma_{k_n}}\frac{\sqrt{k_n}}{\log d_n}\left(\frac{\widehat{x}_{p_n}}{x_{p_n}}-1\right)\xrightarrow[(n\to\infty)]{\text{a.s.}}Z\sim\mathcal{N}(0,1).$$

(Note the slight abuse of notation here.) A Taylor expansion of the functions $f_j(x)=x^{a_j}$ around 1

thus implies

$$\frac{1}{\sigma_{k_n}}\frac{\sqrt{k_n}}{\log d_n}\sum_{j=1}^{J+1} b_j\left(\frac{\widehat{x}_{p_n}^{a_j}}{x_{p_n}^{a_j}}-1\right)\xrightarrow[(n\to\infty)]{\text{a.s.}}\sum_{j=1}^{J+1} b_j a_j Z.$$

Going back to the original probability space, the conclusion follows. ■

Proof of Theorem 2: Since $\log(1+x)\sim x$ as $x\to 0$, it suffices to show

$$\sup_{t\in[\underline{t},\overline{t}]}\left|\frac{1}{\widehat{\sigma}_{k_n}}\frac{\sqrt{k_n}}{\log d_n(t)}\left(\frac{\widehat{x}_{p_n}(t)}{x_{p_n}(t)}-1\right)\right|\xrightarrow[(n\to\infty)]{\mathcal{D}}|Z|.$$

Due to $\widehat{x}_{p_n}(t)=\widehat{x}_{p_n}t^{-\widehat{\gamma}}$ and $\log d_n(t)/\log d_n=1+o(1)$ uniformly in $t\in[\underline{t},\overline{t}]$, we can expand

$$\frac{\sqrt{k_n}}{\log d_n(t)}\left(\frac{\widehat{x}_{p_n}(t)}{x_{p_n}(t)}-1\right)=(1+o(1))\frac{\sqrt{k_n}}{\log d_n}\left(\frac{\widehat{x}_{p_n}}{x_{p_n}}t^{\gamma-\widehat{\gamma}}\frac{x_{p_n}}{x_{p_n}(t)}t^{-\gamma}-1\right). \tag{A.7}$$

Apply the mean value theorem with $(\partial/\partial x)t^x=t^x\log(t)$ to derive $t^{\gamma-\widehat{\gamma}}=1+(\gamma-\widehat{\gamma})t^{\gamma+\nu(\widehat{\gamma}-\gamma)}\log(t)$ for any $t\in[\underline{t},\overline{t}]$ for some $\nu=\nu_t\in[0,1]$. Since $\widehat{\gamma}-\gamma=\mathcal{O}_P(1/\sqrt{k_n})$, this implies

$$t^{\gamma-\widehat{\gamma}}=1+\mathcal{O}_P(1/\sqrt{k_n})\qquad\text{uniformly in } t\in[\underline{t},\overline{t}]. \tag{A.8}$$

Writing (7) in terms of the quantile function $U(\cdot)$, we obtain from de Haan and Ferreira (2006, Thm. 2.3.9) that, uniformly in $t\in[\underline{t},\overline{t}]$,

$$\left|\frac{x_{p_n}(t)}{x_{p_n}}-t^{-\gamma}\right|=\left|\frac{U\left(1/(p_nt)\right)}{U\left(1/p_n\right)}-t^{-\gamma}\right|=\mathcal{O}\left(A(U(1/p_n))\right)=\mathcal{O}\left(A(U(n/k_n))\right)=o(1/\sqrt{k_n}). \tag{A.9}$$

Combining (A.7)–(A.9) gives

$$\frac{\sqrt{k_n}}{\log d_n(t)}\left(\frac{\widehat{x}_{p_n}(t)}{x_{p_n}(t)}-1\right)=\frac{\sqrt{k_n}}{\log d_n}\left(\frac{\widehat{x}_{p_n}}{x_{p_n}}-1\right)+o_P(1)\qquad\text{uniformly in } t\in[\underline{t},\overline{t}].$$

The conclusion now follows from Theorem 1. ■