Where does the tail begin? An approach based on scoring rules

Yannick Hoga*

September 17, 2019

Abstract

Learning about the tail shape of time series is important in, e.g., economics, finance and risk management. However, it is well known that estimates of the tail index can be very sensitive to the choice of the number k of tail observations used for estimation. We propose a procedure that determines where the tail begins by choosing k in a data-driven fashion using scoring rules. So far, scoring rules have mainly been used to compare density forecasts. We also demonstrate how our proposal can be used in multivariate applications in the system risk literature. The advantages of our choice of k are illustrated in simulations and an empirical application to Value-at-Risk forecasts for five U.S. blue-chip stocks.

Keywords: Hill estimator, Optimal Sample Fraction, Tail Index, Risk Management, Value-at-Risk

JEL classification: C13 (Estimation), C14 (Semiparametric and Nonparametric Methods), G32 (Financial Risk and Risk Management)

*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. R code and data files to reproduce the numerical and empirical results in this paper are available from https://www.oek.wiwi.uni-due.de/team/yannick-hoga/. The author is indebted to two anonymous referees for their detailed comments, that greatly improved the quality of the paper. Furthermore, the author would like to thank Christoph Hanck and Till Massing for carefully reading an early draft of this manuscript. This work was supported by the German Research Foundation (DFG) under Grant HO 6305/1-1.

1 Motivation

Extreme value theory (EVT) has numerous applications in economics, finance and risk management (Embrechts *et al.*, 1997; Longin, 2017). For instance, it can be used to extrapolate outside the range of available data to assess the likelihood of extreme events in financial markets (Novak and Beirlant, 2006). Gupta and Liang (2005) use EVT to examine the capital adequacy of hedge funds, that is naturally linked to extreme events. McNeil and Frey (2000) propose to use EVT for forecasting Value-at-Risk (VaR) and Expected Shortfall (ES), which are arguably the two most widely used risk measures in the financial industry. In a comparison of a wide range of forecasting procedures, Kuester *et al.* (2006) find extreme value-enhanced techniques to produce excellent VaR and ES predictions. Using EVT, Straetmans *et al.* (2008) investigate sectoral contagion risk in the US economy.

All the above mentioned applications require an estimate of the extreme value index. The *extreme value index* $\gamma > 0$ is the key parameter determining the rate of decay in the power law assumption for the tail, viz.

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

Slow variation means that $L(tx)/L(x) \underset{(x\to\infty)}{\longrightarrow} 1$ for all $t > 0$. Intuitively, $L(\cdot)$ behaves asymptotically like a function converging at infinity. Hence, the tail behavior of $1-F(x)$ in (1) is essentially determined by the power law decay of $x^{-1/\gamma}$, rendering γ the key tail parameter. The larger γ, the heavier the tail. The inverse of the extreme value index, $\alpha = 1/\gamma$, is called the *tail index*. In finance, tail index estimates of returns on speculative assets in developed markets often lie in the interval $(2, 4)$, implying finite variances and infinite 4th moments (Gabaix *et al.*, 2006).[1] Infinite variance returns may be observed in emerging markets for stock indices and exchange rates (Hill, 2015). In economics, Zipf's law for city size and firm size distributions implies $\alpha = 1$ (Gabaix, 1999, 2009).

Perhaps the most popular estimator of the extreme value index is due to Hill (1975). For observations $X_1 \ldots, X_n$ it is given by

$$\widehat{\gamma}(k) = \frac{1}{k} \sum_{i=1}^{k} \log(X_{(i)}/X_{(k+1)}), \tag{2}$$

where $X_{(1)} \geq \ldots \geq X_{(n)}$ denote the order statistics and $1 \leq k < n$; see Embrechts *et al.* (1997) for a derivation and an overview. To establish asymptotic properties of $\widehat{\gamma}(k)$, one typically requires that k be an *intermediate sequence*, i.e., $k = k(n) \to \infty$ and $k/n \to 0$, as $n \to \infty$. The first requirement ensures that estimation is ultimately based on an infinite number of observations. The second forces

[1] If a random variable $|X|$ has tail index α, then $\mathrm{E}\,|X|^{\underline{\alpha}} < \infty$ for $\underline{\alpha} < \alpha$ and $\mathrm{E}\,|X|^{\overline{\alpha}} = \infty$ for $\overline{\alpha} > \alpha$ (de Haan and Ferreira, 2006, Ex. 1.16).

the order statistics to lie in the 'tail'. In finite samples, these asymptotic requirements are of little help when choosing a particular value for k. So a practitioner might ask: Where does the tail begin? Or, put differently: How do we choose k to obtain a precise estimate of γ?

Generally, the choice of k involves a bias–variance trade-off. When choosing k too large, a bias in $\widehat{\gamma}(k)$ may appear, since non-tail observations are used. When choosing k too small, $\widehat{\gamma}(k)$ may have unnecessarily large variance, as too few observations are exploited in estimation.

Advantages of the Hill estimator include certain optimality properties (Csörgő *et al.*, 1985, Theorem 4), the available limit theory under very general conditions (Hill, 2010), and some theory-guided rules on the choice of k in practice (Drees and Kaufmann, 1998; Danielsson *et al.*, 2001). Because of these advantages and its widespread use in empirical work (Wagner and Marsh, 2005; Fagiolo *et al.*, 2008; Straetmans *et al.*, 2008; Trapani, 2016; Gu and Ibragimov, 2018; Sun and de Vries, 2018; Hoga, 2018a), we focus on the Hill estimator in the following. As we point out below, our proposal may be readily adapted to the choice of k for other estimators of γ, e.g., Peaks-over-Threshold estimators.

A drawback of the Hill estimator—but also of any other estimator of γ—is its sensitivity to the choice of the sample fraction, as evidenced by numerous Hill 'horror' plots; see, e.g., Embrechts *et al.* (1997, Fig. 4.1.13) or Resnick (2007, Fig. 4.2). We propose to choose k such that the resulting Hill estimate produces right-tail quantile estimates—via the Weissman (1978) estimator—that are in some sense adequate. To judge adequacy *in-sample*, we use scoring rules, previously used for *out-of-sample* density forecast evaluation (Gneiting and Raftery, 2007). Specifically, we use the quantile-weighted continuous-ranked probability score (QCRPS) of Gneiting and Ranjan (2011) that allows one to evaluate the density forecast in a particular region of interest. For us, the region of interest naturally is the tail. As is common in this literature, scores are negatively oriented, so that a lower score is preferred. Hence, we choose the k that produces the quantile estimates with the lowest QCRPS.

Extant routines to choose k can be classified into two groups—one containing heuristic approaches and the other containing theoretical approaches. One prominent heuristic approach is the automated Eye-Ball method (Resnick and Stărică, 1997), which identifies stable regions in the Hill plot $k \mapsto \widehat{\gamma}(k)$. Further suggestions include the quantile- and probability-driven methods of Daníelsson *et al.* (2016) and Gonzalo and Olmo (2004), respectively. These are based on comparing extreme value index-based semiparametric quantile and probability estimates with their nonparametric counterparts. The theoretical approaches—such as those of Drees and Kaufmann (1998) and Dacarogna *et al.* (2001)—are based on minimizing the theoretical asymptotic mean squared error of the extreme value index estimate.

Our proposal differs significantly from the above routines. Unlike the automated Eye-Ball method,

we choose k such that $\widehat{\gamma}(k)$ 'explains' the most extreme observations well in the sense of providing 'adequate' extreme quantile estimates. Unlike the quantile- and probability driven methods of Daníelsson *et al.* (2016) and Gonzalo and Olmo (2004), we do not rely on nonparametric (quantile and probability) estimates as a yardstick for a good tail fit. Such nonparametric estimates may be problematic because they are known to be highly unreliable in the tail. Instead, we use objective scoring rules to assess the quality of extreme quantile estimates. In our approach, the k that produces the 'best' (in a minimum score sense) extreme quantile estimates is chosen. Unlike the theoretical approaches, our choice of k is data-driven and not based on asymptotic arguments, which are only valid for specific tail index estimators and independent and identically distributed (i.i.d.) data.

The remainder of the paper proceeds as follows. In Section 2, we present our method of choosing k. Subsection 2.1 introduces some notation and the leading competitor of our choice of k, which is due to Daníelsson *et al.* (2016). Then, Subsection 2.2 presents our scoring rule-based approach to the choice of k. We also discuss possible applications of our proposal in the context of estimating co-crash probabilities, relevant in the systemic risk literature, in Subsection 2.3. Section 3 illustrates the good finite-sample properties of our choice of k and compares it with that of Daníelsson *et al.* (2016). An empirical application to returns on a diverse set of U.S. blue chip stocks in Section 4 demonstrates the advantages of our approach in VaR forecasting. The final section concludes.

2 Choosing k

2.1 Preliminaries

We consider $X_1, \ldots, X_n$ with common distribution function (d.f.) $F(\cdot)$. The left-continuous inverse $F^{\leftarrow}(p) = \inf\{x \in \mathbb{R} \mid F(x) \geq p\}$ denotes the p-quantile. Assume $F(\cdot)$ satisfies (1) or, equivalently,

$$\lim_{t\to\infty} \frac{U(ty)}{U(t)} = y^{\gamma} \qquad \text{for all } y > 0, \tag{3}$$

where $U(y) = F^{\leftarrow}(1 - 1/y)$ denotes the $(1 - 1/y)$-quantile and $\gamma > 0$ is again the extreme value index.[2] This semi-parametric assumption on the (right) tail is satisfied for a wide range of parametric distributions—such as Student's t-, Burr-, or Pareto-distribution (Hua and Joe, 2011)—and also for the stationary distribution of GARCH models (Davis and Mikosch, 1998; Mikosch and Stărică, 2000). Assumption (3) is a plausible modelling assumption in numerous disciplines, e.g., finance, insurance and teletraffic modelling.

[2]Since $L(\cdot)$ is slowly varying, (1) is equivalent to $\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-1/\gamma}$ for all $x > 0$. From this and the fact that for non-decreasing functions $f_n(\cdot)$ and $g(\cdot)$ the relation $\lim_{n\to\infty} f_n(x) = g(x)$ implies $\lim_{n\to\infty} f_n^{\leftarrow}(x) = g^{\leftarrow}(x)$ under some regularity conditions, the equivalence of (1) and (3) may be shown; see the proof of Theorem 1.2.1.1 in de Haan and Ferreira (2006) for details.

An estimate of γ—for instance the Hill (1975) estimate $\widehat{\gamma}(k)$—is crucial in estimating other tail-related quantities. Suppose, for instance, one has to estimate the (right-tail) p-quantile $F^{\leftarrow}(p)$ from $X_1, \ldots, X_n$ for some p close to 1. This need arises, for instance, in financial risk management, where the popular *Value-at-Risk (VaR) at level p* of some portfolio is simply the p-quantile of its profit & loss (P&L). Assume that p is extreme in the sense that $p \gg 1 - k/n$. Then, an estimate of the p-quantile $F^{\leftarrow}(p)$ can be motivated from (3) as follows. Inserting $t = n/k$ and $y = k/(n[1-p])$ in (3) gives

$$F^{\leftarrow}(p) = U(1/[1-p]) \approx U(n/k) \left(\frac{k}{n[1-p]}\right)^{\gamma} \approx X_{(k+1)} \left(\frac{k}{n[1-p]}\right)^{\widehat{\gamma}(k)} =: \widehat{x}_p(k). \tag{4}$$

The idea of the Weissman (1978) estimator $\widehat{x}_p(k)$ is to estimate the extreme (right-tail) p-quantile of interest by first estimating the less extreme, and hence more easily estimated, $(1 - k/n)$-quantile (via $X_{(k+1)}$) and then to use some extrapolation exploiting the tail shape in (3) to the desired level (via $\big(k/(n[1-p])\big)^{\widehat{\gamma}(k)}$). This estimator is semi-parametric in the sense that it exploits the semi-parametric assumption (1), which specifies the tail decay via the parameter γ in $x^{-1/\gamma}$, but otherwise leaves the precise form of the tail unspecified, since $L(\cdot)$ is merely assumed to be slowly varying.

The estimator $\widehat{x}_p(k)$ has several advantages over completely non-parametric estimators of the p-quantile, such as a (possibly kernel-smoothed) order statistic $X_{(\lfloor n(1-p) \rfloor + 1)}$; see Chen and Tang (2005) for the corresponding limit theory. First, it can be more precise than non-parametric alternatives for levels p of practical interest in risk management. Hoga (2019b+) shows that the advantage of EVT-based estimators over non-parametric estimators is the larger, the heavier the tail and/or the more extreme the probability level p and/or the more observations are available. Second, $\widehat{x}_p(k)$ allows for extrapolation outside the range of available observations, i.e., it produces consistent estimates even for $p > 1 - 1/n$, where $X_{(\lfloor n(1-p) \rfloor + 1)}$ fails (Drees, 2003).

Remark 1. The Weissman (1978) estimator is designed to estimate right-tail quantiles. However, in many applications in finance, the lower tail of some variable X is of more interest. In this case, one can simply consider the negated variable $-X$ and proceed as outlined above.

Daníelsson *et al.* (2016) propose to choose k in $\widehat{\gamma}(k)$, such that the semi-parametric estimate $\widehat{x}_p(k)$ of the p-quantile is 'close' (here, in terms of the sup-norm) to the non-parametric estimate $X_{(\lfloor n(1-p) \rfloor + 1)}$ over a range of large values of p (here, $p = 1 - 1/n, \ldots, 1 - k_{\max}/n$). More precisely, k is chosen data-adaptively as

$$k^Q := \underset{k=1,\ldots,k_{\max}}{\arg\min} \left[\sup_{p=1-1/n,\ldots,1-k_{\max}/n} \left| X_{(\lfloor n(1-p) \rfloor + 1)} - \widehat{x}_p(k) \right| \right],$$

where $k_{\max}$ is the maximal value of k one is willing to entertain. Daníelsson *et al.* (2016) show that

k^Q is robust to the choice of $k_{\max}$. They also demonstrate that k^Q compares favorably with the other heuristic and theoretical choices mentioned in the Motivation. For this reason, and to keep the present paper concise, we focus on k^Q as the main competitor of our proposal in the remainder of this paper.

A drawback of k^Q is that there is some arbitrariness in the choice of the sup-metric to measure the difference between $X_{(\lfloor n(1-p)\rfloor+1)}$ and $\widehat{x}_p(k)$. While Daníelsson *et al.* (2016) show in simulations that this metric works particularly well, other possibilities include the mean absolute deviation and the mean squared error. Furthermore, and more importantly, the choice k^Q is designed to produce a high agreement between $\widehat{x}_p(k)$ and the order statistics $X_{(\lfloor n(1-p)\rfloor+1)}$, where the latter are, however, well-known to be highly noisy in the tail region of interest. This in turn may produce some undesirable volatility in the choice of k.

Closely related to the quantile-driven idea of Daníelsson *et al.* (2016) is the probability-driven idea of Gonzalo and Olmo (2004). It is well-known (e.g., Embrechts *et al.*, 1997, Theorem 3.4.13 (b)) that under (1) the conditional excess distribution $(X-u) \mid X > u$ has an approximate generalized Pareto distribution (GPD), $\mathrm{GP}(\sigma, \xi)$, with parameters $\sigma > 0$ and $\xi = \gamma$ for large enough u. Gonzalo and Olmo (2004) suggest to choose k in $u = X_{(k)}$, such that the empirical conditional excess distribution of $(X-u) \mid X > u$ is close in sup-distance to the GPD fitted (via maximum likelihood) to the excesses above u, i.e., given by $X_{(1)} - u, \ldots, X_{(k)} - u$.

2.2 Choosing k via scoring rules

Our proposal addresses the above mentioned two drawbacks of k^Q by evaluating $\widehat{x}_p(k)$ using *scoring rules*. These are primarily used in the evaluation of density forecasts. By choosing a *proper* scoring rule, the arbitrariness in the choice of the metric is avoided. Furthermore, we do not simply rely on order statistics as the yardstick by which to measure adequacy of the choice of k.

To describe our choice of k, we proceed in two steps. First, let $X_1, \ldots, X_n$ be drawn from a common d.f. $F(\cdot)$ with finite first moment and let X denote a generic element of the sequence. Furthermore, denote by $I_{\{\cdot\}}$ the indicator function, which equals 1 if the event in brackets is true and 0 otherwise. To estimate $F^{\leftarrow}(p)$ via $\widehat{x}_p(k)$, the integer k has to be chosen. It is well known that the quantile score

$$\mathrm{QS}_p(F^{\leftarrow}(p), x) = 2\left(I_{\{y \le F^{\leftarrow}(p)\}} - p\right)\left(F^{\leftarrow}(p) - x\right)$$

is a strictly consistent scoring function for the p-quantile $F^{\leftarrow}(p)$ of any distribution with finite first moment (Gneiting, 2011, Thm. 9). This implies that the expected score $\mathrm{E}_F[\mathrm{QS}_p(x, X)]$ assumes its unique minimum at $x = F^{\leftarrow}(p)$. Thus, the choice of k in the estimate $\widehat{x}_p(k)$ of $F^{\leftarrow}(p)$ should render $\mathrm{E}_F[\mathrm{QS}_p(\widehat{x}_p(k), X)]$ small. Replacing the 'population' quantity $\mathrm{E}_F[\mathrm{QS}_p(\widehat{x}_p(k), X)]$ with its sample

counterpart, a reasonable choice of k minimizes

$$\overline{S}^{(k)}_{p,\mathrm{QS}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{QS}_p(\widehat{x}_p(k), X_i) \tag{5}$$

for some fixed $p \gg 1 - k/n$.[3]

Remark 2. One of the main uses of strictly consistent scoring functions is to compare different point forecasts (Gneiting, 2011). For instance, in the context of quantile estimation, strict consistency of the quantile score allows for sensible comparisons of L different sets of point forecasts $\{f_i^{(\ell)}\}_{i=1,\ldots,n}$ ($\ell = 1, \ldots, L$) of the p-quantile via the average score $\frac{1}{n}\sum_{i=1}^{n} \mathrm{QS}_p(f_i^{(\ell)}, x_i)$, where the $\{x_i\}_{i=1,\ldots,n}$ are verifying observations. Since scoring functions are negatively oriented, the set of forecasts with the lowest average score is preferred.

Remark 3. We are by no means the first to use scoring functions—primarily developed for *out-of-sample* forecast comparisons—for *in-sample* purposes; here, the choice of k. Other in-sample uses of consistent scoring functions include M-estimation (Huber and Ronchetti, 2009) and quantile regression (Koenker and Bassett, 1978). For instance, the basic assumption in quantile regression is that the p-th conditional quantile of some variable Y given X is linear, i.e., $X\beta_p$. Then, the quantile regression estimate of β_p is

$$\widehat{\beta}_p = \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} \mathrm{QS}_p(X_i\beta, Y_i),$$

where Y_i and X_i are drawn from Y and X, respectively.

Choosing a specific value of p in (5) involves some subjectivity and, at any rate, the $\widehat{\gamma}(k)$ with optimal k should produce 'good' Weissman estimates $\widehat{x}_p(k)$ for a range of values of p—not just one. So in a second step, we evaluate $\widehat{x}_p(k)$ over some interval $p \in (\beta, 1)$ for different choices of k. It appears intuitive to simply average the quantile scores QS_p over p and choose k to minimize

$$\begin{aligned}\overline{S}^{(k)}_{\mathrm{QCRPS}} &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1-\beta}\int_{\beta}^{1} \mathrm{QS}_p(\widehat{x}_p(k), X_i), \qquad \beta \in (0,1) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{QCRPS}_{\beta}(\widehat{x}_p(k), X_i).\end{aligned}$$

Here, $\mathrm{QCRPS}_{\beta}(\cdot,\cdot) = 1/(1-\beta)\int_{\beta}^{1} \mathrm{QS}_p(\cdot,\cdot)$ is the quantile-weighted continuous-ranked probability score (QCRPS) of Gneiting and Ranjan (2011). The specific form of the QCRPS has been used by Holzmann and Klar (2017). Following Daníelsson *et al.* (2016), we take $\beta = 1 - k_{\max}/n$ for some $k_{\max}$ to be specified below.

[3]Recall that $p \gg 1 - k/n$ is required for $\widehat{x}_p(k)$ in (4) to make sense.

The QCRPS was initially developed as a scoring *rule* for evaluating *density* forecasts in specific regions of interest (Gneiting and Ranjan, 2011). Note that scoring *functions*, on the other hand, are used for evaluating *point* forecasts. Generally, a scoring rule is a loss function $S(F^{\leftarrow}(\cdot), x)$ taking as arguments the quantile function forecast $F^{\leftarrow}(\cdot)$ (which implicitly defines the density forecast $f(\cdot)$) and the verifying realization x of X. In analogy with strictly consistent scoring functions, the scoring rule is said to be *proper* if the expected score under this distribution, $\mathrm{E}_F[S(H^{\leftarrow}(\cdot), X)]$ assumes its (not necessarily unique) minimum at $H^{\leftarrow}(\cdot) \equiv F^{\leftarrow}(\cdot)$. According to Gneiting and Ranjan (2011), the QCRPS is a proper scoring rule. Hence, a sensible choice of k to minimizes $\mathrm{E}_F[\mathrm{QCRPS}(\widehat{x}_p(k), X)]$. Thus, propriety provides the theoretical foundation of the heuristic to choose k to minimize $\overline{S}^{(k)}_{\mathrm{QCRPS}}$, i.e., the sample counterpart of $\mathrm{E}_F[\mathrm{QCRPS}(\widehat{x}_p(k), X)]$.

Remark 4. As pointed out above, the primary use of the QCRPS is in forecast evaluation. The QCRPS assesses the density forecast in the tail based on verifying realizations (here, $x_1, \ldots, x_n$). Propriety of the scoring rule QCRPS_β again allows for valid comparisons via the average scores

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{QCRPS}_\beta(\widehat{F}_i^{(\ell),\leftarrow}, x_i), \qquad \ell = 1, \ldots, L,$$

where $\widehat{F}_i^{(\ell),\leftarrow}(p)$ for $p \in (\beta, 1)$ define the quantile function forecasts. By propriety, the forecast with the lowest score is preferred.

In practice, closed-form expressions for QCRPS may not be available. Yet, the integral can be approximated to any degree of accuracy. We discretize it at the points $p = 1-1/n, \ldots, 1-k_{\max}/n = \beta$ to obtain

$$\widetilde{S}^{(k)}_{\mathrm{QCRPS}} = \frac{2}{nk_{\max}} \sum_{i=1}^{n} \sum_{p=1-1/n}^{1-k_{\max}/n} \left[I_{\{X_i \leq \widehat{x}_p(k)\}} - p \right] \left[\widehat{x}_p(k) - X_i \right], \tag{6}$$

where $\widehat{x}_p(k) = X_{(k+1)} \left(k/\{n[1-p]\} \right)^{\widehat{\gamma}(k)}$ by (4). Finally, this leads to our choice

$$k^{\mathrm{QCRPS}} := \underset{k=1,\ldots,k_{\max}}{\arg\min}\ \widetilde{S}^{(k)}_{\mathrm{QCRPS}}. \tag{7}$$

Just like k^{Q}, this is a data-adaptive choice. The idea of k^{QCRPS} is to choose k, such that the quantile estimates minimize the QCRPS. Hence, the choice of k^{QCRPS} is tailored not only for tail index estimation via the Hill (1975) estimator, but also for subsequent extreme quantile estimation via the Weissman (1978) estimator.

Remark 5. (a) Our approach can also be used to choose k for other estimators than Hill's, e.g., the Pickands (1975) estimator or its refinement by Drees (1995). To that end, simply replace

the Hill estimate in $\widehat{x}_p(k)$ from (6) with the desired estimator and proceed as before.

(b) The choice of k also appears in Peaks over Threshold (POT) estimation, where a generalized Pareto distribution (GPD) is fitted to exceedances above a high threshold, typically given by an order statistic $X_{(k+1)}$ (McNeil and Frey, 2000, Sec. 2.2). Again, the choice of k involves a bias–variance tradeoff: Choosing k too small increases the variance, while choosing k too large, the GPD approximation may not be valid, introducing bias.

Estimates of the p-quantile, say, $\widetilde{x}_p^{\mathrm{POT}}(k)$, are then obtained via the quantiles of the fitted GPD; see McNeil and Frey (2000, Eqn. (10)) for the formula. For this to make sense, we require $p \geq 1 - k/n$, because the GPD approximation is only valid above $X_{(k+1)}$, which estimates the $(1 - k/n)$-quantile. Hence, POT quantile estimates can be obtained as

$$\widehat{x}_p^{\mathrm{POT}}(k) = \begin{cases} \widetilde{x}_p^{\mathrm{POT}}(k) & \text{for } p \geq 1 - k/n, \\ X_{(\lfloor n(1-p) \rfloor + 1)} & \text{for } p < 1 - k/n. \end{cases}$$

Now, replacing $\widehat{x}_p(k)$ in (6) with the POT quantile estimates $\widehat{x}_p^{\mathrm{POT}}(k)$, our method can also be used to choose k in POT estimation of the GPD parameters.

Remark 6. Clearly, k^Q and k^{QCRPS} are non-deterministic sequences. However, limit theory for $\widehat{\gamma}(k)$ (and also limit theory for $\widehat{x}_p(k)$) typically relies on a *deterministic* intermediate sequence k. To the best of our knowledge, only Drees *et al.* (2018) derive asymptotic properties of $\widehat{\gamma}(\widetilde{k}^Q)$ for a stochastic sequence $\widetilde{k}^Q$. The data-driven stochastic sequence $\widetilde{k}^Q$ is due to Clauset *et al.* (2009) and is closely related to k^Q. Drees *et al.* (2018) show that for i.i.d. data with exact Pareto distribution, the usual asymptotic normality of $\widehat{\gamma}(\widetilde{k}^Q)$ no longer holds. Deriving a similar result for $\widehat{\gamma}(k^{\mathrm{QCRPS}})$ for possibly dependent and non-Pareto data is likely to be difficult and is thus left for future research. We provide some simulation evidence on the asymptotic normality of $\widehat{\gamma}(k^{\mathrm{QCRPS}})$ in Subsection 3.4.

2.3 Extensions to measures of systemic risk

So far we discussed the choice of k in a univariate setting. Yet, extreme value methods have also become popular in the analysis of multivariate data, particularly in the systemic risk literature (Poon *et al.*, 2004; Hartmann *et al.*, 2006; Bosma *et al.*, 2019; Nolde and Zhang, 2019+). In this strand of the literature, significant interest attaches to the estimation of co-crash probabilities of the form

$$\begin{aligned} \tau_p &= P\{X > F_X^{\leftarrow}(p) \mid Y > F_Y^{\leftarrow}(p)\} = \frac{P\{X > F_X^{\leftarrow}(p), Y > F_Y^{\leftarrow}(p)\}}{P\{Y > F_Y^{\leftarrow}(p)\}} \\ &= \frac{P\{X > F_X^{\leftarrow}(p), Y > F_Y^{\leftarrow}(p)\}}{1 - p}, \end{aligned} \tag{8}$$

where X and Y denote returns on two risky assets with continuous d.f.s $F_X(\cdot)$ and $F_Y(\cdot)$, respectively. Here, p is typically close to 1 in order for the events $X > F_X^{\leftarrow}(p)$ and $Y > F_Y^{\leftarrow}(p)$ to be interpretable as crashes. If Y denotes the returns on a market portfolio (e.g., the S&P 500), then this co-crash probability can be seen as an extension of the CAPM-β. Hence, τ_p is often referred to as a tail-β (Straetmans *et al.*, 2008). If X and Y denote the returns on, say, two bank stocks, τ_p can be used as a measure for contagion risk (Straetmans and Chaudhry, 2015). Finally, τ_p may be applied within banks for stress testing purposes. In this case, X represents the returns on the managed portfolio and $Y > F_Y^{\leftarrow}(p)$ the stress event.

Remark 7. The limit $\chi = \lim_{p\uparrow 1} \tau_p$ is the well-known tail dependence coefficient, which dates back to Sibuya (1960), but continues to be actively studied (Bücher *et al.*, 2015; Hoga, 2018b). Poon *et al.* (2004) consider χ 'a true measure of systemic risk in international stock markets'. The scalar χ may also be interpreted in terms of the copula $C(\cdot,\cdot)$, i.e., the unique function satisfying

$$P\{X \leq x, Y \leq y\} = C(F_X(x), F_Y(y)), \qquad x, y \in \mathbb{R}.$$

Then, $\chi = \lim_{u\downarrow 0} \widehat{C}(u,u)/u$ may be viewed as a directional derivative of the survival copula $\widehat{C}(u,v) = u + v - 1 + C(1-u, 1-v)$ $(u, v \in [0,1])$ at the origin. We refer to Embrechts *et al.* (2003) and Dey and Yan (2016) for more on copulas and tail dependence, including their practical applications.

We now discuss how our proposal to choose k can also be applied in the estimation of τ_p. To estimate τ_p from a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, we proceed in three steps. First, we reduce the estimation problem to one dimension. To do so, transform X and Y to unit Pareto marginals via $\widetilde{X} = 1/[1 - F_X(X)]$ and $\widetilde{Y} = 1/[1 - F_Y(Y)]$. In practice, we use

$$\widehat{X}_i = \frac{1}{1 - \widehat{F}_X(X_i)}, \qquad \text{where} \quad \widehat{F}_X(x) = \frac{1}{n+1} \sum_{i=1}^{n+1} I_{\{X_i \leq x\}}.$$

We divide by $(n+1)$ in $\widehat{F}_X$ to avoid division by zero in $\widehat{X}_i$. The $\widehat{Y}_i$ are calculated analogously. The Pareto transformation allows us to write

$$\begin{aligned} P\{X > F_X^{\leftarrow}(p), Y > F_Y^{\leftarrow}(p)\} &= P\{\widetilde{X} > 1/(1-p), \widetilde{Y} > 1/(1-p)\} \\ &= P\{\min(\widetilde{X}, \widetilde{Y}) > 1/(1-p)\} = P\{Z > s\}, \end{aligned}$$

where $Z = \min(\widetilde{X}, \widetilde{Y})$ and $s = 1/(1-p)$. Thus, we have reduced the estimation of $\tau_p = P\{Z > s\}/(1-p)$ to one dimension.

In the second step, we introduce an extreme value-type assumption for the distribution of Z, that allows us to extrapolate outside the range of available observations. Going back to Ledford and Tawn

(1996, 1997), a widely applicable semi-parametric model for the joint tail of (X, Y) is

$$P\{Z > s\} = L_Z(s)s^{-1/\gamma_Z}, \qquad \gamma_Z \leq 1, \tag{9}$$

where $L_Z(\cdot)$ again denotes a slowly varying function. Heffernan (2000) gives an extensive list of bivariate distributions satisfying (9). The parameter γ_Z measures the amount of extremal dependence in (X, Y). The smaller (larger) γ_Z, the lighter (heavier) the tail of Z and, hence, the smaller (larger) the probability of joint extremes of X and Y.

Finally, we derive an estimator of $P\{Z > s\}$ based on (9). By inverting the steps leading to the Weissman (1978) estimator in (4), $p_s = P\{Z > s\}$ can be estimated for large s via

$$\widehat{p}_s = \frac{k}{n}\left(\frac{\widehat{Z}_{(k+1)}}{s}\right)^{1/\widehat{\gamma}_Z(k)},$$

where $\widehat{Z}_{(k+1)}$ is the k-th largest value of $\widehat{Z}_i = \min(\widehat{X}_i, \widehat{Y}_i)$, $i = 1, \ldots, n$, and $\widehat{\gamma}_Z(k)$ is the Hill estimator based on $\widehat{Z}_{(1)}, \ldots, \widehat{Z}_{(k+1)}$. From this and (8), we get

$$\widehat{\tau}_p = \frac{k}{n}s\left(\frac{\widehat{Z}_{(k+1)}}{s}\right)^{1/\widehat{\gamma}_Z(k)}.$$

This estimator has been used in applied work by, e.g., Hartmann *et al.* (2006), Straetmans *et al.* (2008) and Straetmans and Chaudhry (2015).

The choice of k in $\widehat{\gamma}_Z(k)$ is again crucial in calculating $\widehat{\tau}_p$. In principle, any of the methods to choose k discussed above can be used here as well. For instance, Hartmann *et al.* (2006, p. 175) use Hill plots, and Straetmans *et al.* (2008, p. 22) rely on the exponential regression algorithm of Beirlant *et al.* (1999). Alternatively, we suggest to choose k as the argument that minimizes $\widetilde{S}^{(k)}_{\text{QCRPS}}$. Of course, the variables X_i in the definition of $\widetilde{S}^{(k)}_{\text{QCRPS}}$ in (6) need to be replaced by $\widehat{Z}_i$. Thus, our procedure can be applied not only in a univariate setting, but also more broadly in the estimation of contagion risk measures.

3 Monte Carlo Simulations

3.1 Simulation setup

We compare the two choices of $k \in \{k^Q, k^{\text{QCRPS}}\}$ regarding their ability to produce accurate estimates $\widehat{\gamma}(k)$ of γ (Subsection 3.2) and accurate extreme quantile estimates $\widehat{x}_p(k)$ over the range $p = 0.90, \ldots, 0.9999$ (Subsection 3.3). As mentioned above, we confine ourselves to this comparison for brevity and, more importantly, because Daníelsson *et al.* (2016) show in extensive simulations that

k^Q works well compared to the most popular existing alternatives.

In our simulations we consider all data-generating processes (DGPs) used by Daníelsson *et al.* (2016) and some additional DGPs. We use the following models for i.i.d. data $X_1, \ldots, X_n$:

(M1) t_α distribution with $\alpha = 1, 3, 5, 7$ degrees of freedom;

(M2) Fréchet(α) distribution with d.f. $F(x) = \exp(-x^{-\alpha})$, $x > 0$, $\alpha = 1, 3, 5, 7$;

(M3) Symmetric stable (SS) distribution with index parameter $\alpha = 0.5, 1.0, 1.5, 1.9$;

(M4) Pareto distribution, Pa(α), with $1 - F(x) = x^{-\alpha}$ for $x > 1$ and $\alpha = 1, 3, 5, 7$;

(M5) Burr(τ, λ) distribution of type XII, $1 - F(x) = 1/(1 + x^\tau)^\lambda$ for $x > 0$, with $\tau = 2$, $\lambda = 1/2, 3/2, 5/2, 7/2$.

The parameter α in **(M1)**–**(M4)** indicates the tail index, and for **(M5)**, the tail index is $\alpha = \lambda\tau$. For more detail on the above distributions, we refer to Table 2.1 in Beirlant *et al.* (2004).

For dependent data we use (G)ARCH models $\{X_i = \sigma_i \varepsilon_i\}_{i=1,\ldots,n}$ with $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $\sigma_i^2 = \omega^\circ + \alpha^\circ X_{i-1}^2 + \beta^\circ \sigma_{i-1}^2$. We set:

(M6) ARCH: $\omega^\circ = 10^{-6}$, $\alpha^\circ = 0.6, 0.7, 0.8, 0.9$ and $\beta^\circ = 0$, leading to tail indices of the stationary distributions of $\alpha = 3.82, 3.17, 2.68, 2.30$, respectively;[4]

(M7) GARCH: $\omega^\circ = 10^{-6}$, $\alpha^\circ = 0.4$ and $\beta^\circ = 0.27, 0.43, 0.53, 0.599$, leading to tail indices of the stationary distributions of $\alpha = 4.99, 3.96, 2.98, 2.03$, respectively.

Finally, we consider GARCH-filtered residuals for tail estimation.

(M8) Filtered GARCH: Consider $\{\widehat{\varepsilon}_i = X_i/\widehat{\sigma}_i\}$, were X_i is generated according to the GARCH model in **(M7)** with parameters $\omega^\circ = 10^{-6}$, $\alpha^\circ = 0.4$, $\beta^\circ = 0.53$ and (standardized) t_α-distributed errors ε_i with $\alpha = 5, 7$. Volatility estimates $\widehat{\sigma}_i^2 = \widehat{\omega^\circ} + \widehat{\alpha^\circ} X_{i-1}^2 + \widehat{\beta^\circ}\widehat{\sigma}_{i-1}^2$ are based on quasi-maximum likelihood (QML) estimates $(\widehat{\omega^\circ}, \widehat{\alpha^\circ}, \widehat{\beta^\circ})$ of the GARCH parameters.

The distributions in **(M1)**–**(M3)** and **(M6)** are those used by Daníelsson *et al.* (2016). We also consider a GARCH model, because it is empirically more relevant than the simpler ARCH specification in **(M6)**. The parameters in **(M6)** and **(M7)** are chosen such that the implied tail indices are roughly between 2 and 4, which is the range of implied tail indices of fitted GARCH models in practice; see,

[4]Mikosch and Stărică (2000, Thm. 2.1) show that the true tail indices of the (G)ARCH models can be computed as the unique positive solution $\alpha > 0$ of $\mathrm{E}[\alpha^\circ \varepsilon_1^2 + \beta^\circ]^{\alpha/2} = 1$. We have obtained the true tail indices of models **(M6)** and **(M7)** by solving this equation for α.

e.g., Mikosch and Stărică (2000) for foreign exchange returns. We use model **(M8)**, because it is often of interest in risk management to estimate the tail index of GARCH-filtered residuals; see, e.g., Chan *et al.* (2007) and Hoga (2019a+). Unlike in **(M1)**, we do not consider degrees of freedom $\alpha = 1, 3$ for the t-distributed errors in **(M8)**, since finite fourth moments of the errors are required for the asymptotic normality of QML estimates in GARCH models (Francq and Zakoïan, 2010).

We use sample sizes of $n = 500, 1000, 2000, 5000$ that are typically used in practice, and choose $k_{\max} = \lfloor n^{0.6} \rfloor$ in calculating k^Q and k^{QCRPS}. This choice leads to almost uniformly better results across all distributions and sample sizes considered here than, e.g., $k_{\max} = \lfloor n^{0.5} \rfloor$ or $k_{\max} = \lfloor n^{0.7} \rfloor$. Furthermore, since $k^{\mathrm{QCRPS}} \leq k_{\max} = \lfloor n^{0.6} \rfloor$, the theoretical requirement that $k^{\mathrm{QCRPS}}/n \to 0$ is automatically satisfied. All results in this section are based on $R = 10000$ replications.

3.2 Estimation of γ

For the DGPs in **(M1)**–**(M8)**, Tables 1 and 2 display the simulation root mean squared error (RMSE) and the bias of $\widehat{\gamma}(k)$ with $k \in \{k^Q, k^{\mathrm{QCRPS}}\}$. The RMSE is calculated as $\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\gamma}^{(r)}(k^{(r)}) - \gamma)^2}$, and the bias as $\frac{1}{R}\sum_{r=1}^{R}\widehat{\gamma}^{(r)}(k^{(r)}) - \gamma$ for R simulation runs of $\widehat{\gamma}^{(r)}(k^{(r)})$. Table 3 shows the average values of the chosen $k \in \{k^Q, k^{\mathrm{QCRPS}}\}$ over all R replications and Table 4 shows their standard errors.

We draw the following conclusions from Tables 1–4:

1. RMSE: We make two observations. First, as expected, RMSEs tend to be lower for larger sample sizes for both choices of k. The reduction in RMSEs is particularly marked for k^{QCRPS}, whereas the improvements are relatively small for k^Q. For instance, for the Fréchet($\alpha = 3$) distribution, the RMSE using k^{QCRPS} drops in half from 5.6 ($n = 500$) to 2.8 ($n = 5000$). For k^Q the RMSE only decreases from 11.5 to 10.4. Thus, the relative advantage of k^{QCRPS} increases with the sample size.

   This may be explained by the small increase in the average number of order statistics k^Q when n increases. For instance, as Table 3 shows for the Fréchet($\alpha = 3$) distribution, the effective (average) sample size using k^Q increases by a factor of 2.5 from 13 ($n = 500$) to 32 ($n = 5000$). Yet, using k^{QCRPS} it jumps from 28 to 107, representing a 3.8-fold increase.

   Second, RMSEs of estimates $\widehat{\gamma}(k)$ ($k \in \{k^Q, k^{\mathrm{QCRPS}}\}$) tend to improve, the lighter the tail, i.e., the larger α. This is as expected, because the theoretical asymptotic variance of $\widehat{\gamma}(k)$ is $1/\alpha^2$ (Beirlant *et al.*, 2004, p. 111). However, this general tendency is sometimes reversed for k^{QCRPS}, most notably for the t_α-distribution. Table 2 suggests that this may be due to the bias increasing with α.

Model	α	$n = 500$		$n = 1000$		$n = 2000$		$n = 5000$	
		k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}
t_α	1	79.4	**19.0**	76.3	**16.2**	73.6	**13.0**	69.6	**9.9**
	3	12.3	**10.3**	10.8	**8.2**	10.4	**6.7**	10.6	**5.1**
	5	**8.5**	13.3	**7.2**	11.4	**6.0**	9.7	**5.4**	7.8
	7	**9.6**	15.6	**7.7**	13.5	**6.3**	11.7	**5.1**	9.7
Fréchet	1	79.6	**19.3**	74.6	**15.9**	69.9	**13.1**	69.8	**10.1**
	3	11.5	**5.6**	10.9	**4.6**	10.9	**3.7**	10.4	**2.8**
	5	5.8	**3.4**	5.6	**2.7**	5.2	**2.2**	5.1	**1.7**
	7	3.8	**2.4**	3.6	**1.9**	3.4	**1.6**	3.3	**1.2**
SS	0.5	172	**60.5**	177	**58.0**	175	**55.0**	167	**50.0**
	1.0	79.8	**19.5**	75.4	**15.8**	71.8	**13.0**	63.1	**10.1**
	1.5	34.4	**14.4**	31.9	**12.3**	30.9	**10.2**	31.3	**7.8**
	1.9	29.5	**24.4**	28.3	**25.2**	26.0	**25.7**	**23.3**	25.3
Pareto	1	79.8	**19.2**	72.9	**15.8**	70.2	**13.0**	69.1	**10.0**
	3	11.0	**5.7**	10.9	**4.6**	10.5	**3.8**	10.7	**2.8**
	5	5.8	**3.3**	5.3	**2.7**	5.2	**2.2**	5.0	**1.7**
	7	3.8	**2.4**	3.6	**1.9**	3.4	**1.5**	3.2	**1.2**
Burr	1	79.8	**19.0**	71.6	**15.9**	73.7	**12.9**	70.9	**9.9**
	3	11.1	**6.6**	10.8	**5.4**	10.9	**4.3**	10.3	**3.3**
	5	**6.5**	7.1	**5.6**	6.1	5.4	**5.2**	5.1	**4.3**
	7	**6.0**	8.3	**5.1**	7.3	**4.4**	6.5	**3.8**	5.5
ARCH	2.30	17.3	**12.0**	16.9	**10.1**	16.7	**8.4**	16.4	**6.3**
	2.68	14.0	**10.5**	14.0	**8.8**	14.2	**7.0**	13.8	**5.4**
	3.17	11.9	**10.0**	11.6	**7.9**	11.7	**6.6**	11.4	**5.0**
	3.82	**9.4**	9.8	9.2	**8.1**	9.3	**6.7**	8.6	**5.2**
GARCH	2.03	20.9	**15.0**	21.0	**14.3**	21.0	**13.7**	20.9	**12.6**
	2.98	10.8	**10.6**	10.9	**9.5**	11.1	**8.3**	10.9	**6.7**
	3.96	**8.4**	10.2	**7.9**	8.6	7.7	**7.0**	8.0	**5.4**
	4.99	**7.7**	10.8	**6.9**	9.0	**6.2**	7.4	6.1	**5.9**
Filtered	5	**7.9**	13.2	**6.7**	11.2	**5.7**	9.5	**5.2**	7.8
GARCH	7	**9.2**	15.3	**7.5**	13.4	**6.1**	11.5	**4.8**	9.7

Table 1: RMSE($\times 10^{-2}$) of $\widehat{\gamma}(k)$ with k chosen as k^Q or k^{QCRPS} for models **(M1)**–**(M8)** with true tail index α. Lower values for RMSE are set in boldface.

We also observe that k^{QCRPS} improves more vis-à-vis k^Q, the smaller α, i.e., the heavier the tail. A reason for this may be as follows. As pointed out in Section 2.1, k^Q is designed to produce a high agreement between $\widehat{x}_p(k^Q)$ and the order statistics $X_{(\lfloor n(1-p)+1 \rfloor)}$ in the tail. Yet, these large order statistics may be very volatile for heavy-tailed data and, hence, k^Q may be chosen to fit the 'noise' in the tail. Note that the influence of a single large observation $X_{(\lfloor n(1-p)+1 \rfloor)}$ is reduced in (6), since the score is an average over all observations and large p.

Model	α	$n = 500$		$n = 1000$		$n = 2000$		$n = 5000$	
		k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}
t_α	1	7.0	**−0.8**	5.3	**−0.2**	3.1	**−0.4**	0.2	**−0.1**
	3	**1.0**	8.1	**−0.3**	6.5	**−1.0**	5.2	**−1.5**	4.1
	5	**5.3**	12.4	**3.8**	10.8	**2.6**	9.2	**1.5**	7.6
	7	**8.1**	15.0	**6.3**	13.1	**4.9**	11.4	**3.6**	9.6
Fréchet	1	8.2	**0.1**	4.1	**0.1**	1.8	**−0.1**	0.4	**0.0**
	3	−2.7	**−0.0**	−2.6	**0.0**	−2.6	**0.0**	−2.6	**0.0**
	5	−1.5	**0.0**	−1.7	**−0.0**	−1.5	**0.0**	−1.4	**0.0**
	7	−1.1	**−0.0**	−1.1	**0.0**	−1.0	**0.0**	−1.0	**0.0**
SS	0.5	31.2	**3.7**	31.6	**3.9**	28.7	**5.0**	24.6	**4.8**
	1.0	7.6	**−0.4**	3.8	**−0.4**	2.8	**−0.3**	−1.4	**−0.1**
	1.5	**−8.6**	−9.3	**−7.5**	−8.3	**−6.8**	−6.9	−5.8	**−5.3**
	1.9	−25.7	**−23.8**	**−23.0**	−24.9	**−18.2**	−25.4	**−10.8**	−25.1
Pareto	1	7.2	**−1.5**	2.7	**−0.7**	1.0	**−0.7**	0.3	**−0.1**
	3	−3.1	**−0.5**	−2.8	**−0.4**	−2.8	**−0.3**	−2.6	**−0.1**
	5	−1.8	**−0.4**	−1.7	**−0.1**	−1.5	**−0.1**	−1.4	**−0.0**
	7	−1.3	**−0.2**	−1.2	**−0.1**	−1.1	**−0.1**	−1.0	**−0.0**
Burr	1	6.7	**−1.2**	2.8	**−1.1**	2.6	**−0.7**	2.0	**−0.2**
	3	**−1.4**	2.8	**−1.8**	2.4	−2.0	**2.0**	−2.3	**1.7**
	5	**1.9**	5.9	**1.1**	5.2	**0.6**	4.5	**0.1**	3.8
	7	**3.9**	7.6	**3.0**	6.8	**2.3**	6.2	**1.6**	5.3
ARCH	2.30	−5.6	**2.8**	−6.3	**2.4**	−6.5	**1.9**	−6.9	**1.4**
	2.68	**−3.3**	4.1	−3.9	**3.5**	−3.9	**2.9**	−4.4	**2.3**
	3.17	**−0.8**	5.8	**−1.6**	4.8	**−2.0**	4.1	**−2.4**	3.1
	3.82	**1.0**	7.3	**0.2**	6.2	**−0.4**	5.2	**−1.0**	4.2
GARCH	2.03	−18.2	**−4.8**	−18.6	**−3.6**	−18.6	**−2.3**	−18.7	**−1.1**
	2.98	−5.5	**3.7**	−6.0	**3.4**	−6.3	**3.0**	−6.5	**2.4**
	3.96	**0.2**	7.4	**−0.7**	6.2	**−1.4**	5.1	**−1.8**	4.1
	4.99	**3.2**	9.4	**1.9**	7.9	**1.0**	6.6	**0.4**	5.3
Filtered GARCH	5	**4.8**	12.2	**3.4**	10.6	**2.2**	9.1	**1.2**	7.5
	7	**7.7**	14.7	**6.0**	13.0	**4.7**	11.3	**3.4**	9.6

Table 2: Bias($\times 10^{-2}$) of $\widehat{\gamma}(k)$ with k chosen as k^Q or k^{QCRPS} for models **(M1)**–**(M8)** with true tail index α. Lower absolute values for bias are set in boldface.

2. Bias: Just like the RMSE, the bias tends to decrease the longer the sample. The simulation bias in Table 2 varies less systematically than the RMSE: It can be larger for larger α (as for the t_α distribution) or for smaller α (k^Q for GARCH). Also, it can either be mostly negative (Pareto distribution) or mostly positive (t_α distribution). On balance, there do not appear to be marked differences in bias between k^Q and k^{QCRPS}. The latter choice tends to lead to lower bias for heavy-tailed distributions, while the former tends to produce estimates with lower bias

Model	α	$n = 500$		$n = 1000$		$n = 2000$		$n = 5000$	
		k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}
t_α	1	9.7	26	12	38	17	57	24	96
	3	10	28	14	42	18	63	25	108
	5	9.9	28	12	43	15	65	22	113
	7	9.5	28	11	44	14	66	19	116
Fréchet	1	9.4	26	13	38	16	57	23	96
	3	13	28	18	42	23	63	32	107
	5	15	29	19	43	25	63	36	108
	7	15	29	20	43	26	64	38	109
SS	0.5	6.4	20	8.7	27	12	38	17	57
	1.0	9.4	26	12	39	16	57	24	97
	1.5	13	28	18	41	23	61	33	102
	1.9	13	27	19	41	26	59	33	105
Pareto	1	9.6	26	13	38	17	56	24	96
	3	14	28	18	42	24	63	33	106
	5	15	29	20	43	26	64	36	108
	7	15	29	21	44	27	64	37	109
Burr	1	9.7	26	13	39	17	57	23	96
	3	12	28	16	42	21	62	29	106
	5	12	29	15	43	20	64	27	109
	7	11	29	14	44	18	65	25	113
ARCH	2.30	9.8	28	12	42	15	62	20	106
	2.68	10	28	12	42	15	63	21	107
	3.17	10	28	13	43	16	63	22	108
	3.82	10	28	13	43	17	64	23	110
GARCH	2.03	8.2	28	9.9	42	11	62	13	106
	2.98	9.1	28	11	43	14	64	18	109
	3.96	9.7	28	12	43	15	65	21	111
	4.99	9.8	28	12	44	16	65	22	113
Filtered	5	9.5	28	12	43	15	65	21	113
GARCH	7	9.2	28	11	44	14	66	19	116

Table 3: Average number of k used in $\widehat{\gamma}(k)$ with k chosen as k^Q or k^{QCRPS} for models **(M1)**–**(M8)** with true tail index α.

for lighter-tailed distributions. Since it is for heavy-tailed distributions that (semiparametric) extreme value methods have a larger relative advantage over nonparametric methods (Hoga, 2019b+), the choice k^{QCRPS} appears to be preferable in terms of bias.

3. Average number of k: By definition, $1 \leq k^Q, k^{\mathrm{QCRPS}} \leq k_{\max}$. For $n = 500, 1000, 2000, 5000$ we have $k_{\max} = \lfloor n^{0.6} \rfloor = 41, 63, 95, 165$, respectively. Hence, Table 3 shows that both k^Q and

Model	α	$n = 500$		$n = 1000$		$n = 2000$		$n = 5000$	
		k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}
t_α	1	11	10	16	16	23	25	38	46
	3	9.5	7.2	14	11	20	16	34	28
	5	8.4	6.2	11	9.2	17	13	27	23
	7	7.8	5.7	10	8.2	15	11	23	19
Fréchet	1	10	10	16	16	23	25	38	46
	3	11	7.8	17	12	25	18	41	31
	5	11	7.5	17	11	25	17	42	30
	7	11	7.4	17	11	25	17	42	30
SS	0.5	9.5	11	14	18	21	27	34	48
	1.0	10	10	16	16	23	25	38	46
	1.5	12	8.8	18	13	27	20	44	36
	1.9	10	7.3	16	11	24	18	36	26
Pareto	1	11	10	16	16	24	25	38	46
	3	11	7.9	17	12	25	18	41	31
	5	11	7.5	17	11	25	17	42	30
	7	11	7.5	18	11	26	17	42	30
Burr	1	11	10	16	16	23	25	38	46
	3	11	7.7	16	12	23	18	38	31
	5	10	7.2	14	10	21	16	34	28
	7	9.5	6.7	13	10	19	14	30	25
ARCH	2.30	9.1	7.5	13	11	19	17	31	29
	2.68	9.1	7.1	13	11	18	16	31	29
	3.17	9.0	6.9	13	10	19	16	31	28
	3.82	8.8	6.5	13	10	19	15	31	26
GARCH	2.03	7.2	6.3	10	9.8	13	14	21	26
	2.98	7.9	6.3	11	10	16	15	27	26
	3.96	8.4	6.1	12	9.6	18	14	29	26
	4.99	8.1	5.8	11	8.9	17	13	29	24
Filtered	5	8.1	5.9	11	9.0	16	13	27	22
GARCH	7	7.6	5.5	10	8.0	15	11	22	18

Table 4: Standard deviation of k's used in $\widehat{\gamma}(k)$ with k chosen as k^Q or k^{QCRPS} for models **(M1)**–**(M8)** with true tail index α.

k^{QCRPS} do not have a strong tendency to pick either of the corner solutions 1 or $k_{\max}$ of the minimization problem in (7).

Furthermore, for fixed sample size n, there is little variation in average k^{QCRPS} across models, unlike for k^Q. For instance, for $n = 500$ and excluding the SS distribution with $\alpha = 0.5$, the average k^{QCRPS} varies between 26 and 29, while the average k^Q is between 8.2 and 15. The small variation in k^{QCRPS} is quite remarkable, as the underlying distributions exhibit quite different

tail behaviour.

4. Standard deviation of k's: Table 4 shows that, with the exception of extremely heavy-tailed data with $\alpha \leq 1$, the standard deviations of the choice k^{QCRPS} tend to be lower than those for k^{Q}. This supports the intuition that the choice k^{Q} is more volatile, since it is designed to produce a high agreement between $\widehat{x}_p(k)$ and the noisy tail observations $X_{(\lfloor n(1-p)+1 \rfloor)}$.

5. Comparing the results for the t_5- and t_7-distribution with those for the filtered GARCH residuals, that are only approximately (standardized) t_5- and t_7-distributed, we find the differences in Tables 1–4 to be very small. This is as expected, because Chan *et al.* (2007) show that tail index estimates of the GARCH-errors are asymptotically not affected by the GARCH-filter. This is because model parameters can be estimated $\sqrt{n}$-consistently, whereas tail index estimates converge at a slower $\sqrt{k}$-rate.

6. The models with $\alpha \leq 1$ in Table 1 have infinite first moments. However, finite first moments are required for propriety of the scoring rule underlying the choice k^{QCRPS}. Nonetheless, k^{QCRPS} significantly outperforms k^{Q} in terms of RMSE even for those distributions with $\alpha \leq 1$.

We conclude that k^{QCRPS} leads to more precise Hill estimates of γ than k^{Q} for heavy-tailed distributions with small tail index α and also for lighter-tailed data when the sample size n is sufficiently large. But k^{QCRPS} also leads to good results in the remaining cases. The reason for the good performance of k^{QCRPS} may be that more observations are used when employing k^{QCRPS}—thus, reducing the variance of the tail index estimates—while at the same time keeping the bias constant, since only additional observations from the tail are included. Thus, for particularly heavy-tailed data, which are often found in economics and finance, k^{QCRPS} often offers a superior bias–variance trade-off.

Remark 8. (a) The above conclusions may depend on the particular tail index estimator being used. To provide further evidence on the merit inherent in the choice k^{QCRPS}, we also consider the log-log rank-size estimator of Gabaix and Ibragimov (2011) in Appendix B of the Online Supplemental Appendices. The results in Appendix B demonstrate even larger efficiency gains from using k^{QCRPS} than those reported here for the Hill (1975) estimator.

(b) The smallest sample size of $n = 500$ we consider in the simulations seems to be a lower bound for applications of EVT in practice; see, e.g., Quintos *et al.* (2001) and Hill (2015). The theoretical literature supporting the use of EVT in smaller samples is in its infancy; see Müller and Wang (2017). We speculate that even in these shorter samples our approach to the choice of k works well, even though in the above simulations we observe some tendency of k^{QCRPS} to work better

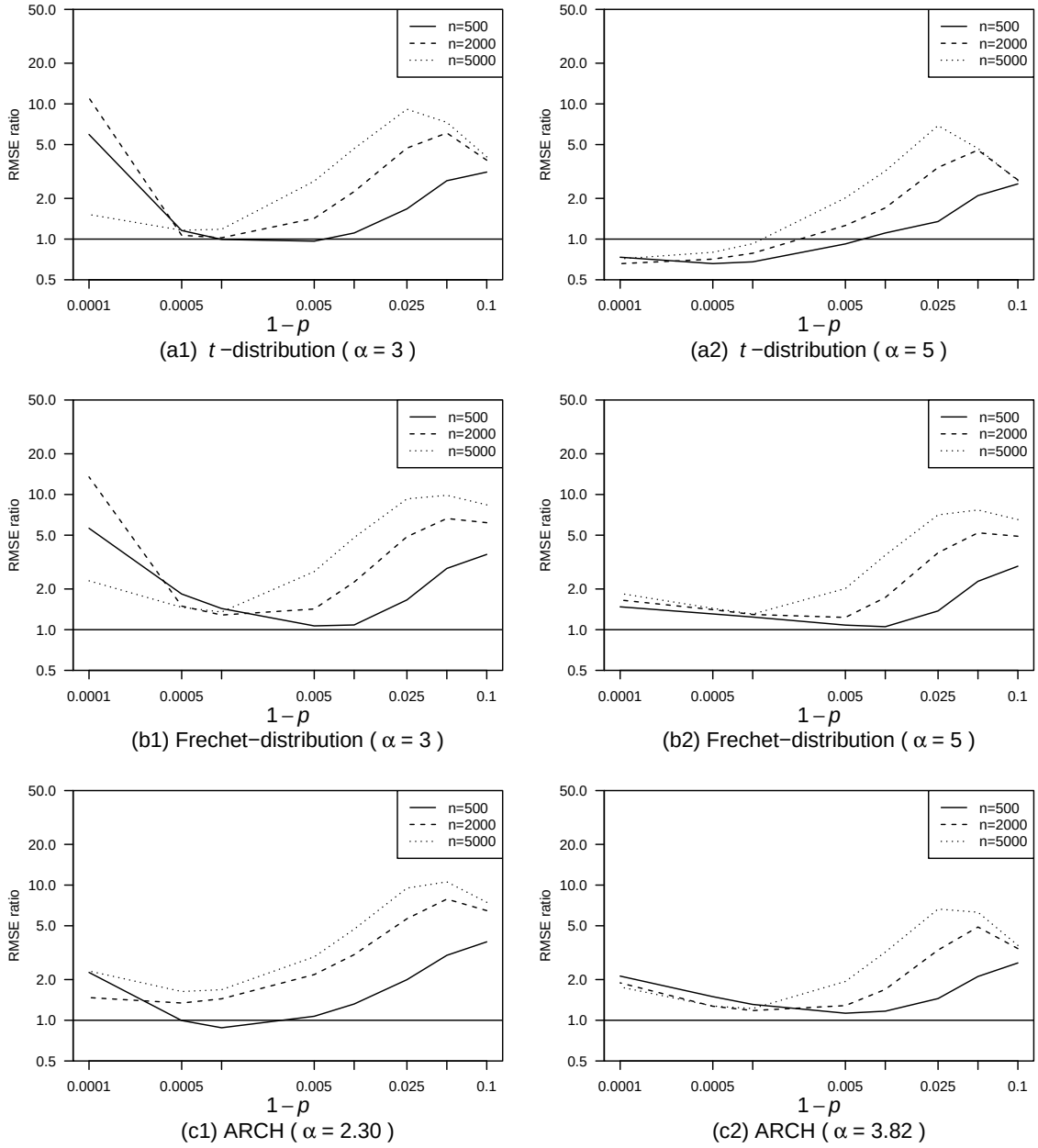

Figure 1: Log-log plots of $\mathrm{RMSE}_p(k^Q)/\,\mathrm{RMSE}_p(k^{\mathrm{QCRPS}})$ for different values of $1-p$. Here, $\mathrm{RMSE}_p(k)$ is the RMSE of $\widehat{x}_p(k)$ ($k \in \{k^Q, k^{\mathrm{QCRPS}}\}$). Results shown for models **(M1)**, **(M2)**, and **(M6)** with indicated α.

in larger samples. We think so because the results in Appendix B suggest that k^{QCRPS} may sometimes lead to more precise estimates of γ in smaller samples.

3.3 Estimation of extreme VaR

In typical financial applications, interest centers not so much on γ, but rather on risk measures like the Value-at-Risk at level p (with p typically close to 1), i.e., the p-quantile of the P&L of the portfolio.

VaR at level p can be estimated by the Weissman (1978) estimator $\widehat{x}_p(k)$ based on losses $X_1, \ldots, X_n$.

To assess if the choice k^{QCRPS} improves VaR estimation at different levels, Figure 1 shows the ratio of the RMSEs of $\widehat{x}_p(k^Q)$ over those of $\widehat{x}_p(k^{\text{QCRPS}})$ for different values of $n \in \{500, 2000, 5000\}$ and $p \in \{0.90, 0.95, 0.975, 0.99, 0.995, 0.999, 0.9995, 0.9999\}$. To conserve space, we only consider the models **(M1)**, **(M2)**, and **(M6)** of Daníelsson *et al.* (2016) and the values of α indicated in the panel captions. The plots for the remaining models **(M3)**–**(M5)** and **(M7)**–**(M8)** can be found in Appendix A of the Online Supplemental Appendices.

In general, the good properties of $\widehat{\gamma}(k^{\text{QCRPS}})$ carry over to $\widehat{x}_p(k^{\text{QCRPS}})$. Since most lines are above 1, the choice k^{QCRPS} leads to more precise quantile estimates than k^Q in most cases. The superiority is particularly marked for the small levels $p = 0.90, 0.95$ and the most extreme quantiles of the very heavy-tailed t_2- and Fréchet(2)-distribution. Also, the observations from Subsection 3.2 that the relative advantage of the tail shape estimates $\widehat{\gamma}(k^{\text{QCRPS}})$ over $\widehat{\gamma}(k^Q)$ get larger for larger n and smaller α, manifest themselves in Figure 1. There, the improvements in VaR estimation precision also tend to be larger for longer samples and heavier tails.

In Appendix A of the Online Supplemental Appendices we complement Figure 1 with a similar plot for estimates of Expected Shortfall—another popular risk measure in the financial industry. This plot further supports our choice k^{QCRPS}.

3.4 Inference on γ with k^{QCRPS}

For i.i.d. data with tails obeying a second-order refinement of (1), it is well-known (see, e.g., Resnick, 2007, Proposition 9.3) that

$$\sqrt{k}(\widehat{\gamma}(k) - \gamma) \xrightarrow{d} N(0, \gamma^2) \tag{10}$$

for some intermediate sequence k. For a wide range of dependent data, a similar result holds with a different asymptotic variance than γ^2 (Hill, 2010). Our choice of k is obviously stochastic and, hence, the asymptotic approximation in (10) may be misleading when drawing inference on γ in practice.

To judge the possible distortions caused by our stochastic choice k^{QCRPS}, we compare the kernel density estimates of

$$\frac{\sqrt{k^{\text{QCRPS},(r)}}}{\gamma}(\widehat{\gamma}^{(r)}(k^{\text{QCRPS},(r)}) - \gamma) \quad \text{and} \quad \frac{\sqrt{\overline{k}}}{\gamma}(\widehat{\gamma}^{(r)}(\overline{k}) - \gamma), \tag{11}$$

where $r = 1, \ldots, R$ again runs over all $R = 10000$ replications, and $\overline{k}$ is the average value of k^{QCRPS} reported for the respective model in Table 3. The discrepancy between both kernel density plots—shown in Figure 2—may then be regarded as a measure for the distortions incurred by using the

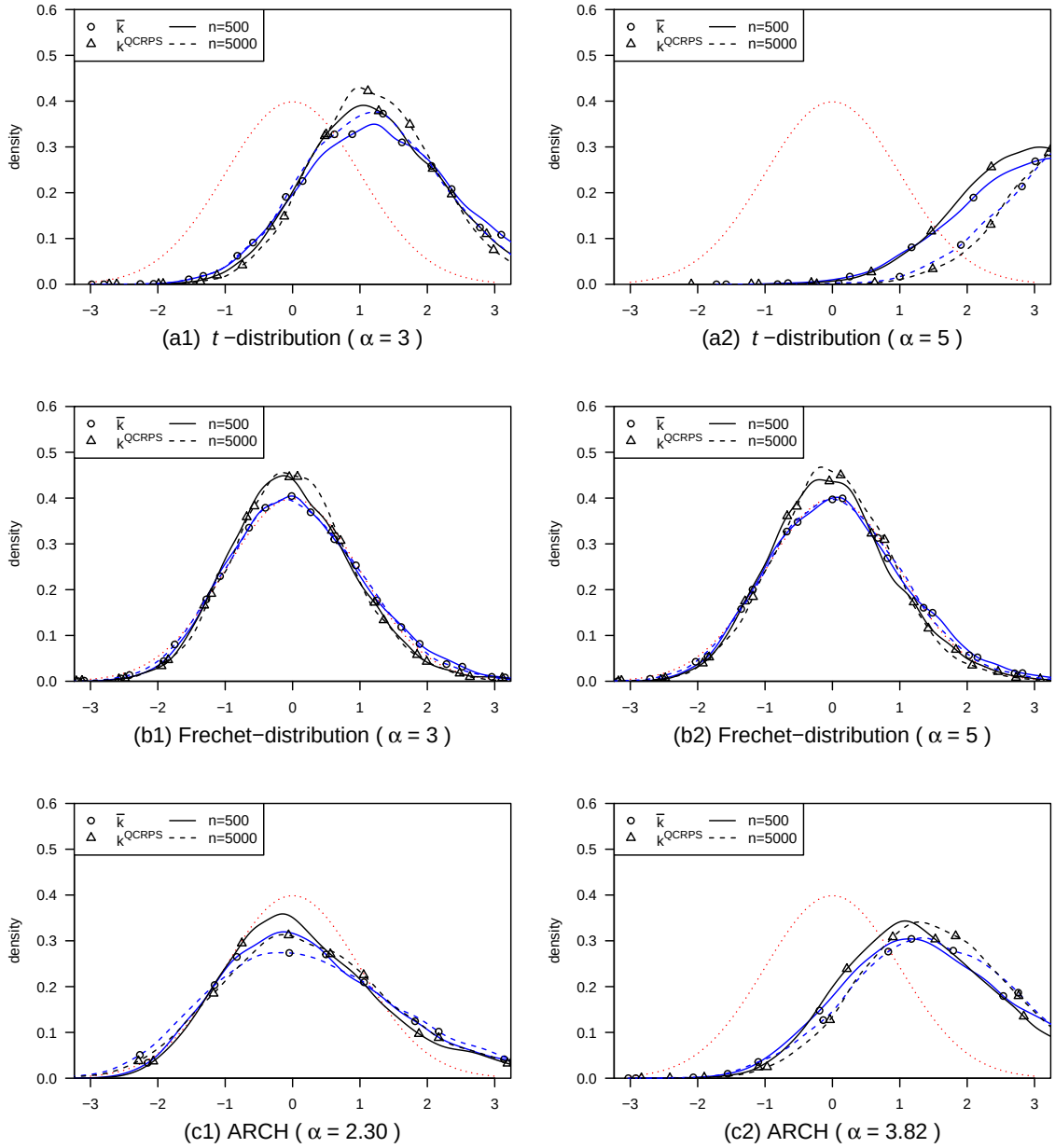

Figure 2: Kernel density plots of $\frac{\sqrt{k^{\mathrm{QCRPS},(r)}}}{\gamma}(\widehat{\gamma}^{(r)}(k^{\mathrm{QCRPS},(r)}) - \gamma)$ (black line marked with triangles) and $\frac{\sqrt{\overline{k}}}{\gamma}(\widehat{\gamma}^{(r)}(\overline{k}) - \gamma)$ (blue line marked with circles) for models **(M1)**, **(M2)**, and **(M6)** with indicated α. The dotted red line depicts the standard normal density. See the web version of this article for colour.

stochastic k^{QCRPS} instead of a deterministic one.

Clearly, it would also be possible to compare the kernel density plot based on $\sqrt{k^{\mathrm{QCRPS},(r)}}(\widehat{\gamma}^{(r)}(k^{\mathrm{QCRPS},(r)}) - \gamma)$ with the density of a $N(0, \gamma^2)$-distribution. Yet, depending on the choice of k and second-order behaviour of $F(\cdot)$, bias terms in the mean of the asymptotic distribution may appear (see, e.g., de Haan

and Ferreira, 2006, Theorem 3.2.5). Also, for dependent data (e.g., the (G)ARCH models in **(M6)** and **(M7)**) the asymptotic variance γ^2 is inflated by some hard-to-calculate factor (Hoga, 2017). For these reasons we present the comparison as outlined above.

Figure 2 shows that the finite-sample distribution of the quantities in (11) does not differ significantly. Hence, choosing the stochastic k^{QCRPS} does not invalidate inference for the tail index, when compared with the fixed $\overline{k}$. Rather, other factors—most notably asymptotic bias terms, but possibly also asymptotic variances different from γ^2 for the ARCH processes—have a larger impact on inference and thus invalidate the standard normal limit for the two quantities in (11).

4 Application

We explore the benefits of using k^{QCRPS} in a risk management application. Consider the $N = 5032$ log-losses of the five Dow Jones stocks Apple (AAPL), Disney (DIS), Microsoft (MSFT), Nike (NKE) and Pfizer (PFE). These shares represent a highly diversified set of U.S. blue-chips. The log-losses are sampled in the 20 year period from 1/1/1998 through 12/31/2017 and are calculated as $X_i = -\log(P_i/P_{i-1})$, where P_i denotes the closing price on day i.[5] In this application, we consider rolling window forecasts—based on some window length n—of the Value-at-Risk, which is of key importance in risk management.

For ease of exposition, consider the losses $X_1, \ldots, X_n$ in the first window. The Value-at-Risk is defined as the loss that is only exceeded with some small probability $(1 - p)$ by 'tomorrow's' loss X_{n+1} conditional on $X_1, \ldots, X_n$. Clearly, one could use the estimate $\widehat{x}_p(k)$ of the unconditional quantile $F^{\leftarrow}(p)$ based on the raw losses $X_1, \ldots, X_n$, as was done in the simulations for models **(M1)**–**(M7)**. Yet, such an estimate, while accurate on average, is likely to be too low when market conditions—embodied by $X_1, \ldots, X_n$—are volatile and too high when markets are calm. Hence, to take into account this volatility clustering of the losses—shown exemplarily for the Apple shares in Figure 3—we focus on the conditional Value-at-Risk, $\mathrm{VaR}_{p,n+1}$, i.e., the p-quantile of the conditional d.f. $F_{n+1}(x) := P\left\{X_{n+1} \leq x \mid X_n, X_{n-1}, \ldots\right\}$. The conditional d.f. $F_n(\cdot)$ is more more informative than the unconditional d.f. $F(\cdot)$, as it incorporates the current state of the market. We describe next how the methods proposed in this paper can be used to compute $\mathrm{VaR}_{p,n+1}$.

Much, if not all, of the variation in the conditional d.f. $F_{n+1}(x)$ of losses on speculative assets is due to changes in the variance. The benchmark models to incorporate such changes are Bollerslev's

[5] All data have been taken from *finance.yahoo.com*.

(1986) GARCH(1,1) processes

$$X_i = \sigma_i \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} (0,1),$$
$$\sigma_i^2 = \omega^\circ + \alpha^\circ X_{i-1}^2 + \beta^\circ \sigma_{i-1}^2, \qquad \omega^\circ > 0,\ \alpha^\circ \geq 0,\ \beta^\circ \geq 0. \tag{12}$$

For GARCH models, it is easy to show that

$$\mathrm{VaR}_{p,n+1} = \sigma_{n+1} F_\varepsilon^{\leftarrow}(p),$$

where $F_\varepsilon(\cdot)$ is the d.f. of ε_i.

To obtain VaR forecasts from log-losses $X_1, \ldots, X_n$ in practice, McNeil and Frey (2000) suggest a two-step procedure based on EVT. While refinements of McNeil and Frey's (2000) popular two-step procedure exist (Laurini and Tawn, 2008), it has been shown to be theoretically sound (Chan *et al.*, 2007; Hoga, 2019a+) and robust to misspecification (Jalal and Rockinger, 2008). In the first step, we obtain QML parameter estimates to forecast volatility $\widehat{\sigma}_{n+1}^2 = \widehat{\omega}^\circ + \widehat{\alpha}^\circ X_n^2 + \widehat{\beta}^\circ \widehat{\sigma}_n^2$. Then, the standardized residuals $\{\widehat{\varepsilon}_i = X_i/\widehat{\sigma}_i\}_{i=11,\ldots,n}$ are used to estimate $F_\varepsilon^{\leftarrow}(p)$ using the Weissman (1978) estimator, say $\widehat{x}_p^{\widehat{\varepsilon}}(k)$.[6] Thus, in this application, the method of choosing k is not applied to the raw data $\{X_i\}$, but rather to the GARCH-filtered residuals $\{\widehat{\varepsilon}_i\}$, similarly as in model **(M8)** in the simulations. The resulting VaR forecast is then

$$\widehat{\mathrm{VaR}}_{p,n+1}^{k} = \widehat{\sigma}_{n+1} \widehat{x}_p^{\widehat{\varepsilon}}(k).$$

We repeat this procedure based on a moving window $X_j, \ldots, X_{j+n-1}$ $(j = 1, \ldots, N-n)$ to obtain VaR forecasts $\widehat{\mathrm{VaR}}_{p,n+1}^{k,(j)}$ $(j = 1, \ldots, N-n)$. We compare the choices $k = k^Q$ and $k = k^{\mathrm{QCRPS}}$. Furthermore, we let $n = 500, 1000, 2000$, such that for each moving window we have, respectively, $490, 990, 1990$ standardized residuals for extreme quantile estimation. We consider the levels $p = 95\%, 99\%, 99.5\%, 99.9\%, 99.95\%, 99.99\%$.

To illustrate this procedure, Figure 3 displays the rolling window estimates of the GARCH parameters $(\omega^\circ, \alpha^\circ, \beta^\circ)$ for the Apple log-losses with $n = 2000$. There is some evidence for a structural break in the GARCH parameters during the financial crisis in 2008. As is frequently found for financial data, $\alpha^\circ + \beta^\circ$ is close to, but less than, one and the tail index estimates fluctuate around four. The bottom plot shows the log-losses together with the 99%-VaR forecasts, which are calculated based on the previous $n = 2000$ observations.

As a benchmark, we also consider VaR forecasts with a nonparametric (NP) estimator of $F_\varepsilon^{\leftarrow}(p)$ to illustrate the advantages of the semiparametric Weissman (1978) estimator. Specifically, the forecast

[6] We discard the first 10 standardized residuals $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_{10}$, because they are unreliable due to initialization effects in the variance equation (12).

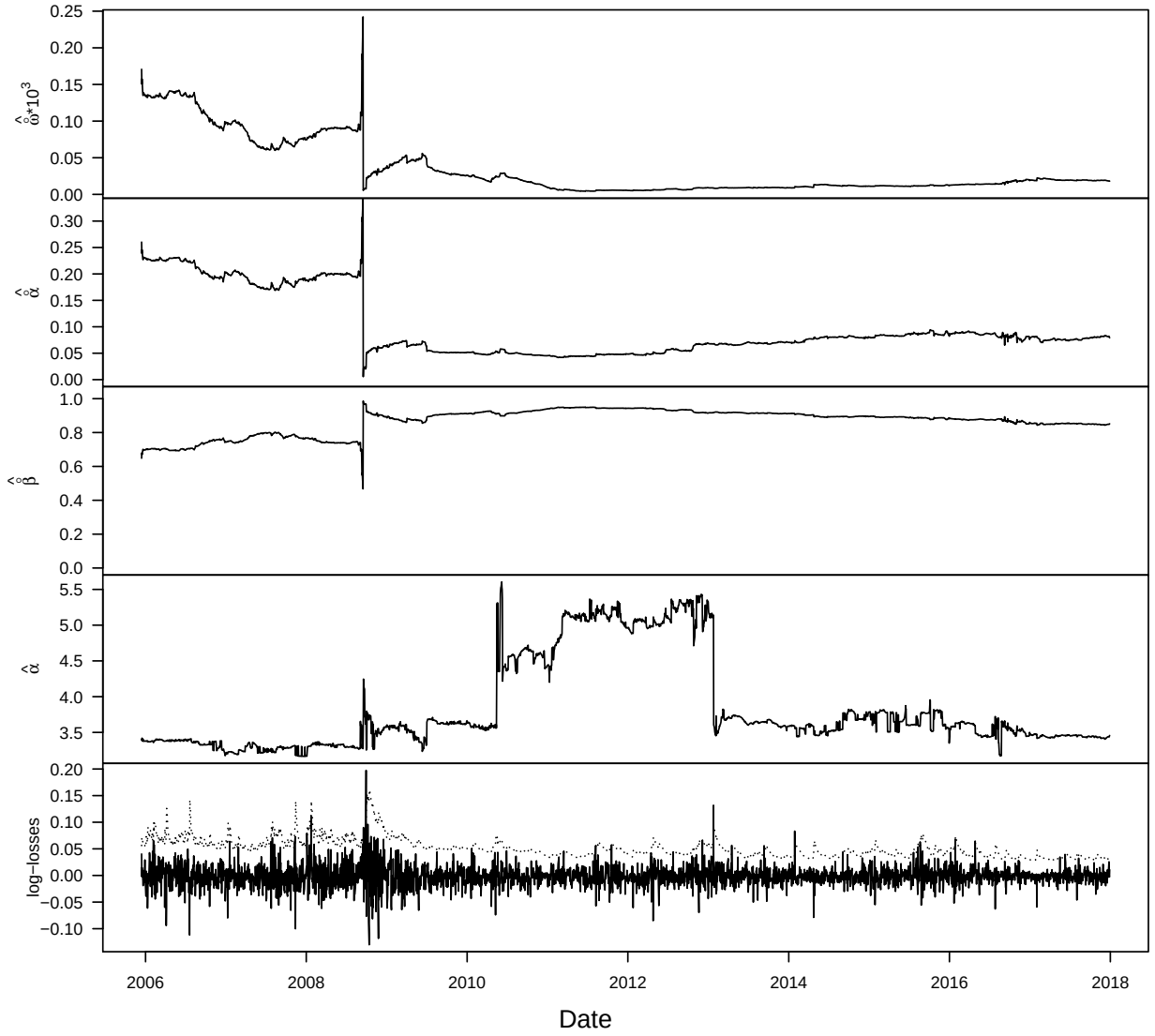

Figure 3: Rolling window estimates of GARCH parameters $(\widehat{\omega}^{\circ}, \widehat{\alpha}^{\circ}, \widehat{\beta}^{\circ})$ and tail index $\widehat{\alpha} = 1/\widehat{\gamma}(k^{\mathrm{QCRPS}})$ for Apple returns with $n = 2000$. The bottom plot displays the resulting rolling window VaR forecasts $\widehat{\mathrm{VaR}}_{p=0.99}^{k^{\mathrm{QCRPS}},(j)}$ (dotted line) together with the log-losses (solid line) X_{j+n} $(j = 1, \ldots, N-n)$.

of $\mathrm{VaR}_{p,n+1}$ based on $X_1, \ldots, X_n$ is given by

$$\widehat{\mathrm{VaR}}_{p,n+1}^{\mathrm{NP}} = \widehat{\sigma}_{n+1}\widehat{\varepsilon}_{(\lfloor n(1-p)\rfloor+1)},$$

where $\widehat{\varepsilon}_{(\lfloor n(1-p)\rfloor+1)}$ is the $(\lfloor n(1-p)\rfloor + 1)$-largest value of $\{\widehat{\varepsilon}_i = X_i/\widehat{\sigma}_i\}_{i=11,\ldots,n}$. This procedure is iterated over moving windows as before to yield $(N-n)$ VaR forecasts $\widehat{\mathrm{VaR}}_{p,n+1}^{\mathrm{NP},(j)}$, $j = 1, \ldots, N-n$.

We compare the three different sets of forecasts via their average scores (see Remark 2)

$$\overline{S}^{M}_{p,\mathrm{QS}} = \frac{1}{N-n}\sum_{j=1}^{N-n} \mathrm{QS}_p(\widehat{\mathrm{VaR}}^{M,(j)}_{p,n+1}, X_{n+j}),$$

where $M \in \{k^Q, k^{\mathrm{QCRPS}}, \mathrm{NP}\}$. Table 5 shows the ratios of the average scores $\overline{S}^{k^{\mathrm{QCRPS}}}_{p,\mathrm{QS}}/\overline{S}^{k^Q}_{p,\mathrm{QS}}$ and $\overline{S}^{k^{\mathrm{QCRPS}}}_{p,\mathrm{QS}}/\overline{S}^{\mathrm{NP}}_{p,\mathrm{QS}}$ for the different stocks and $n \in \{500, 1000, 2000\}$. Since scores are negatively oriented, values below 1 indicate a superior performance of the forecasts $\widehat{\mathrm{VaR}}^{k^{\mathrm{QCRPS}},(j)}_{p,n+1}$.

To judge the statistical significance of the VaR score differences $\overline{S}^{k^{\mathrm{QCRPS}}}_{p,\mathrm{QS}} - \overline{S}^{k^Q}_{p,\mathrm{QS}}$ and $\overline{S}^{k^{\mathrm{QCRPS}}}_{p,\mathrm{QS}} - \overline{S}^{\mathrm{NP}}_{p,\mathrm{QS}}$, we follow Taylor (2019) and Hoga (2019a+) and apply a simple Diebold and Mariano (1995) test. More precisely, to judge the statistical significance of the score difference $\overline{S}^{k^{\mathrm{QCRPS}}}_{p,\mathrm{QS}} - \overline{S}^{k^Q}_{p,\mathrm{QS}}$ (the other score difference can be handled similarly), we use the test statistic $\sqrt{N-n}\frac{\overline{S}^{k^{\mathrm{QCRPS}}}_{p,\mathrm{QS}} - \overline{S}^{k^Q}_{p,\mathrm{QS}}}{\widehat{\sigma}_{N-n}}$ suggested by Diebold and Mariano (1995), where $\widehat{\sigma}_{N-n}$ denotes the sample standard deviation of the score differences $\Delta_j = \mathrm{QS}_p(\widehat{\mathrm{VaR}}^{k^{\mathrm{QCRPS}},(j)}_{p}, X_{n+j}) - \mathrm{QS}_p(\widehat{\mathrm{VaR}}^{k^Q,(j)}_{p}, X_{n+j})$. Under the null of equal forecast accuracy, the test statistic is standard normally distributed. Score ratios with statistically significant score differences at the $10\%, 5\%, 1\%$-level are marked with a $^*/^{**}/^{***}$ in Table 5. We make the following observations from Table 5.

1. For the smallest level $p = 95\%$ there is little difference between the methods. However, as expected, the extreme value methods improve significantly upon the nonparametric estimates the more extreme p. The ratios can go as low as 0.04 for the Apple stocks with $n = 500$ and $p = 99.99\%$.

2. Comparing only the extreme value methods, we find that choosing k^{QCRPS} leads to better forecasts than k^Q in most cases, particularly for more extreme levels of p.

   This may be explained as follows. As already seen in the simulations in Section 3, k^Q tends to choose fewer upper order statistics for tail estimation than k^{QCRPS}. This may lead to quantile estimates that are too volatile and hence are outperformed by k^{QCRPS}.

The results clearly demonstrate the superior performance of EVT-enhanced methods. They also suggest that one can improve these methods by a judicious choice of k, such as k^{QCRPS} proposed in this paper.

The choice k^{QCRPS} *and* the forecast evaluation method rely on the quantile score QS_p. Hence, one may be concerned that the forecast evaluation method is more favourable to k^{QCRPS} than to k^Q. To confirm that this is not the reason for the out-performance of the new method, we carry out the forecast comparison using a different scoring function. Gneiting (2011, Theorem 9 (c)) shows that all

Stock	n	Average k		Method	p					
		k^Q	k^{QCRPS}		95%	99%	99.5%	99.9%	99.95%	99.99%
AAPL	500	14	28	k^Q	1.00	**0.99**	**0.98**	**0.94**	**0.96**	**0.12**
				NP	**0.99**	**0.99**	**0.97*****	**0.83*****	**0.79***	**0.04**
	1000	25	43	k^Q	**0.99****	**0.99***	**0.97***	**0.91**	**0.86**	1.18***
				NP	**0.99**	1.01	**0.99**	**0.89**	**0.73****	**0.10**
	2000	31	63	k^Q	**0.99*****	**0.99**	**0.98***	**0.97**	**0.94**	1.09***
				NP	1.00	1.00	**0.99**	**0.94**	**0.95**	**0.45*****
DIS	500	10	28	k^Q	**0.94*****	1.00	1.00	**0.91**	**0.83**	**0.80*****
				NP	**0.99**	**0.99**	1.00	**0.90**	**0.77**	**0.07***
	1000	10	43	k^Q	**0.94*****	1.00	1.00	**0.90**	**0.82****	**0.42**
				NP	1.00	1.00	1.00	**0.81****	**0.88***	**0.15**
	2000	26	58	k^Q	**0.96****	**0.99**	**0.99**	**0.97**	**0.93***	**0.84*****
				NP	1.00	**0.99**	**0.99**	**0.94**	**0.93****	**0.15**
MSFT	500	11	28	k^Q	**0.96*****	**0.99**	**0.99**	**0.95**	**0.84**	**0.76*****
				NP	**0.99**	**0.99**	**0.96*****	1.00	**0.85**	**0.09****
	1000	14	39	k^Q	**0.91*****	**0.94*****	**0.98**	**0.96**	**0.88**	**0.41**
				NP	**0.99*****	**0.98****	**0.99**	**0.92***	**0.85***	**0.09**
	2000	41	52	k^Q	**0.99**	1.00	1.00	**0.96****	**0.93****	**0.87*****
				NP	1.00	1.00	1.00	**0.93**	**0.85**	**0.18**
NKE	500	13	28	k^Q	**0.99**	1.00	**0.98**	**0.87***	**0.74***	1.05***
				NP	**0.99**	**0.99**	**0.97***	**0.87***	**0.73***	**0.07****
	1000	19	44	k^Q	**0.99**	1.00	1.00	**0.91**	**0.90**	1.01*
				NP	**0.99**	1.00	**0.98**	**0.84****	**0.89**	**0.22**
	2000	30	67	k^Q	**0.97*****	**0.98****	**0.98**	**0.98**	**0.99**	1.03***
				NP	**0.99****	1.00	**0.99**	**0.91**	**0.95**	**0.79**
PFE	500	8	28	k^Q	**0.93*****	**0.98**	1.00	1.11	1.06	**0.43*****
				NP	**0.99**	**0.99**	1.00	1.06	1.00	**0.09***
	1000	11	41	k^Q	**0.90*****	**0.96****	**0.99**	1.01	1.00	**0.61*****
				NP	1.00	**0.99**	1.00	**0.99**	**0.97**	**0.19**
	2000	25	64	k^Q	**0.96*****	**0.99**	1.00	**0.95*****	**0.90*****	**0.66*****
				NP	1.00***	1.00***	1.02***	**0.93**	**0.96*****	**0.36**

Table 5: Ratios $\overline{S}^{k^{\text{QCRPS}}}_{p,\text{QS}}/\overline{S}^{k^Q}_{p,\text{QS}}$ and $\overline{S}^{k^{\text{QCRPS}}}_{p,\text{QS}}/\overline{S}^{\text{NP}}_{p,\text{QS}}$ of the average scores for different stocks. Values below 1 are set in boldface. Average score ratios marked with $^{*}/^{**}/^{***}$ indicate statistically significant score differences at the $10\%, 5\%, 1\%$-level. The column "Average k" displays the average values of k^Q and k^{QCRPS} over all $(N-n)$ VaR forecasts.

scoring functions of the form

$$S(F^{\leftarrow}(p), x) = [I_{\{y \leq F^{\leftarrow}(p)\}} - p][g(F^{\leftarrow}(p)) - g(x)]$$

are strictly consistent for the p-quantile, if $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing. The quantile score obtains

for $g(x) = x$. In a related context, Ziegel *et al.* (2019) suggest $g(x) = \exp(x)/[1 + \exp(x)]$. Carrying out the forecast comparison with a scoring function based on this latter choice, we find the results of Table 5 to be confirmed. Indeed, for all but the most extreme p, the score ratios are almost unchanged. In the interest of space the results are not included in the paper, but are available from the author upon request.

5 Summary

Extreme value methods have gained significant popularity in economics, finance and beyond. However, in practical applications it is difficult to tell where 'the tail begins', i.e., to determine the number k of large observations for which extreme value methods can be validly applied. Choosing k thus becomes a crucial part of applying EVT in practice. In this paper, we introduce a novel idea to choose k based on proper scoring rules. Such scoring rules have hitherto been mainly applied in forecast evaluation (Gneiting and Ranjan, 2011). In simple terms, we propose a choice of k that leads to a minimum score of the Weissman (1978) quantile estimates over probability levels in the right tail. This idea may also be applied for POT estimation and in the estimation of co-crash probabilities.

We show in simulations that—particularly for heavy-tailed data of primary interest in applications—our choice k^{QCRPS} often leads to more precise estimates of the tail index than its main competitor k^{Q}. These more precise tail index estimates directly translate into more reliable (extreme) risk measure estimates.

In an application to returns on a diversified set of U.S. blue-chip stocks, we find that k^{QCRPS} also leads to better VaR forecasts in risk management practice. The increased precision of tail estimates may be explained by the fact that our choice of k provides a better bias-variance trade-off: It tends to suggest more upper order statistics than k^{Q} (thus decreasing variance), while—at the same time—not increasing the bias, that may be introduced by using non-tail observations.

References

Beirlant J, Dierckx G, Goegebeur Y, Matthys G. 1999. Burr regression and portfolio segmentation. *Extremes* **2**: 177–200.

Beirlant J, Goegebeur Y, Segers J, Teugels J. 2004. *Statistics of Extremes*. Chichester: Wiley.

Bollerslev T. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* **31**: 307–327.

Bosma JJ, Koetter M, Wedow M. 2019. Too connected to fail? inferring network ties from price co-movements. *Journal of Business & Economic Statistics* **37**: 67–80.

Bücher A, Jäschke S, Wied D. 2015. Nonparametric tests for constant tail dependence with an application to energy and finance. *Journal of Econometrics* **187**: 154–168.

Chan NH, 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.

Chen SX, Tang SX. 2005. Nonparametric inference of value-at-risk for dependent financial returns. *Journal of Financial Econometrics* **3**: 227–255.

Clauset A, Shalizi CR, Newman MEJ. 2009. Power-law distributions in empirical data. *SIAM Review* **51**: 661–703.

Csörgő S, Deheuvels P, Mason D. 1985. Kernel estimates of the tail index of a distribution. *The Annals of Statistics* **13**: 1050–1077.

Dacarogna MM, Gençay R, Müller UA, Olsen RB, Pictet OV. 2001. *An Introduction to High-Frequency Finance*. San Diego: Academic Press.

Daníelsson J, de Haan L, Ergun LM, de Vries CG. 2016. Tail index estimation: Quantile driven threshold selection (Available at SSRN: *http://dx.doi.org/10.2139/ssrn.2717478*).

Danielsson J, de Haan L, Peng L, de Vries CG. 2001. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis* **76**: 226–248.

Davis RA, Mikosch T. 1998. The sample autocorrelations of heavy-tailed processes with applications to ARCH. *The Annals of Statistics* **26**: 2049–2080.

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

Dey DK, Yan J (eds.) . 2016. *Extreme Value Modeling and Risk Analysis: Methods and Applications*. Boca Raton: Chapman & Hall.

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* **13**: 253–263.

Drees H. 1995. Refined Pickands estimators for the extreme value index. *The Annals of Statistics* **23**: 2059–2080.

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

Drees H, Janßen A, Resnick SI, Wang T. 2018. On a minimum distance procedure for threshold selection in tail analysis. *arXiv e-prints* : arXiv:1811.06433.

Drees H, Kaufmann E. 1998. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications* **75**: 149–172.

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

Embrechts P, Lindskog F, McNeil A. 2003. Modelling dependence with copulas and applications to risk management. In Rachev ST (ed.) *Handbook of Heavy Tailed Distributions in Finance*, vol. 1 of *Handbooks in Finance*, Amsterdam: North-Holland, pages 329–384.

Fagiolo G, Napoletano M, Roventini A. 2008. Are output growth-rate distributions fat-tailed? Some evidence from OECD countries. *Journal of Applied Econometrics* **23**: 639–669.

Francq C, Zakoïan JM. 2010. *GARCH Models: Structure, Statistical Inference and Financial Applications*. Chichester: Wiley.

Gabaix X. 1999. Zipf's law for cities: An explanation. *Quarterly Journal of Economics* **114**: 739–767.

Gabaix X. 2009. Power laws in economics and finance. *Annual Review of Economics* **1**: 255–293.

Gabaix X, Gopikrishnan P, Plerou V, Stanley HE. 2006. Institutional investors and stock market volatility. *The Quarterly Journal of Economics* **121**: 461–504.

Gabaix X, Ibragimov R. 2011. Rank – 1/2: A simple way to improve the OLS estimation of tail exponents. *Journal of Business & Economic Statistics* **29**: 24–39.

Gneiting T. 2011. Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**: 746–762.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**: 359–378.

Gneiting T, Ranjan R. 2011. Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* **29**: 411–422.

Gonzalo J, Olmo J. 2004. Which extreme values are really extreme? *Journal of Financial Econometrics* **2**: 349–369.

Gu Z, Ibragimov R. 2018. The "cubic law of the stock returns" in emerging markets. *Journal of Empirical Finance* **46**: 182–190.

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

Hartmann P, Straetmans S, de Vries CG. 2006. Banking system stability: A cross-atlantic perspective. In Carey M, Stulz RM (eds.) *The Risks of Financial Institutions*, Chicago and London: The University of Chicago Press, 1 edn., pages 133–193.

Heffernan JE. 2000. A directory of coefficients of tail dependence. *Extremes* **3**: 279–290.

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

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

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

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

Hoga Y. 2018a. Detecting tail risk differences in multivariate time series. *Journal of Time Series Analysis* **39**: 665–689.

Hoga Y. 2018b. A structural break test for extremal dependence in β-mixing random vectors. *Biometrika* **105**: 627–643.

Hoga Y. 2019a+. Confidence intervals for conditional tail risk measures in ARMA–GARCH models. Forthcoming in *Journal of Business & Economic Statistics* (DOI: 10.1080/07350015.2017.1401543) **XX**: 1–12.

Hoga Y. 2019b+. Extreme conditional tail moment estimation under serial dependence. Forthcoming in *Journal of Financial Econometrics* (DOI: 10.1093/jjfinec/nby016) **XX**: 1–29.

Holzmann H, Klar B. 2017. Discussion of "Elicitability and backtesting: Perspectives for banking regulation". *The Annals of Applied Statistics* **11**: 1875–1882.

Hua L, Joe H. 2011. Second order regular variation and conditional tail expectation of multiple risks. *Insurance: Mathematics and Economics* **49**: 537–546.

Huber PJ, Ronchetti EM. 2009. *Robust Statistics*. Hoboken: Wiley, 2 edn.

Jalal A, Rockinger M. 2008. Predicting tail-related risk measures: The consequences of using GARCH filters for non-GARCH data. *Journal of Empirical Finance* **15**: 868–877.

Koenker R, Bassett G. 1978. Regression quantiles. *Econometrica* **46**: 33–50.

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

Laurini F, Tawn JA. 2008. Regular variation and extremal dependence of GARCH residuals with application to market risk measures. *Econometric Reviews* **28**: 146–169.

Ledford AW, Tawn JA. 1996. Statistics for near independence in multivariate extreme values. *Biometrika* **83**: 169–187.

Ledford AW, Tawn JA. 1997. Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**: 475–499.

Longin F (ed.) . 2017. *Extreme Events in Finance: A Handbook of Extreme Value Theory and its Applications*. Hoboken: Wiley.

McNeil AJ, 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.

Mikosch T, Stărică C. 2000. Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process. *Annals of Statistics* **28**: 1427–1451.

Müller UK, Wang Y. 2017. Fixed-k asymptotic inference about tail properties. *Journal of the American Statistical Association* **112**: 1334–1343.

Nolde N, Zhang J. 2019+. Conditional extremes in asymmetric financial markets. *Journal of Business & Economic Statistics* : 1–13.

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

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

Poon SH, Rockinger M, Tawn J. 2004. Extreme value dependence in financial markets: Diagnostics, models, and financial implications. *Review of Financial Studies* **17**: 581–610.

Quintos C, Fan Z, Phillips PCB. 2001. Structural change tests in tail behaviour and the Asian crisis. *Review of Economic Studies* **68**: 633–663.

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

Resnick S, Stărică C. 1997. Smoothing the Hill estimator. *Advances in Applied Probability* **29**: 271–293.

Sibuya M. 1960. Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics* **11**: 195–210.

Straetmans S, Chaudhry SM. 2015. Tail risk and systemic risk of US and Eurozone financial institutions in the wake of the global financial crisis. *Journal of International Money and Finance* **58**: 191–223.

Straetmans STM, Verschoor WFC, Wolff CCP. 2008. Extreme US stock market fluctuations in the wake of 9/11. *Journal of Applied Econometrics* **23**: 17–42.

Sun P, de Vries CG. 2018. Exploiting tail shape biases to discriminate between stable and Student t alternatives. *Journal of Applied Econometrics* **33**: 708–726.

Taylor JW. 2019. Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *Journal of Business & Economic Statistics* **37**: 121–133.

Trapani L. 2016. Testing for (in)finite moments. *Journal of Econometrics* **191**: 57–68.

Wagner N, Marsh TA. 2005. Measuring tail thickness under GARCH and an application to extreme exchange rate changes. *Journal of Empirical Finance* **12**: 165–185.

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

Ziegel JF, Krüger F, Jordan A, Fasciati F. 2019. Robust forecast evaluation of expected shortfall. Forthcoming in *Journal of Financial Econometrics* (DOI: 10.1093/jjfinec/nby035) **XX**: 1–26.

Online Supplemental Appendices

In Appendix A, we complement the simulation results in Section 3 of the main paper. To show that our method of choosing k also works well for other tail index estimators than the Hill (1975) estimator, we re-run the Monte Carlo simulations for the popular log-log rank-size estimator due to Gabaix and Ibragimov (2011). These results are reported in Appendix B.

Appendix A Complementary Simulations

A.1 Estimation of Expected Shortfall

The arguably two most important risk measures in the financial industry are the Value-at-Risk and the Expected Shortfall (ES). The ES at level p is the average loss given that VaR at level p (VaR_p) is exceeded; formally, $\text{ES}_p = \text{E}[X \mid X > \text{VaR}_p]$. Hua and Joe (2011) show that under (1) ES and VaR behave similarly far out in the tail, viz.

$$\frac{\text{ES}_p}{\text{VaR}_p} \to \frac{1}{1-\gamma}, \qquad \text{as } p \to \infty.$$

This motivates the estimator $\widehat{\text{ES}}_p(k) = \widehat{x}_p(k)/(1-\widehat{\gamma}(k))$ studied by Hoga (2019b+).

Figure A.1 is the analogue of Figure 1 in the main paper for ES estimation via $\widehat{\text{ES}}_p(k)$. It confirms the finding that risk measure estimates may be significantly improved by using k^{QCRPS}. For panels (a1), (b1), (b2) and (c2), where VaR estimates are more precise throughout, the improvements for ES estimates are even more pronounced.

A.2 Additional Results for VaR Estimation

In the main paper, Figure 1 only compares the VaR estimates $\widehat{x}_p(k^{\text{QCRPS}})$ and $\widehat{x}_p(k^Q)$ for models **(M1)**, **(M2)** and **(M6)**. Here, we provide similar figures for the remaining models **(M3)**–**(M5)** and **(M7)**–**(M8)**. With RMSE ratios mostly above one, Figure A.2 confirms the good performance of $\widehat{x}_p(k^{\text{QCRPS}})$ for models **(M3)**–**(M5)**. The main patterns of Figure 1 also emerge again. The advantage of $\widehat{x}_p(k^{\text{QCRPS}})$ vis-à-vis $\widehat{x}_p(k^Q)$ is often particularly marked for 'small' $p \in \{0.90, 0.95, 0.975\}$ and 'large' $p \in \{0.9995, 0.9999\}$. Additionally, the advantages become larger, the larger the sample size n and the smaller α.

Figure A.3 shows the RMSE ratios for the remaining models **(M7)**–**(M8)**. Table 1 shows that these models present challenges for k^{QCRPS} for larger α. This is reflected in the VaR estimates $\widehat{x}_p(k^{\text{QCRPS}})$, which are often inferior to $\widehat{x}_p(k^Q)$ for large p. It is perhaps a bit surprising to find that this is reversed

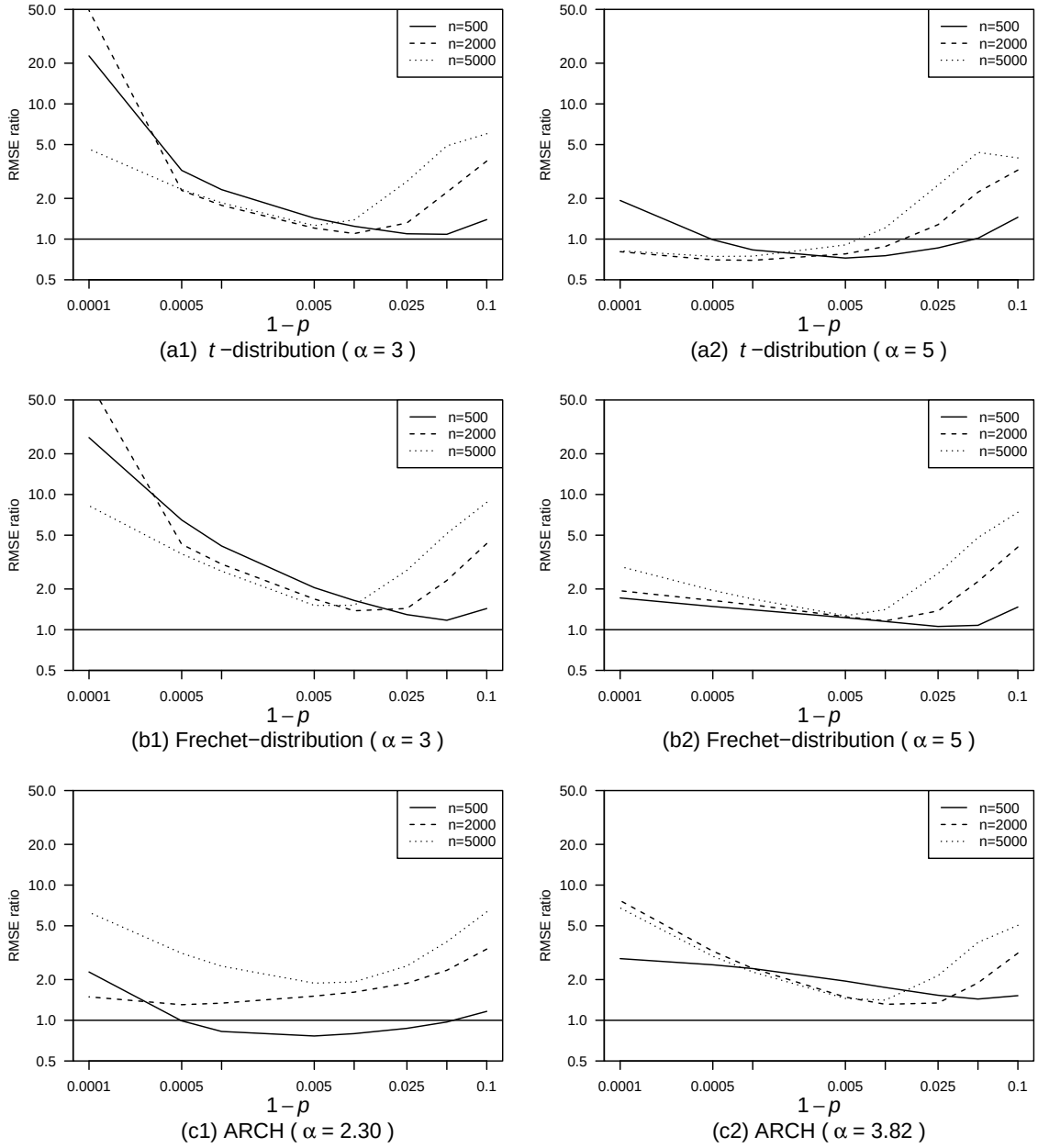

Figure A.1: Log-log plots of $\mathrm{RMSE}_p(k^Q)/\,\mathrm{RMSE}_p(k^{\mathrm{QCRPS}})$ for different values of $1-p$. Here, $\mathrm{RMSE}_p(k)$ is the RMSE of $\widehat{\mathrm{ES}}_p(k)$ ($k \in \{k^Q, k^{\mathrm{QCRPS}}\}$). Results shown for models **(M1)**, **(M2)**, and **(M6)** with indicated α.

for 'small' $p \in \{0.90, 0.95, 0.975\}$. The reason for this is that k^Q is—on average—much smaller than k^{QCRPS}. For instance, for the model in panel (b1) the average k^Q is 9.5, 15, 21 for $n = 500$, 2000, 5000, respectively. Thus, it does not make sense to use $\widehat{x}_p(k^Q)$ for $p \leq 1 - k^Q/n = 0.981, 0.9925, 0.9958$, which amounts to estimating a more extreme $(1 - k^Q/n)$-quantile and then extrapolating 'backwards' to the desired less extreme p-quantile. This reverses the logic of the Weissman (1978) estimator. Thus,

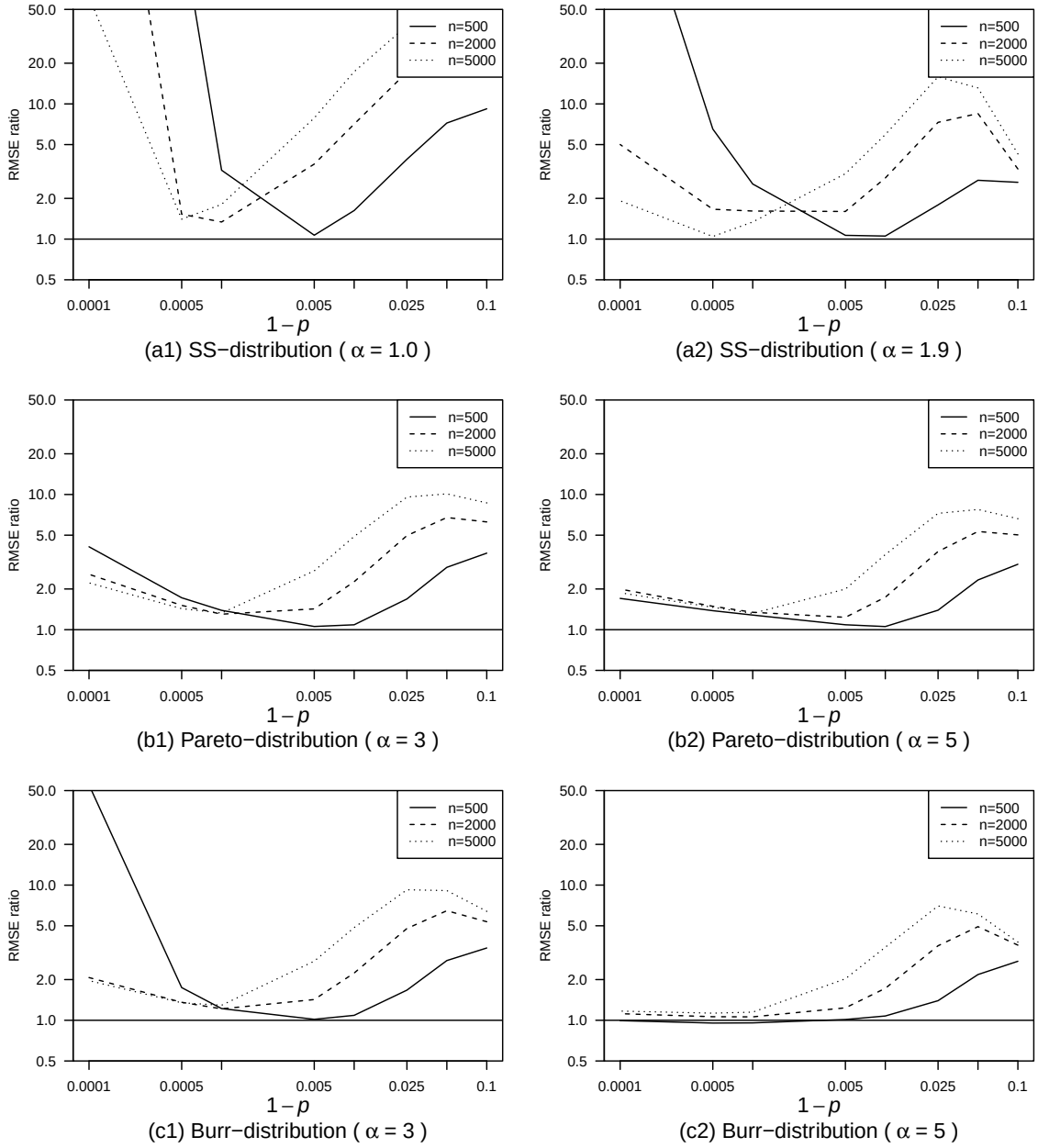

Figure A.2: Log-log plots of $\mathrm{RMSE}_p(k^Q)/\,\mathrm{RMSE}_p(k^{\mathrm{QCRPS}})$ for different values of $1-p$. Here, $\mathrm{RMSE}_p(k)$ is the RMSE of $\widehat{x}_p(k)$ ($k \in \{k^Q, k^{\mathrm{QCRPS}}\}$). Results shown for models **(M3)**–**(M5)** with indicated α.

$\widehat{x}_p(k^Q)$ is at a disadvantage when estimating quantiles at 'small' levels $p \in \{0.90, 0.95, 0.975\}$.

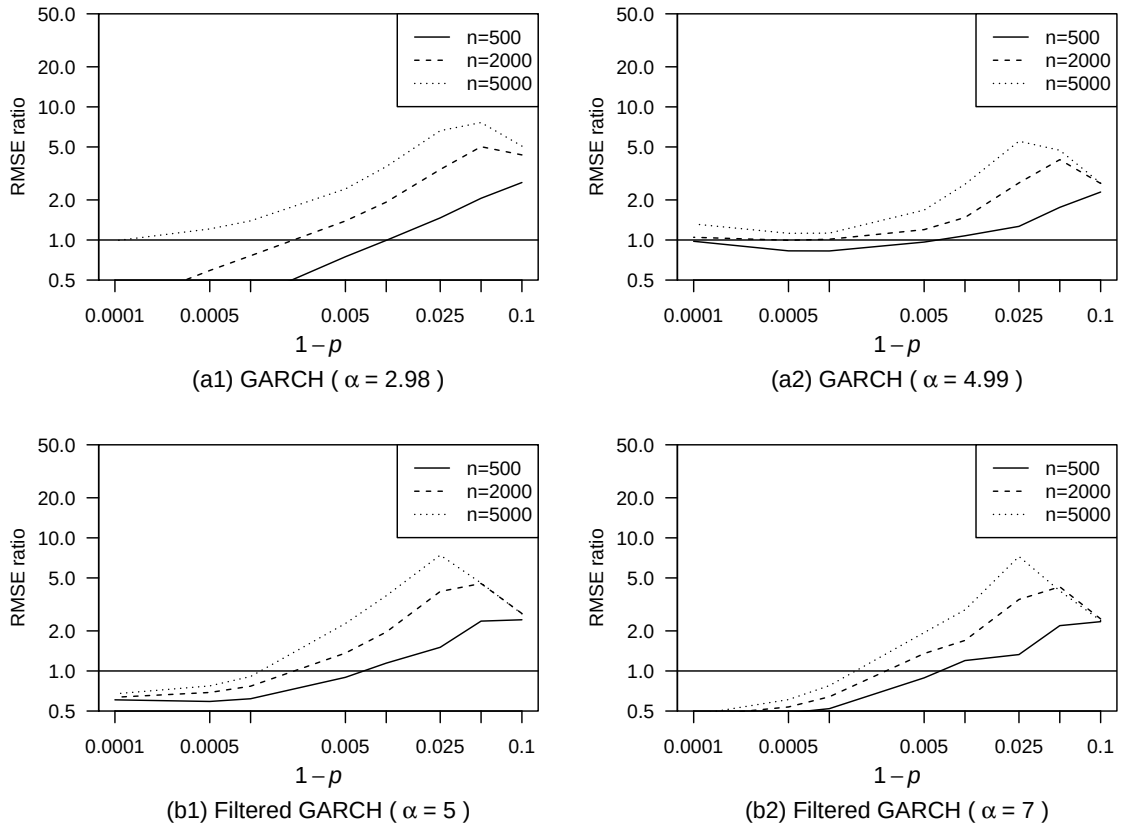

Figure A.3: Log-log plots of $\mathrm{RMSE}_p(k^Q)/\mathrm{RMSE}_p(k^{\mathrm{QCRPS}})$ for different values of $1-p$. Here, $\mathrm{RMSE}_p(k)$ is the RMSE of $\widehat{x}_p(k)$ ($k \in \{k^Q, k^{\mathrm{QCRPS}}\}$). Results shown for models **(M7)**–**(M8)** with indicated α.

Appendix B Simulations with the Log-Log Rank-Size Estimator

To introduce the log-log rank-size estimator of Gabaix and Ibragimov (2011), we again consider $X_1, \ldots, X_n$ with common d.f. $F(\cdot)$ satisfying (1), i.e.,

$$1 - F(x) = x^{-1/\gamma} L(x) \tag{B.1}$$

for $\gamma > 0$ and slowly varying $L(\cdot)$. As before, the ordered observations are denoted by $X_{(1)} \geq \ldots X_{(n)}$. The index i in $X_{(i)}$ is consequently a rank. Now apply the logarithm to each side of (B.1) and insert $x = X_{(i)}$, the empirical $(1 - i/n)$-quantile, to obtain

$$\log\left(\frac{i}{n}\right) \approx \log(L(X_{(i)})) - \frac{1}{\gamma}\log(X_{(i)}). \tag{B.2}$$

This relation motivates the regression of log-ranks on log-size in

$$\log(i - 1/2) = a - b\log(X_{(i)}). \tag{B.3}$$

Gabaix and Ibragimov (2011) explain the reason for shifting the ranks by 1/2. An estimate of γ in (B.2) can be obtained from (B.3) via the inverse of the OLS estimate of b:

$$\widehat{b}(k) = -\frac{\sum_{i=1}^{k}(x_i - \overline{x}_k)(y_i - \overline{y}_k)}{\sum_{i=1}^{k}(x_i - \overline{x}_k)^2},$$

where $x_i = \log(X_{(i)})$, $y_i = \log(i - 1/2)$, $\overline{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$, $\overline{y}_k = \frac{1}{k}\sum_{i=1}^{k} y_i$ and k denotes the number of tail observations for which (B.1) approximately holds.

The choice of k in the extreme value index estimate $\widehat{\gamma}(k) = 1/\widehat{b}(k)$ is again crucial. In the remainder of this appendix we compare the two choices k^Q and k^{QCRPS}. Tables B.1–B.4 are the analogues of Tables 1–4 for the log-log rank-size estimator.

We draw the following conclusions from Tables B.1–B.4:

1. RMSE: Table B.1 presents even stronger evidence in favour of k^{QCRPS} than Table 1. With a few exceptions, notably for the t-distribution and the filtered GARCH innovations, using k^{QCRPS} leads to more precise estimates of γ. Again, the advantage tends to be larger, the heavier the tail. Comparing the RMSEs of the Hill estimator in Table 1 with those of the log-log rank-size estimator in Table B.1, we find that most of the time the Hill estimator delivers more precise estimates for both choices k^Q and k^{QCRPS}.

2. Bias: On balance, we find that bias is lower for $\widehat{\gamma}(k^{\mathrm{QCRPS}})$ than for $\widehat{\gamma}(k^Q)$ in Table B.2. The bias of $\widehat{\gamma}(k^{\mathrm{QCRPS}})$ is particularly low for very heavy-tailed data, whereas for $\widehat{\gamma}(k^Q)$ this is true for lighter tails. In comparison with Table 2, bias tends to be smaller for the Hill estimator, but

Model	α	$n = 500$		$n = 1000$		$n = 2000$		$n = 5000$	
		k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}
t_α	1	85.0	**29.1**	81.3	**21.5**	80.9	**17.2**	81.3	**13.2**
	3	24.5	**10.7**	24.0	**8.6**	21.7	**6.9**	17.3	**5.3**
	5	14.7	**11.9**	12.4	**9.8**	8.8	**8.1**	**6.3**	6.5
	7	**12.4**	13.4	**9.8**	11.3	**7.6**	9.6	**5.6**	7.9
Fréchet	1	83.9	**27.5**	80.7	**21.4**	82.1	**17.3**	81.3	**13.3**
	3	25.0	**8.2**	23.9	**6.6**	21.3	**5.4**	16.1	**4.1**
	5	11.9	**5.0**	8.9	**3.9**	7.0	**3.2**	5.1	**2.4**
	7	6.1	**3.4**	5.3	**2.9**	3.3	**2.2**	2.7	**1.7**
SS	0.5	169	**106**	172	**101**	165	**85.4**	161	**66.9**
	1.0	82.4	**28.3**	81.0	**20.9**	81.5	**17.4**	81.8	**13.2**
	1.5	53.9	**17.5**	55.6	**14.3**	52.6	**11.7**	53.0	**8.7**
	1.9	40.0	**24.3**	42.6	**23.4**	42.0	**22.1**	36.7	**19.8**
Pareto	1	82.3	**27.0**	82.8	**21.5**	80.2	**17.0**	79.9	**13.1**
	3	24.2	**8.1**	23.3	**6.5**	21.4	**5.4**	16.1	**4.0**
	5	12.1	**4.8**	9.5	**3.9**	6.8	**3.2**	5.2	**2.3**
	7	6.2	**3.4**	4.3	**2.8**	3.4	**2.2**	2.8	**1.7**
Burr	1	80.4	**26.3**	82.4	**22.2**	81.0	**17.3**	80.2	**12.9**
	3	25.1	**8.8**	23.7	**7.2**	21.8	**5.7**	17.3	**4.3**
	5	13.2	**7.5**	10.7	**6.3**	7.3	**5.2**	5.4	**4.1**
	7	9.7	**7.9**	7.1	**6.8**	**5.3**	5.8	**4.0**	4.8
ARCH	2.30	19.1	**13.5**	18.3	**11.9**	17.7	**10.3**	17.1	**8.3**
	2.68	16.8	**11.7**	15.9	**10.1**	15.2	**8.6**	14.8	**6.9**
	3.17	14.8	**10.5**	13.7	**8.8**	12.4	**7.4**	11.5	**5.9**
	3.82	12.4	**9.5**	11.9	**8.1**	9.8	**6.7**	9.3	**5.3**
GARCH	2.03	19.9	**16.5**	19.7	**16.0**	19.5	**15.1**	19.6	**13.9**
	2.98	11.1	**10.6**	10.6	**9.6**	10.6	**8.5**	10.4	**7.5**
	3.96	9.7	**9.4**	8.9	**7.9**	8.2	**6.6**	7.5	**5.4**
	4.99	10.0	**9.5**	8.2	**7.8**	7.1	**6.6**	6.1	**5.2**
Filtered	5	12.3	**11.2**	10.1	**9.4**	**7.5**	7.8	**6.0**	6.4
GARCH	7	**11.5**	13.0	**9.2**	11.0	**7.1**	9.5	**5.5**	7.8

Table B.1: RMSE($\times 10^{-2}$) of $\widehat{\gamma}(k)$ with k chosen as k^Q or k^{QCRPS} for models **(M1)**–**(M8)** with true tail index α. Lower values for RMSE are set in boldface.

the differences are slight.

3. Average number of k: Table B.3 and Table 3 show that the average k^{QCRPS} is slightly higher for the log-log rank size estimator than for the Hill estimator. For instance, for the GARCH model and $n = 5000$, the average k^{QCRPS} is between 106 and 113 for the Hill estimator, and 110 and 121 for the log-log rank-size estimator, while the respective numbers for k^Q are 13 and 22, and

Model	α	$n = 500$		$n = 1000$		$n = 2000$		$n = 5000$	
		k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}
t_α	1	21.3	**4.3**	19.5	**2.9**	19.5	**2.3**	18.3	**1.6**
	3	8.2	**6.9**	6.7	**5.5**	**4.2**	4.3	**2.2**	3.3
	5	**8.5**	10.4	**6.3**	8.7	**4.3**	7.3	**2.7**	5.9
	7	**9.9**	12.4	**7.7**	10.7	**6.0**	9.1	**4.4**	7.6
Fréchet	1	20.5	**4.5**	19.0	**3.3**	18.2	**2.3**	18.9	**1.7**
	3	5.6	**1.3**	4.4	**0.9**	2.4	**0.7**	**0.4**	0.5
	5	1.7	**0.8**	**0.3**	0.6	**−0.3**	0.4	−0.8	**0.3**
	7	**0.3**	0.5	**−0.0**	0.4	−0.4	**0.3**	−0.7	**0.2**
SS	0.5	31.8	**20.1**	34.4	**18.4**	30.0	**12.2**	29.9	**7.6**
	1.0	19.5	**4.1**	19.9	**2.7**	19.5	**2.5**	18.6	**1.6**
	1.5	12.4	**−4.0**	13.1	**−3.3**	12.3	**−2.9**	10.9	**−2.3**
	1.9	**−10.5**	−21.6	**−6.5**	−21.0	**−5.1**	−20.0	**−6.3**	−18.2
Pareto	1	19.6	**3.6**	20.0	**3.0**	18.7	**1.8**	18.9	**1.6**
	3	5.1	**0.8**	4.4	**0.9**	2.5	**0.5**	0.2	**0.3**
	5	1.4	**0.5**	**0.2**	0.4	−0.3	**0.3**	−0.8	**0.2**
	7	**0.2**	0.3	**−0.2**	0.3	−0.5	**0.2**	−0.7	**0.1**
Burr	1	19.4	**3.3**	18.7	**2.6**	18.9	**2.1**	17.8	**1.2**
	3	6.3	**3.3**	5.1	**2.7**	3.5	**2.2**	**1.2**	1.6
	5	**5.2**	5.4	**3.5**	4.7	**2.1**	3.9	**1.0**	3.2
	7	**5.7**	6.8	**4.3**	6.0	**3.2**	5.2	**2.3**	4.4
ARCH	2.30	−2.5	**1.3**	−3.1	**1.3**	−3.5	**1.0**	−4.4	**0.8**
	2.68	**−0.0**	2.7	**−0.9**	2.4	**−1.7**	2.0	−2.2	**1.5**
	3.17	**1.9**	4.2	**0.6**	3.5	**−0.2**	2.9	**−1.0**	2.4
	3.82	**3.3**	5.5	**2.4**	4.8	**0.9**	3.9	**0.2**	3.1
GARCH	2.03	−16.5	**−9.0**	−17.0	**−7.3**	−16.9	**−5.7**	−17.1	**−3.9**
	2.98	−3.6	**1.1**	−4.6	**0.9**	−5.1	**0.6**	−5.2	**0.9**
	3.96	**1.7**	4.9	**0.7**	4.1	**−0.1**	3.3	**−0.9**	2.7
	4.99	**4.8**	7.2	**3.2**	5.9	**2.1**	4.9	**1.2**	4.0
Filtered	5	**7.4**	9.8	**5.4**	8.3	**3.8**	7.0	**2.5**	5.8
GARCH	7	**9.3**	12.1	**7.4**	10.4	**5.8**	9.0	**4.3**	7.5

Table B.2: Bias($\times 10^{-2}$) of $\widehat{\gamma}(k)$ with k chosen as k^Q or k^{QCRPS} for models **(M1)**–**(M8)** with true tail index α. Lower absolute values for bias are set in boldface.

23 and 40. Clearly, the classification as a tail observation (through the choice of k) should not depend on the particular estimator being used by the practitioner. Thus, the choice k^Q seems inferior, as it is much more influenced by the idiosyncrasies of the tail index estimator.

4. Standard deviation of k's: The standard deviations reported in Table B.4 are increased relative to those in Table 4, particularly so for large n and large α. Again, the standard deviations are

Model	α	$n = 500$		$n = 1000$		$n = 2000$		$n = 5000$	
		k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}
t_α	1	7.0	28	8.6	44	10	66	16	113
	3	13	29	18	44	24	66	38	112
	5	16	28	22	43	30	63	42	105
	7	18	28	23	42	31	62	43	104
Fréchet	1	7.1	28	8.5	44	10	65	15	113
	3	12	31	17	47	26	69	43	117
	5	16	31	24	46	34	69	53	117
	7	18	31	26	46	37	69	57	116
SS	0.5	5.9	21	7.4	29	8.5	42	11	64
	1.0	7.1	29	8.5	44	10	65	15	113
	1.5	8.4	30	10	46	14	68	25	114
	1.9	15	29	23	47	38	73	75	121
Pareto	1	7.0	29	8.6	44	10	66	16	114
	3	12	30	17	47	25	70	42	118
	5	16	31	24	47	34	69	54	117
	7	18	31	26	46	37	69	57	117
Burr	1	6.9	29	8.5	44	10	66	15	113
	3	12	30	17	46	25	68	41	117
	5	16	30	22	45	32	67	47	113
	7	18	29	24	44	33	66	46	111
ARCH	2.30	13	29	16	45	20	67	28	115
	2.68	13	29	17	45	22	68	31	115
	3.17	14	29	18	45	24	67	33	115
	3.82	15	29	20	44	26	66	36	114
GARCH	2.03	14	29	17	45	20	69	23	121
	2.98	15	29	19	45	23	67	29	115
	3.96	16	28	21	44	26	65	35	113
	4.99	17	28	22	43	29	64	40	110
Filtered	5	16	28	23	42	29	63	41	106
GARCH	7	17	28	24	42	31	62	43	103

Table B.3: Average number of k used in $\widehat{\gamma}(k)$ with k chosen as k^Q or k^{QCRPS} for models **(M1)**–**(M8)** with true tail index α.

smaller for k^{QCRPS} than for k^Q relative to their average values. For instance, while the standard deviations of k^Q and k^{QCRPS} are 22 and 37 for the GARCH($\alpha = 2.03$) models with $n = 5000$, the respective average values are 23 and 121. This implies a much higher relative standard deviation of $22/23 = 0.96$ for k^Q than for k^{QCRPS}, where $27/121 = 0.31$.

The simulation results in Tables B.1–B.4 further strengthen the case for the choice k^{QCRPS}. Not only does k^{QCRPS} lead to more precise estimates of γ again, but this choice also classifies roughly the

Model	α	$n = 500$		$n = 1000$		$n = 2000$		$n = 5000$	
		k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}	k^Q	k^{QCRPS}
t_α	1	8.4	9.7	12	15	17	24	32	43
	3	10	10	16	16	25	26	46	49
	5	10	10	15	16	24	26	41	48
	7	9.5	10	14	16	21	25	36	47
Fréchet	1	8.6	9.7	12	15	17	24	30	43
	3	10	9.9	17	16	28	26	52	48
	5	11	10	19	16	30	26	54	49
	7	11	10	19	16	30	26	54	50
SS	0.5	8.4	10	12	17	15	26	26	45
	1.0	8.4	9.7	12	15	18	24	31	43
	1.5	9.2	9.7	13	16	22	25	43	47
	1.9	12	10	22	16	38	24	70	43
Pareto	1	8.4	9.6	12	15	18	24	31	43
	3	10	10	16	16	27	25	51	48
	5	11	10	19	16	30	26	54	49
	7	12	10	19	17	30	26	54	50
Burr	1	8.4	9.6	12	15	18	24	30	43
	3	10	10	16	16	27	26	49	48
	5	10	10	17	16	27	26	47	49
	7	10	10	16	16	25	26	43	49
ARCH	2.30	9.5	9.9	14	15	21	25	35	46
	2.68	9.7	9.8	14	16	21	25	37	46
	3.17	10	9.9	15	16	22	25	39	47
	3.82	10	10	15	16	23	25	40	47
GARCH	2.03	8.5	9.1	11	14	16	22	22	37
	2.98	9.1	9.5	13	15	19	24	31	45
	3.96	9.6	9.8	14	15	21	25	36	46
	4.99	9.7	10	15	16	23	25	39	47
Filtered	5	10	10	15	16	23	26	40	48
GARCH	7	9.3	10	14	16	21	26	36	47

Table B.4: Standard deviation of k's used in $\widehat{\gamma}(k)$ with k chosen as k^Q or k^{QCRPS} for models **(M1)**–**(M8)** with true tail index α.

same number of observations as belonging to the tail for different tail index estimators.

The analogues of Figures 1–2 and Figures A.1–A.3 are largely similar for the log-log rank-size estimator and add no additional insight. Hence, we omit these plots, which are available upon request.