Unbiased Estimation of Tail Properties in Small Samples with Complete, Censored, or Truncated Data

Yulong Wang*

Syracuse University

Department of Economics and Center of Policy Research

January 19, 2018

Abstract

This paper considers estimating tail properties such as tail index, extreme quantile and tail conditional expectation with small sample size, say only 250 observations. We provide new asymptotically (quantile) unbiased estimators that are applicable to (i) complete data; (ii) tail censored (top-coded) data with known or unknown censoring value; and (iii) tail truncated data with known or unknown truncation value. The new method only requires regularly varying tails and delivers excellent small sample bias and risk properties as shown by Monte Carlo simulations. The empirical relevance is illustrated by estimating (i) the tail index of macroeconomic disasters as studied by Barro and Jin (2011) and (ii) extreme quantiles of the U.S. earthquake fatality.

Keywords: catastrophe, macroeconomic disaster, tail index, extreme quantile, tail conditional expectation, extreme value theory, heavy-tailed distribution, power law

*Yulong Wang is assistant professor of economics, Syracuse University, Syracuse, NY 13210 (e-mail: ywang402@maxwell.syr.edu). The author thanks Ulrich Müller, Bo Honoré, Mark Watson, Kirill Evdokimov, Xiaohong Chen, and seminar participants at Princeton, Syracuse, and SUNY-Albany for inspiring discussions and helpful comments. Working in progess, please do not circulate.

1 Introduction

Estimating tail properties such as tail index, extreme quantile, and tail conditional expectation (TCE) has been an important issue in economics and finance (see, for example, McNeil and Frey (2000), Engle and Manganelli (2004), Kuester, Mittnik, and Paolella (2006), Jorion (2007), Bollerslev and Todorov (2011a,b), Fissler and Ziegel (2016), and Patton, Ziegel, and Chen (2017)). Due to the limited numer of observations in the tail, existing methods typically require a large sample and completely observed data, which are unavailable in many empirical applications such as natural catastrophe and macroeconomic disaster. This paper focuses on the small sample situation, say only 100 observations, and develops new estimators that are unbiased and optimal in a well-defined sense. Furthermore, they are tailored to cover incomplete data due to censoring or truncation.

When the data are complete, a large number of estimators have been developed based on the extreme value theorem and tail regularity conditions. See Embrechts, Klupperberg, and Mikosch (1997), Reiss and Thomas (2007), Resnick (2007), and de Haan and Ferreira (2007), Beirlant, Caeiro, and Gomes (2012), Gomes and Guillou (2015) for reviews and references. Given the assumption that the underlying distribution F is regularly varying, its tail can be well approximated by a generalized Pareto distribution (cf. Pickands (1975)), and then common tail properties of interest, including high quantile and TCE, are functions of three parameters only, at least approximately. These parameters include the tail index ξ which is the exponential component, the scale, and the location. Along this line, numerous suggestions have been made about estimating these parameters, and the corresponding estimators of high quantile and TCE can be constructed by plugging in estimators of these parameters.

One concern of the above mentioned methods is that they rely on the "increasing-k" asymptotics, which, however, may lead to a poor small sample approximation when the sample size is only moderate. More specifically, methods in the existing literature require the asymptotics under $k \to \infty$ and $k/n \to 0$ where k is the number of tail observations and n is the sample size. So given a certain sample, the choice of k can be difficult to keep the delicate balance that (i) k has to be large enough for the asymptotics to hold on estimating ξ; and (ii) k has to be so small relative to n that the largest k observations are well approximated by Pareto. Therefore, in many empirical applications such as financial daily data collected from one year, there could be no choice of k that results in a satisfactory small sample performance (cf. Diebold, Schuermann, and Stroughair (2000)).

In addition to the concern about choosing k, the above mentioned methods cannot be applied to incomplete data due to censoring or truncation. In particular, data censoring usually exists in surveys about earnings and wealth such as the Current Population Survey, and the problems subject to data truncation can be found in finance, hydrology, fire ecology, and seismology (see, for example, Groisman, Knight, Karl, Easterling, Sun, and Lawrimore (2004) and Malamud, Morein, and Turcotte (1998)). In particular, recent literature on the macroeconomic disasters (see Barro and Jin (2011) and reference therein) concerns that the largest observations of disaster size (measured by percentage of GDP or consumption decline) can be missing due to government collapse or fighting a war. Estimating tail properties with such incomplete data is even more difficult, since the largest observations are very informative about tail but unfortunately unavailable. Existing literature typically makes parametric assumptions on the whole distribution (cf. Aban, Meerschaert, and Panorska (2006) and Jenkins, Burkhauser, Feng, and Larrimore (2010)) but this approach may suffer severe misspecification. Recently, Beirlant, Alves, and Gomes (2016) and Zou, Davis, and Samorodnitsky (2017) develop estimators for truncated data and they also require a large sample for satisfactory approximation of their asymptotics.

To overcome the above issues, this paper develops a unified framework to accommodate all three type of data and considers the "fixed-k" asymptotic embedding developed by Müller and Wang (2017) under the sole assumption that the k largest order statistics jointly converge to k jointly extreme value distributed variables, for fixed and given k. This means we only require a fixed number of tail observations are approximately stemming from a (generalized) Pareto while leaving the main body of the underlying distribution unspecified. Consequently, it is asymptotically equivalent to deal with a small sample problem in which we are estimating a quantity as a function of the underlying distribution based on k observations. Regarding incomplete data, we model censoring as that in the sample, the largest m observations ($m < k$) are unobserved, for a known and fixed m, and model truncation as that approximately only a fixed number of the largest draws are unavailable. In neither case does the fixed-k method require the knowledge of the censoring or truncation value. We start with an i.i.d. sample and then extend the results to stochastic volatility models by establishing that the innovations of fitting a GARCH model still satisfy the joint extreme value theorem.

To precisely capture the difficulty in small samples, we do not aim for consistency but focus instead on the unbiasedness, which is well motivated since it is naturally embedded in the definition of quantile and TCE. In particular, in estimating the quantile of the underlying distribution F, it makes sense to require the estimator of the p quantile, $\hat{Q}(p)$, to satisfy the quantile unbiasedness: $P\left(Y_i > \hat{Q}(p)\right) = 1 - p$ where Y_i is another independent draw from

F. For TCE, we show that the mean unbiasedness is equivalent to some average quantile unbiasedness that measures both the size and the likelihood of the tail above a certain confidence level. This is especially important in finance as it corresponds to the reason why Basel III suggests switching from value at risk (VaR) to expected shortfall (ES) for measuring financial risk. See Basel Committee on Banking Supervision (2013) for more details. For the tail index, we impose median unbiasedness which has been studied by the influential works by Andrews (1993) and Stock and Watson (1998) in estimating the autoregressive coefficient and the coefficient of time variation, respectively. Given the unbiasedness constraint, we then set up and solve a Lagrangian problem to find the optimal estimator that minimizes a weighted average risk criteria and satisfies the corresponding constraint.

As a summary, the new approach has four advantages comparing with existing methods. First, the fixed-k asymptotics provides excellent small sample approximations as shown by Monte Carlo experiments with moderate sample size. Second, our estimator requires no parametric assumption on the underlying distributions and hence is robust to deviation from the power law, which is commonly used. This robustness is valuable since it is difficult to test out a specific parametric distribution in small samples by any goodness-of-fit test (Chernobai, Rachev, and Fabozzi (2012)). Third, our estimator is invariant/equivariant to data shifting. As pointed out by Alves, Gomes, de Haan, and Neves (2009), popular estimators of the tail index such as the Hill (1975) estimator and the log-log regression estimator (see, for example, Gabaix and Ibragimov (2011)) are sensitive to data shifting, especially when the underlying distribution is not exactly Pareto. Finally, the Lagrangian structure gurantees that the new approach is nearly optimal in the sense of minimizing a weighted average risk criteria among all uniformly unbiased estimators that are invariant/equivariant to data shifting. However, imposing invariance/equavariance to data shifting is costly in terms of the risk, suggesting that the knowledge of the location is very informative.

We illustrate the empirical relevance of our approach with two applications. The first is to estimate the tail index of the size distribution of macroeconomic disasters. Barro and Ursua (2008) and Barro and Jin (2011) define a macroeconomic disaster if the consumption/GDP declines by more than 10%. The authors collect a dataset of 157 observations, fit them with a double power law, and claim that only the very large observations are mainly informative about the tail behavior. The parametric assumption is subject to misspecification and the largest observations can be missing due to goverment collapse or fighting wars. We apply the new approach allowing possible data truncation and obtain a substantively different tail index estimate, resulting in a larger coefficient of relative risk aversion, 8.59 instead of 4.

Second, we examine the U.S. earthquake fatality data and estimate the large quantiles, which are important from economic persepctive (see, for example, Anbarci, Escaleras, and Register (2005) and Kahn (2005)). In the earthquake fatality exercise, data might be truncated due to complicated reasons (Burroughs and Tebbens (2001, 2002) and Clark (2013)). Our approach works well for truncated data as verified by Monte Carlos and delivers a much larger estimate of the high quantiles than the increasing k method proposed by Beirlant, Alves, and Gomes (2016).

The rest of the paper is organized as follows. Section 2 illustrate the new approach with estimating extreme quantiles, including deriving the fixed-k asymptotics, setting up and solving the Lagrangian problem for three data categories. Section 3 discusses estimating TCE and tail index and Section 4 reports Monte Carlo simulations. Section 5 applies the new approach to two empirical examples and Section 7 concludes. All the proofs and computational details are included in the Appendix.

2 Extreme Quantile

2.1 Complete Data

We start with estimating a high quantile based a random sample $Y_1, Y_2, ..., Y_n$ drawn from a certain cumulative distribution function (CDF) F. To capture the fact that we only have limited information about the tail, we focus on estimating the $1-h/n$ quantile of F, denoted by $Q(F, 1-h/n)$, for a given and fixed h. This indicates that the object of interest is of the same order of magnitude as the sample maximum and thus cannot be consistently estimated. Typical choices of h can be 0.1, 1, 5, and 10, corresponding to the quantile at levels 99.98%, 99.8%, 99%, and 98% for a sample of 500.

There is a naturally embedded quantile unbiasedness constraint on the estimator $\hat{Q}$, that is $E\left[P\left(Y_i > \hat{Q}\right)\right] = h/n$ for an independent draw Y_i from F. That says the violation probability that another independent draw is larger than p-th quantile should be exactly $1-p$. Hence our objective is to construct the optimal $\hat{Q}$ that satisfies such quantile unbiasedness restriction, at least asymptotically.

To avoid assuming an increasing k that may lead to a poor finite sample approximation (see Section 4 for Monte Carlo results), we follow Müller and Wang (2017) to consider the fixed-k asymptotics. In particular, we use only the largest k order statistics as our effective sample, denoted as

$$\mathbf{Y} = (Y_{(1)}, Y_{(2)}, \cdots, Y_{(k)})$$

where $Y_{(1)} \geq Y_{(2)} \geq \cdots \geq Y_{(k)}$ denote the order statistics.

Our approach relies on the sole assumption that the underlying distribution F is regular varying (see, for example, de Haan and Ferreira (2007)). In particular, a CDF F is called regular varying at infinity if

$$\lim_{t \to \infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\xi} \qquad \text{for } x > 0 \tag{1}$$

where ξ is the tail index, measuring the decay rate of the tail. The regular variation assumption (1) is satisfied for a large range of commonly used distributions, including, normal, Pareto, Student-t, F, and Burr distributions. Given the regular variation, the extreme value theorem (the Fisher–Tippett–Gnedenko theorem) suggests that that there exist sequences a_n and b_n such that

$$\frac{Y_{(1)} - b_n}{a_n} \Rightarrow X_1 \tag{2}$$

where X_1 has the following generalized extreme value distribution

$$G_\xi(x) = \begin{cases} \exp[-(1 + \xi x)^{-1/\xi}], \ 1 + \xi x \geq 0, \text{ for } \xi \neq 0 \\ \exp[-e^{-x}], \ x \in \mathbb{R}, \ \xi = 0. \end{cases} \tag{3}$$

The cases with $\xi < 0$, $\xi = 0$ and $\xi > 0$ correspond to Weibull, Gumbel and Fréchet type, respectively.

In additional to the sample maximum, the extreme value theorem also extends to the first k order statistics such that if (2) holds, then for any fixed k,

$$\left(\frac{Y_{(1)} - b_n}{a_n}, ..., \frac{Y_{(k)} - b_n}{a_n} \right) \Rightarrow \mathbf{X} = (X_1, ..., X_k) \tag{4}$$

where the joint probability density function (PDF) of $\mathbf{X}$ is given by $f_{\mathbf{X}|\xi}(x_1, ..., x_k) = G_\xi(x_k) \prod_{i=1}^k g_\xi(x_i)/G_\xi(x_i)$ on $x_k \leq x_{k-1} \leq \ldots \leq x_1$, where $g_\xi(x) = dG_\xi(x)/dx$.

Suppose the constants a_n and b_n were known, the limiting problem then only involves a k-dimensional draw $\mathbf{X}$ whose distribution is fully characterized by the scalar parameter ξ, and then we can seek to construct an estimator that satisfies the asymptotic quantile unbiasedness whenever (4) holds. But unfortunately a_n and b_n are unknown and even difficult to estimate since they depend on the underlying distribution F, especially on ξ whose knowledge requires a large number of tail observations. To see this, in the case of the standard Pareto distribution, the constant a_n is n^{ξ} and b_n is 0. The estimator of a_n is naturally constructed as $n^{\hat{\xi}}$ for some tail index estimator $\hat{\xi}$, and thus a small estimation error in ξ leads to a substantive error in estimating a_n.

To avoid estimating a_n and b_n (essentially ξ), we impose location and scale equivariance on the estimator $\hat{Q}$, such that for any constants $a \neq 0$ and b,

$$\hat{Q}(a\mathbf{Y} + b) = a\hat{Q}(\mathbf{Y}) + b. \tag{5}$$

Such equivariance can be implemented by constructing the estimator as a function of a self-normalized statistic in the following fashion

$$\hat{Q}\left(\mathbf{Y}\right) = \left(Y_{(1)} - Y_{(k)}\right)\hat{Q}\left(\mathbf{Y}^s\right) + Y_{(k)}$$

where

$$\mathbf{Y}^s = \left(\frac{Y_{(1)} - Y_{(k)}}{Y_{(1)} - Y_{(k)}}, \frac{Y_{(2)} - Y_{(k)}}{Y_{(1)} - Y_{(k)}}, ..., \frac{Y_{(k)} - Y_{(k)}}{Y_{(1)} - Y_{(k)}}\right)$$

is a maximal invariant statistic to the linear transformations (see, for example, Lehmann and Romano (2005)). The continuous mapping theorem and (4) imply

$$\mathbf{Y}^s \Rightarrow \mathbf{X}^s \equiv \left(\frac{X_1 - X_k}{X_1 - X_k}, \frac{X_2 - X_k}{X_1 - X_k}, ..., \frac{X_k - X_k}{X_1 - X_k}\right)$$

whose distribution then depends on ξ only.

Also notice that $P(Y_{(1)} > Q(F, 1-h/n)) = (1-h/n)^n \to e^{-h}$ and $(Q(F, 1-h/n) - b_n)/a_n$ converges to the e^{-h} quantile of X_1, denoted as $q(\xi, h)$ in the following. Some calculation shows $q(\xi, h) = (h^{-\xi} - 1)/\xi$ for $\xi \neq 0$ and $q(0, h) = -\log(h)$. Since $Q(F, 1-h/n)$ and $\mathbf{Y}$ share the same normalizing constants a_n and b_n, the original problem asymptotically amounts to determining $\hat{Q}$ as a function of $\mathbf{X}^s$ on the $k-2$ dimensional subset of $\mathbb{R}^k$ where $\mathbf{X}^s$ has first element equal to 1 and last element equal to zero. More specifically, we can derive the quantile unbiasedness, that is, $P\left(Y_i > \hat{Q}\left(\mathbf{Y}\right)\right) = h/n$, as follows

$$\begin{aligned} h &= nP\left(Y_i > \hat{Q}\left(\mathbf{Y}\right)\right) \\ &= n\left(1 - F\left(\left(Y_{(1)} - Y_{(k)}\right)\hat{Q}\left(\mathbf{Y}^s\right) + Y_{(k)}\right)\right) \end{aligned}$$

$$
\begin{aligned}
&= n\left(1-F\left(a_n\left(\frac{Y_{(1)}-Y_{(k)}}{a_n}\hat{Q}\left(\mathbf{Y}^s\right)+\frac{Y_{(k)}-b_n}{a_n}\right)+b_n\right)\right) \\
&\to E_\xi\left[\left(1+\xi\left(\left(X_1-X_k\right)\hat{Q}\left(\mathbf{X}^s\right)+X_k\right)\right)^{-1/\xi}\right]
\end{aligned}
$$

where the expectation is taken w.r.t. the vector $(X_1 - X_k, X_k, \mathbf{X}^s)$ whose distribution can be derived from the PDF of $\mathbf{X}$ via change of variables, and the convergence follows from Theorem 1.1.6 of de Haan and Ferreira (2007) given F is regularly varying at infinity.

Recall that ξ cannot be consistently estimated as we only have a fixed k number of observations. Alternatively, we impose the asymptotically quantile unbiasedness for all the values of ξ in an empirically relevant set $\Xi \subset \mathbb{R}$. The asymptotic problem is then to construct $\hat{Q}$ that satisfies

$$
E_\xi\left[\left(1+\xi\left(\left(X_1-X_k\right)\hat{Q}\left(\mathbf{X}^s\right)+X_k\right)\right)^{-1/\xi}\right] = h \text{ for all } \xi \in \Xi. \tag{6}
$$

To construct the optimal $\hat{Q}$ among those satisfying (6), we focus on the one that minimizes the weighted average mean absolute deviation (MAD) criterion

$$
\int E_\xi[\left|\left(X_1-X_k\right)\hat{Q}\left(\mathbf{X}^s\right)+X_k-q\left(\xi,h\right)\right|]dW(\xi) \tag{7}
$$

where W is a positive measure with support on Ξ.[1] Thus combining the asymptotic versions of the constraint (6) and the objective (7), the limiting problem can be formulated as

$$
\begin{aligned}
&\min\nolimits_{\hat{Q}(\cdot)} \textstyle\int E_\xi[\left|\left(X_1-X_k\right)\hat{Q}\left(\mathbf{X}^s\right)+X_k-q\left(\xi,h\right)\right|]dW(\xi) \\
&E_\xi\left[\left(1+\xi\left(\left(X_1-X_k\right)\hat{Q}\left(\mathbf{X}^s\right)+X_k\right)\right)^{-1/\xi}\right] = h \text{ for all } \xi \in \Xi.
\end{aligned} \tag{8}
$$

By writing the expectations in (8) in terms of the PDF of $\mathbf{X}^s$, denoted by $f_{\mathbf{X}^s|\xi}$, the above problem can be written in a Lagrangian form

$$
\begin{aligned}
&\min_{\hat{Q}(\cdot)} \int_\Xi E_\xi[\left|\left(X_1-X_k\right)\hat{Q}\left(\mathbf{X}^s\right)+X_k-q\left(\xi,h\right)\right| |\mathbf{X}^s] f_{\mathbf{X}^s|\xi}\left(\mathbf{X}^s\right)dW(\xi) \\
&+\int_\Xi \lambda\left(\xi\right)E_\xi\left[\left(1+\xi\left(\left(X_1-X_k\right)\hat{Q}\left(\mathbf{X}^s\right)+X_k\right)\right)^{-1/\xi}|\mathbf{X}^s\right] f_{\mathbf{X}^s|\xi}\left(\mathbf{X}^s\right)d\xi
\end{aligned} \tag{9}
$$

[1] The mean squared error criterion might not be well-defined for $\xi \geq 1/2$.

where the function $\lambda(\xi)$ denotes the Lagrangian multipler, and the expectations in the above expression can be numerically computed by Gaussian quadrature. Therefore, the limiting problem can be treated as estimating $q(\xi, h)$, a function of the scalar parameter $\xi \in \Xi$, based on a single observation $\mathbf{X}^s$ from a parametric distribution indexed only by ξ. The only remaining challenge is thus to identify suitable Lagrangian multipliers. To this end, we can discretize Ξ into a fine enough grid and resort to the numerical algorithm developed in Müller and Wang (2015). This algorithm delivers the Lagrangian multipliers that lead to uniform unbiasedness up to numerical accuracy. Further details are provided in Appendix A.1.

By construction, the estimator constructed in this way is nearly optimal in the sense of minimizing the weighted risk (7) among all estimators that are location and scale invariant and satisfy the uniform asymptotic unbiasedness (6). Following Arnold, Balakrishnan, and Nagaraja (1992), it is easy to show that the simple empirical quantile is asymptotically quantile unbiased when h takes a positive integer larger than 1. This leads to the existence of the estimator (9), which is in general difficult to establish theoretically. If such uniformly unbiased estimator does not exist, we will expect an arbitrarily large Lagrangian multipler $\lambda(\xi)$ for some values of ξ as the sample size grows, reflecting the possibly infinite cost of imposing the restriction. This is not found in our Monte Carlo simulations.

2.2 Censored Data

The previous analysis can be easily generalized to cover censoring data. Consider an i.i.d. sample of n observations with the largest m being censored with a known m. We focus on the case in which the censoring value (which is recorded for all observations larger than it) is unavailable. The case with a known censoring value follows from similar derivation, and is postponed to Appendix A.2.

Note that for fixed m and k, the extreme value theorem (2) applies to the largest $m+k$ order statistics. Then, to implement the previously introduced approach, we modify the definition of $\mathbf{X}^s$ as

$$\mathbf{X}_m^s = \left(\frac{X_{m+1} - X_{m+k}}{X_{m+1} - X_{m+k}}, \frac{X_{m+2} - X_{m+k}}{X_{m+1} - X_{m+k}}, ..., \frac{X_{m+k} - X_{m+k}}{X_{m+1} - X_{m+k}} \right)$$

which is again invariant to location and scale transformation, and construct the estimator of $q(\xi, h)$ as $(X_{m+1} - X_{m+k})\,\hat{Q}(\mathbf{X}_m^s) + X_{m+k}$. The density of $\mathbf{X}_m^s$ as well as the asymptotic quantile bias (6) and risk (7) can be adjusted accordingly. Therefore, the asymptotic

Lagrangian problem can be rewritten as

$$\min_{\hat{Q}(\cdot)} \int_{\Xi} E_{\xi}[\left|\left(X_{m+1}-X_{m+k}\right)\hat{Q}\left(\mathbf{X}_m^s\right)+X_{m+k}-q\left(\xi,h\right)\right| |\mathbf{X}_m^s] f_{\mathbf{X}_m^s|\xi}\left(\mathbf{X}_m^s\right) dW(\xi) \tag{10}$$

$$+\int_{\Xi} \lambda_m\left(\xi\right) E_{\xi}\left[\left(1+\xi\left(\left(X_{m+1}-X_{m+k}\right)\hat{Q}\left(\mathbf{X}_m^s\right)+X_{m+k}\right)\right)^{-1/\xi} |\mathbf{X}_m^s\right] f_{\mathbf{X}_m^s|\xi}\left(\mathbf{X}_m^s\right) d\xi$$

where the Lagrangian multipler $\lambda_m\left(\cdot\right)$ depends on m. Similarly as before, the finite sample version can be implemented as $\left(Y_{(m+1)}-Y_{(m+k)}\right)\hat{Q}\left(\mathbf{Y}_m^s\right)+Y_{(m+k)}$ where $\mathbf{Y}_m^s$ is defined similarly as $\mathbf{Y}^s$ with $Y_{(i)}$ replaced by $Y_{(m+i)}$ for $i=1,...,k$.

2.3 Truncated Data

Truncated data exist when observations outside a certain range are automatically eliminated. Without loss of generality, we assume the truncation value is $Q\left(F,1-\tilde{h}/n\right)$ for some unknown $\tilde{h}$.[2]This means approximately the largest $\tilde{h}$ order statistics are unavailable, which is coherent with our small sample setup in the sense that both the unknown quantity and the truncation value are of the same order of magnitude as the sample maximum. We assume the truncation value $Q\left(F,1-\tilde{h}/n\right)$ is unobserved (cf. Aban, Meerschaert, and Panorska (2006) and Beirlant, Alves, and Gomes (2016)) and postpone the case in which it is observed ($\tilde{h}$ still unobserved) to Appendix A.2.

We still consider the largest k observations, whose limiting distribution is stated in the following lemma.

Lemma 1 *Suppose data are i.i.d. and generated from a CDF F truncated from above at $Q\left(F,1-\tilde{h}/n\right)$ with some unknown fixed value $\tilde{h}\geq 0$ and F is regularly varying at infinity. Then*

$$\left(\frac{Y_{(1)}-b_n}{a_n},...,\frac{Y_{(k)}-b_n}{a_n}\right)\Rightarrow \widetilde{\mathbf{X}}=\left(\tilde{X}_1,...,\tilde{X}_k\right)$$

where the PDF of $\widetilde{\mathbf{X}}$ is $G_{\xi,\tilde{h}}\left(\tilde{x}_k\right)\prod_{i=1}^k g_{\xi,\tilde{h}}\left(\tilde{x}_i\right)/G_{\xi,\tilde{h}}\left(\tilde{x}_i\right)$ with $G_{\xi,\tilde{h}}\left(x\right)=\exp\left(\tilde{h}\right)G_{\xi}\left(x\right)$ and $g_{\xi,\tilde{h}}\left(x\right)=dG_{\xi,\tilde{h}}\left(x\right)/dx$.

[2] We cannot learn anything if the tail probability under truncation is of a larger order than $1/n$, while the truncation is asymptotically negligible if the truncated mass is of a smaller order.

The proof is in Appendix A.3. The density of the maximal invariant statistic

$$\widetilde{\mathbf{X}}^s = \left(\frac{\tilde{X}_1 - \tilde{X}_k}{\tilde{X}_1 - \tilde{X}_k}, ..., \frac{\tilde{X}_k - \tilde{X}_k}{\tilde{X}_1 - \tilde{X}_k} \right)$$

can be derived similarly as $\mathbf{X}^s$ with more algebra, as well as the asymptotic bias (6) and risk (7). The exact expressions are described in Appendix A.1. Then, the estimator of $q(\xi, h)$ can be constructed as $\left(\tilde{X}_1 - \tilde{X}_k\right) \hat{Q}\left(\widetilde{\mathbf{X}}^s\right) + \tilde{X}_k$, and $\tilde{h}$ shows up in the limiting problem as an additional nuisance parameter. Given a two-dimensional weight $W\left(\xi, \tilde{h}\right)$ and the set of Lagrangian multipliers $\lambda\left(\xi, \tilde{h}\right)$ defined on $\Xi \times H$ for $H = \left[0, \bar{h}\right]$ with some pre-specified $\bar{h}$, the Lagrangian problem can be rewritten as

$$\begin{aligned} &\min_{\hat{Q}(\cdot)} \int_{\Xi \times H} E_{\xi,\tilde{h}}[\left| \left(\tilde{X}_1 - \tilde{X}_k\right) \hat{Q}\left(\widetilde{\mathbf{X}}^s\right) + \tilde{X}_k - q(\xi, h) \right| |\widetilde{\mathbf{X}}^s] f_{\widetilde{\mathbf{X}}^s|\xi,\tilde{h}}\left(\widetilde{\mathbf{X}}^s\right) dW(\xi, \tilde{h}) \qquad (11) \\ &+ \int_{\Xi \times H} \lambda\left(\xi, \tilde{h}\right) E_{\xi,\tilde{h}}\left[\left(1 + \xi\left(\left(\tilde{X}_1 - \tilde{X}_k\right) \hat{Q}\left(\widetilde{\mathbf{X}}^s\right) + \tilde{X}_k\right)\right)^{-1/\xi} |\widetilde{\mathbf{X}}^s \right] f_{\widetilde{\mathbf{X}}^s|\xi,\tilde{h}}\left(\widetilde{\mathbf{X}}^s\right) d\xi d\tilde{h}. \end{aligned}$$

Note that our approach can be easily adapted to estimate the $1 - h/n$ quantile of the truncated distribution, which equals the $1 - \left(h + \tilde{h}\right)/n + h\tilde{h}/n^2$ quantile of the original distribution. Therefore, the same asymptotic problem can be set up with h replaced by $h + \tilde{h}$ since the term $h\tilde{h}/n^2$ is asymptotically negligible. For implementation, the estimator is constructed in the same way as in the complete data case.

2.4 Extension to Time Series Data

In financial applications, data may exhibit time series correlation and heteroskedasticity. To overcome this difficulty, we can resort to the stochastic volatility models with i.i.d. driving innovations. In particular, we assume data are completely observed and generated from a correctly specified AR($\tilde{p}$)-GARCH(p,q) model (cf. McNeil and Frey (2000)). The conditional quantile of a one-step ahead forecast then simply becomes the product of the square root of the conditional heteroskedasticity function and the estimated quantile of the driving innovations. We show that estimation error of the AR and GARCH parameters is negligible for our asymptotic theory, so that we can apply the fixed-k approach to the estimated innovations.

More specifically, let Z_t denote the real data which are assumed to be the following stationary time series

$$\begin{aligned}
Z_t &= \mu_t + \sigma_t Y_t \\
\sigma_t^2 &= \alpha_0 + \alpha_1 Y_{t-1}^2 + ...\alpha_q Y_{t-q}^2 + \beta_1 \sigma_{t-1}^2 + ... + \beta_p \sigma_{t-p}^2 \\
\mu_t &= \bar{\mu} + \phi_1 Z_{t-1} + ...\phi_{\tilde{p}} Z_{t-\tilde{p}},
\end{aligned}$$

where the innovation Y_t is i.i.d. with CDF F. As standard in the literature, we assume that μ_t and σ_t are measurable with respect to $\mathcal{F}_{t-1}$, the information available up to time $t-1$.

To estimate the unknown coefficients, we can apply the pseudo maximum likelihood (PML) estimator, which maximizes the likelihood under the assumption of standard Gaussian innovations. Given the PML estimator, we can back out the estimated conditional mean and standard deviation series, denoted as $\{\hat{\mu}_t\}$ and $\{\hat{\sigma}_t\}$, respectively. Then, the residual is calculated as

$$\hat{Y}_t = \frac{Z_t - \hat{\mu}_t}{\hat{\sigma}_t}$$

which can be used as i.i.d. data for estimating high quantile and TCE. The following theorem shows that the error in fitting the AR-GARCH type models is asymptotically negligible if the estimator of the coefficients is consistent.

Theorem 1 *Suppose there exists a consistent estimator of the AR($\tilde{p}$)-GARCH(p,q) coefficients for some known positive integers ($\tilde{p}$,p,q), then the estimated innovations $\{\hat{Y}_t\}$ satisfy the extreme value theorem, i.e., the largest k innovations with a fixed and given k, $\left(\hat{Y}_{(1)}, ..., \hat{Y}_{(k)}\right)$, satisfy (4).*

The proof is in Appendix A.3. This theorem validates the weak convergence (4) for the ordered estimated innovations, and hence the previously suggested approach is applicable again. As a summary, our estimator can be implemented by the following steps:

Step 1 For time $t > n$, fit the data $\{Z_t, Z_{t-1}, ..., Z_{t-n+1}\}$ with an AR-GARCH type model and obtain the standardized innovations.

Step 2 Compute the fixed-k estimators $\hat{Q}$ by using the largest k standardized innovations, denoted by $\hat{\mathbf{Y}}_t$, and solving the Lagrangian problems (9).

Step 3 Plug in the conditional mean and standard deviation at time t to construct the one-step prediction of the high quantile, that is, $\hat{Q}\left(\hat{\mathbf{Y}}_t\right)\hat{\sigma}_t + \hat{\mu}_t$.

When the data generating process is unknown, it is in general difficult to establish (4) (and even (2)) for strictly stationary and weakly dependent time series data. Leadbetter (1974) (see also Leadbetter (1983) and O'Brien (1987)) establishes that if the sample maximum (normalized by a_n and b_n) has the same extreme value distribution G_ξ if data satisfy a global mixing condition (referred as the D assumption which is implied by strong mixing) and another condition restricting that large observations cannot cluster (referred as the D' condition). Furthermore, Chernick, Hsing, and McCormick (1991) and Hsing (1993) show that if the data do not satisfy the D' condition yet have a limit distribution of the sample maximum, this limit is of the form $G_\xi^\theta(\cdot)$ where $\theta \in (0,1]$ is called the extremal index. This scalar parameter θ captures the tail dependence structure. In particular, the i.i.d. case implies $\theta = 1$ while the identical observation case is reflected by $\theta = 0$.

The consistent estimation of θ typically requires more conditions and an even larger sample size (see Ancona-Navarrete and Tawn (2000) and Süveges (2007) and references therein). So in small samples, we can treat it as an additional nuisance parameter and in principle construct a uniformly unbiased estimation of the high quantile for all values of $\theta \in [\underline{\theta}, 1]$ for some $\underline{\theta} \in (0,1)$. However, the condition for the convergence to the limit distribution $G_\xi^\theta(\cdot)$ and even for determining the tail index can only be verified in some specific case (see, for instance, O'Brien (1987), Leadbetter (1983), Davis (1985), Kearns and Pagan (1997), and Mikosch and Starica (2000)). So it is unclear in how to determine an empirically relevant range of dependence (the value of $\underline{\theta}$) over which we would like to impose the unbiasedness.

3 Tail Conditional Expectation and Tail Index

3.1 Tail Conditional Expectation

The fixed-k estimators (9), (10), and (11) can be modified for the TCE: $T(F, 1-h/n) = E[Y_i | Y_i \geq Q(F, 1-h/n)]$, for given h. Assume F is in regularly varying with tail index $\xi < 1$ (otherwise, the tail conditional expectation does not exist). The limiting problem still amounts to choose the estimator $\hat{T}$ as a function of the limiting observation $\mathbf{X}^s$ to minimize a weighted average risk subject to some unbiasedness restriction. In particular, given $(T(F, 1-h/n) - b_n)/a_n \to \tau(\xi, h) = q(\xi, h)/(1-\xi) - 1/\xi$ for $\xi \neq 0$ and $1 - \log(h)$ for $\xi = 0$ (cf. Müller and Wang (2017), pp. 1336), the limit version of the weighted average risk given a weight W on Ξ thus has the form

$$\int_\Xi E_\xi[\left|(X_1 - X_k)\hat{T}(\mathbf{X}^s) + X_k - \tau(\xi, h)\right| dW(\xi).$$

Regarding the constraint, we impose the mean unbiasedness restriction on the estimator of TCE, that is,

$$E\left[\hat{T}\left(\mathbf{Y}\right)\right]-T\left(F,1-h/n\right)=0, \tag{12}$$

which is motivated in finance and risk management. Recall that for a positive random variable Z with CDF F_Z, $E[Z]=\int(1-F_Z(z))dz$. Denote $\hat{F}^{TCE}\left(\cdot\right)$ as the CDF leading to the TCE that equals $\hat{T}\left(\mathbf{Y}\right)$, then the constraint (12) is equivalent to

$$\int_{Q(F,1-h/n)}^{\infty}\left(\hat{F}^{TCE}\left(z\right)-F\left(z\right)\right)dz=0$$

which can be interpreted as an average quantile unbiasedness above the true quantile. In finance, the high quantile and TCE can be interpreted as the VaR and the ES, respectively. The above expression then suggests that the expected shortfall captures the unbiasedness at all the confidence levels above a certain VaR, which, however, measures the risk at only one particular level. As pointed out by Basel Committee on Banking Supervision (2013): "*A number of weaknesses have been identified with using VaR for determining regulatory capital requirements, including its inability to capture "tail risk". For this reason, the Committee proposed in May 2012 to replace VaR with ES. ES measures the riskiness of a position by considering both the size and the likelihood of losses above a certain confidence level.*"

To impose (12), which is asymptotically equivalent to $E_\xi\left[\left(X_1-X_k\right)\hat{T}\left(\mathbf{X}^s\right)+X_k-\tau\left(\xi,h\right)\right]=0$,[3] we can construct the following Lagrangian problem

$$\begin{aligned}&\min_{\hat{Q}(\cdot)}\int_{\Xi}E_\xi[\left|\left(X_1-X_k\right)\hat{T}\left(\mathbf{X}^s\right)+X_k-\tau\left(\xi,h\right)\right||\mathbf{X}^s]f_{\mathbf{X}^s|\xi}\left(\mathbf{X}^s\right)dW(\xi) \\ &+\int_{\Xi}\tilde{\lambda}\left(\xi\right)E_\xi\left[\left(X_1-X_k\right)\hat{T}\left(\mathbf{X}^s\right)+X_k-\tau\left(\xi,h\right)|\mathbf{X}^s\right]f_{\mathbf{X}^s|\xi}\left(\mathbf{X}^s\right)d\xi\end{aligned} \tag{13}$$

where $\tilde{\lambda}$ is another set of Lagrangian multipliers that are numerically determined. Once $\hat{T}$ is obtained, the estimator of TCE can be implemented as $\left(Y_{(1)}-Y_{(k)}\right)\hat{T}\left(\mathbf{Y}^s\right)+Y_{(k)}$.

When data exhibit censoring or truncation, the same procedure applies by adjusting the density, risk, and bias terms accordingly. We postpone the exact expressions to the Appendix.

3.2 Tail Index

Given the limiting observation $\mathbf{X}^s$, it is even more straightforward to construct an estimator $\hat{\xi}$ which minimizes a weighted average risk criteria and satisfies a certain unbiased restriction. In particular,

[3] A formal derivation is obtained by applying Theorem 5.3.1 of de Haan and Ferreira (2007) and assuming that $\hat{T}\left(\cdot\right)$ is uniformly bounded, which is without loss of generality given $T\left(F,1-h/n\right)$ is of the same order as $Y_{(1)}-Y_{(k)}$.

we choose the median unbiased restriction, that is, $P\left(\hat{\xi}\left(\mathbf{Y}^{s}\right)>\xi\right)=1/2$ at least asymptotically. This restriction has been proposed by Andrews (1993) and Stock and Watson (1998) to learn the parameters that cannot be consistently estimated. Assume that $\xi\in\Xi$ a compact subset of $\mathbb{R}$, the dominated convergence theorem and (4) yield that

$$\begin{aligned} E\left[\left(\hat{\xi}\left(\mathbf{Y}^{s}\right)-\xi\right)^{2}\right]d\xi &\rightarrow E\left[\left(\hat{\xi}\left(\mathbf{X}^{s}\right)-\xi\right)^{2}\right] \\ &= \int\left(\hat{\xi}\left(\mathbf{x}^{s}\right)-\xi\right)^{2}f_{\mathbf{X}^{s};\xi}\left(\mathbf{x}^{s}\right)d\mathbf{x}^{s} \end{aligned}$$

and for any $\xi\in\Xi$

$$\begin{aligned} P\left(\hat{\xi}\left(\mathbf{Y}^{s}\right)>\xi\right) &\rightarrow P\left(\hat{\xi}\left(\mathbf{X}^{s}\right)>\xi\right) \\ &= \int\mathbf{1}\left[\hat{\xi}\left(\mathbf{x}^{s}\right)>\xi\right]f_{\mathbf{X}^{s};\xi}\left(\mathbf{x}^{s}\right)d\mathbf{x}^{s}. \end{aligned}$$

We present the result with the mean standard error since it is well defined given our bounded space Ξ and widely used in comparing different approaches (see, for example, de Haan and Peng (1998)). It can be replaced by MAD without any difficulty. Combine the risk and the bias terms, our problem is asymptotically equivalent to

$$\min_{\hat{\xi}\in\Xi}\int_{\Xi}\left(\hat{\xi}\left(\mathbf{X}^{s}\right)-\xi\right)^{2}f_{\mathbf{X}^{s}|\xi}\left(\mathbf{X}^{s}\right)dW\left(\xi\right)+\int_{\Xi}\lambda\left(\xi\right)\left(\hat{\xi}\left(\mathbf{X}^{s}\right)-\xi\right)^{2}f_{\mathbf{X}^{s}|\xi}\left(\mathbf{X}^{s}\right)d\xi \tag{14}$$

where $\lambda\left(\xi\right)$ is the Lagrangian multiplier, and the density $f_{\mathbf{X}^{s}|\xi}\left(\mathbf{X}^{s}\right)$ can be numerically computed by Gaussian quadrature. Again, the estimator constructed in (9) is nearly the best in the sense of nearly minimizing the weighted risk (7) among all estimators that are invariant to location and scale and satisfy the uniform asymptotic unbiasedness (6). It turns out that being invariant to location or not makes a substantive effect in small sample performance as shown in Monte Carlos.

It is worth mentioning that the fixed-k asymptotics can be easily modified to construct confidence intervals for ξ. Now consider the hypothesis testing problem

$$H_{0}:\xi=\xi_{0}\text{ against }H_{1}:\xi\neq\xi_{0},\ \xi\in\Xi.$$

Given the effective limiting observation $\mathbf{X}^{s}$ whose density $f_{\mathbf{X}^{s};\xi}$ is fully characterized by ξ, we can simply construct the likelihood ratio test if the alternative is simple. To transform the composite alternative into a simple one, we consider a weighted average power critera (see, among others, Andrews and Ploberger (1994), Elliott, Müller, and Watson (2015), and Lehmann and Romano

(2005)). With a positive measure $W(\cdot)$ defined on Ξ, our test is constructed as

$$\varphi(\mathbf{x}^s) = \mathbf{1}\left[\frac{\int_\Xi f_{\mathbf{X}^s|\xi}(\mathbf{x}^s)\, dW(\xi)}{f_{\mathbf{X}^s|\xi_0}(\mathbf{x}^s)} > \kappa(\alpha; \xi_0, k)\right] \tag{15}$$

where $\kappa(\alpha; \xi_0, k)$ is the critival value for significance level α, null value ξ_0 and the length of the effect sample. This is a simplier problem than the case in Müller and Wang (2017) since there is no nuisance parameter under the null hypothesis. We calculate the critical values by simulation and provide the results on the author's website. In finite samples, this test can be implemented by replacing $\mathbf{X}^s$ by $\mathbf{Y}^s$. The continuous mapping theorem and the extreme value theorem (4) yield that $E[\varphi(\mathbf{Y}^s)] \to E[\varphi(\mathbf{X}^s)] = \alpha$ under the null and the confidence interval is obtained by inverting the test (15). Note that by construction, this likelihood ratio test maximizes the weighte average power among all invariant tests that have a converging power function under (4) and that control size under the null.

As a final remark in this section, both the estimator (14) and the test (15) can be easily modified to data with censoring and truncation, and thus omitted for notational ease.

4 Monte Carlo Simulations

4.1 Extreme Quantile and TCE

This section reports some small sample results of estimating extreme quantiles and the corresponding TCEs. In particular, we consider $h = 0.5$ and 5 and $n = 250$, corresponding to the confidence levels at 99.8% and 98%. For expositional ease, we only report the results for $k = 20$. We consider six data generating processes: A Pareto law with tail index equal to $\xi = 0.25$, a standard normal distribution, a standard lognormal distribution, a Student-t distribution with 3 degrees of freedom, and the empirical distributions of the AR(1)-GARCH(1,1) residuals of S&P500 and Nasdaq daily returns from 02/08/1990 to 04/17/2017 (see Section 2.4 for a theoretical justification for using such residuals). For the censored data model, the largest 5 observations are censored. For the truncated data model, we generate the data from those distributions truncated at $Q\left(F, 1 - \tilde{h}/n\right)$ with $\tilde{h} = \{0, 1, 2\}$, and impose the unbiasedness for the data truncated at up to $Q(F, 1 - 2/n)$, i.e., $\bar{h} = 2$.

Tables 1 and 2 present the (quantile) bias and the mean absolute deviation of the fixed-k approach and three other popular estimators: (i) the estimator based on the Smith (1987) estimator (Smith) of ξ: $\hat{Q}_{Smith} = Y_{(k)} + \hat{\sigma}^{Smith}((h/(k-1))^{-\hat{\xi}^{Smith}} - 1)/\hat{\xi}^{Smith}$ and $\hat{T}_{Smith} = (\hat{Q}_{Smith} + \hat{\sigma}^{Smith} - \hat{\xi}^{Smith} Y_{(k)})/\left(1 - \hat{\xi}^{Smith}\right)$ where $\hat{\xi}^{Smith}$ and $\hat{\sigma}^{Smith}$ denote the maximum likelihood estimators of the tail index and the scale, correspondingly, by fitting the exceedances,

$Y_{(1)} - Y_{(k)}, ..., Y_{(k-1)} - Y_{(k)}$ with a generalized Pareto distribution (see also McNeil and Frey (2000)) (ii) the estimators described in Chapter 4 of de Haan and Ferreira (2007) textbook (dH-F): $\hat{Q}_{dHF} = Y_{(k)} + \hat{a}\left(n/(k-1)\right)\left((h/(k-1))^{-\hat{\xi}^{M}} - 1\right)/\hat{\xi}^{M}$ and $\hat{T}_{dHF} = Y_{(k)} + \hat{a}\left(n/(k-1)\right)\left((h/(k-1))^{-\hat{\xi}^{M}} - 1 + \hat{\xi}^{M}\right)/(\hat{\xi}^{M}(1-\hat{\xi}^{M}))$ where $\hat{a}\left(n/(k-1)\right)$ and $\hat{\xi}^{M}$ are moment estimators of the scale and the tail index, respectively, (see also Dekkers and de Haan (1989) and de Haan and Rootzén (1993)); and (iii) the classic Weissman (1978) estimator (W-H): $\hat{Q}_{WH} = Y_{(k)}\left(h/(k-1)\right)^{-\hat{\xi}^{H}}$ and $\hat{T}_{WH} = (Y_{(k)}\left(h/(k-1)\right)^{-\hat{\xi}^{H}})/(1-\hat{\xi}^{H})$ where $\hat{\xi}^{H}$ denotes the classic Hill (1975) estimator. For all four methods we use the same parameter space $\Xi = [-1/2, 1/2]$ for imposing unbiasedness or estimating ξ.

The bias for quantile is reported as $100\left(E\left[1 - F\left(\hat{Q}\left(\mathbf{Y}\right)\right)\right] - h/n\right)$, that is, the probability measured in percentage that an additional random draw from F is larger than the quantile estimator minus the target tail probability. For TCE, we report the bias $E\left[\hat{T}\left(\mathbf{Y}\right) - E\left[Y_i | Y_i > \hat{Q}\left(\mathbf{Y}\right)\right]\right]$ to reflect the fact that the error in estimating high quantiles should be included in estimating the corresponding TCE. The mean bias relative to the true TCE exhibits a similar pattern, and hence is not reported for ease of presentation. In addition, we also report the bias measured in h. More precisely, the h bias is defined as $\hat{h} - h$ where $\hat{h}$ is value of h that satisfies $E\left[\hat{T}\left(\mathbf{Y}\right)\right] = T\left(F, 1-h/n\right)$. Linear interpolation is implemented for the AR(1)-GARCH(1,1) residuals.

We find that the new method has much smaller bias across all h in contrast to the other three methods. In particular, the quantile bias of the Smith estimator is approximately 0.2% at $h = 0.5$, which means the Smith method approximately delivers the 99.6% quantile while the true target is 99.8%. The other two estimators exhibit small sample biases that differ a lot across distributions, indicating that $k = 20$ is still too small for their increasing-k asymptotics to perform satisfactorily. For relatively large h such as 5, the unbiasedness restriction is much easier to impose as reflected by substantially smaller MADs. This is because the quantity is closer to the central part of the distribution and therefore more observations can be collected from the right side of the true quantile.

Table 3 lists the small sample bias of the fixed-k and the Smith (adjusted to fit a truncated Pareto) methods for censored data (the other two estimators are not applicable). These results suggest that the Smith method always substantially underestimate the high quantile. In particular, the Smith estimator delivers approximately the 97.8% quantile while the true target is 99.8%.

Table 4 depicts the performance of the fixed-k method with data truncation. We compare a recently developed method (TP-WH) by Beirlant, Alves, and Gomes (2016), who suggest fitting the largest k observations with a truncated Pareto distribution to estimate ξ and the truncation value and applying the Weissman-Hill type estimator. These numbers suggest that the new approach has an excellent small sample performance in terms of the quantile unbiasedness while TP-WH method tends to underestimate and hence has a smaller MAD. Note that although the bias is not large from the perspective of violation probability, the actural level of bias can be substantive given the

Table 1: Small Sample Properties for the $1 - h/n$ Quantile

Quantile	$h = 0.5$							
	fixed-k		Smith		dH-F		W-H	
	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
Pareto	0.00	2.89	0.23	1.11	0.22	0.93	0.08	0.83
Normal	0.02	0.58	0.24	0.30	0.20	0.30	-0.15	0.88
Lognormal	-0.01	13.6	0.22	5.25	0.20	4.64	-0.02	4.48
Student-t	0.01	6.29	0.23	2.51	0.22	2.23	-0.01	2.00
SP500	0.01	0.84	0.24	0.42	0.20	0.40	-0.13	0.93
Nasdaq	0.02	0.84	0.24	0.39	0.20	0.36	-0.12	0.79
	$h = 5$							
	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
Pareto	-0.01	0.22	0.02	0.21	0.18	0.20	0.20	0.20
Normal	0.02	0.13	0.17	0.13	0.27	0.13	0.39	0.13
Lognormal	-0.01	1.08	0.17	1.01	0.27	0.94	0.40	0.83
Student-t	-0.04	0.48	0.14	0.44	0.23	0.42	0.35	0.37
SP500	-0.00	0.15	0.15	0.15	0.30	0.14	0.38	0.15
Nasdaq	-0.01	0.13	0.15	0.13	0.27	0.13	0.37	0.13

Note: Entries are quantile biases and mean absolute deviations of estimators in a sample of size $n = 250$ about the $1 - h/n$ quantile of the underlying distribution F, based on the largest 20 order statistics. See the main text for a description of the four types of estimators. Based on 5,000 Monte Carlo simulations.

heavy-tail.

The substantial difference across three data types in MAD indicates that the largest observations are very informative about the right tail and thus being precisely quantile unbiased is expensive as measured by risk. Hence if they are unobserved due to either censoring or truncation, the unbiasedness has to hold at a much larger cost.

4.2 Tail Index

Since tail index is of particular empirical imporance, this section exclusively examines some small sample behavior for our estimator (14) and confidence interval by inverting (15) (fixed-k) and three other popular approaches in the literature. In particular, we implement: (i) the Smith (1987) estimator (Smith), which treats the exceedances $Y_{(i)} - Y_{(k)}$ for $i = 1, ..., k-1$ as i.id. draws from a generalized Pareto distribution and estimate ξ by maximizing the likelihood; (ii) the classic Hill (1975) estimator (Hill); and (iii) the bias-reduced log-log regression estimator proposed by Gabaix and Ibragimov (2011) (GI), who regress the log rank $\log(i - 1/2)$ on a constant and the log size, $\log\left(Y_{(i)}\right)$, for $i = 1, ..., k$. All three estimators are asymptotically normal with convergence rate $k^{-1/2}$ but different asymptotic variances, which can be easily estimated by plugging in the tail

Table 2: Small Sample Properties for TCE above the $1 - h/n$ Quantile

TCE	$h = 0.5$											
	fixed-k			Smith			dH-F			W-H		
	Bias	h Bias	MAD	Bias	h Bias	MAD	Bias	h Bias	MAD	Bias	h Bias	MAD
Pareto	-0.07	0.02	2.29	0.86	-0.20	3.39	0.10	-0.03	1.94	-0.16	0.06	1.53
Normal	-0.04	0.07	0.44	-0.15	0.33	0.48	0.02	-0.03	0.46	-1.20	14.9	2.02
Lognormal	0.40	-0.03	10.3	4.82	-0.24	15.6	-1.85	0.15	10.4	-9.45	1.90	14.5
Student-t	-0.53	0.07	5.53	1.74	-0.16	8.32	0.04	-0.00	5.23	-2.83	0.59	5.70
SP500	-0.05	0.06	0.67	-0.12	0.16	0.79	0.04	-0.05	0.67	-1.07	5.20	2.23
Nasdaq	-0.08	0.09	0.63	-0.20	0.28	0.75	0.12	-0.11	0.65	-0.72	2.25	1.74
	$h = 5$											
	Bias	h Bias	MAD	Bias	h Bias	MAD	Bias	h Bias	MAD	Bias	h Bias	MAD
Pareto	0.02	-0.11	0.52	0.06	-0.32	0.62	0.10	-0.53	0.49	-0.04	0.23	0.48
Normal	0.00	0.00	0.18	-0.05	0.72	0.18	0.03	-0.34	0.18	-0.35	7.23	0.36
Lognormal	0.08	-0.09	2.45	0.28	-0.32	2.94	0.04	-0.05	2.43	-1.82	2.87	2.42
Student-t	0.03	-0.07	1.21	0.13	-0.31	1.43	0.16	-0.39	1.13	-0.69	2.22	1.10
SP500	0.01	-0.11	0.22	-0.05	0.61	0.23	0.02	-0.28	0.22	-0.36	6.14	0.39
Nasdaq	0.01	-0.12	0.20	-0.04	0.53	0.21	0.03	-0.30	0.20	-0.30	5.39	0.33

Note: Entries are biases and mean absolute deviations of estimators in a sample of size $n = 250$ about the tail conditional expectation above the $1 - h/n$ quantile of the underlying distribution F, based on the largest 20 order statistics. See the main text for a description of the three types of estimators and the definitions of biases. Based on 5,000 Monte Carlo simulations.

Table 3: Small Sample Properties for the $1 - h/n$ Quantile with Censored Data

Quantile	$h = 0.5$				$h = 5$			
	fixed-k		Smith		fixed-k		Smith	
	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
Pareto	0.01	5.19	2.20	2.12	0.00	0.28	2.32	0.44
Normal	0.04	3.12	2.10	0.86	0.00	0.16	2.50	0.35
Lognormal	-0.02	25.6	2.18	10.2	-0.02	1.38	2.29	2.10
Student-t	0.04	10.7	2.17	4.67	-0.02	0.59	2.30	0.91
SP500	0.06	3.67	2.16	1.07	0.02	0.19	2.47	0.39
Nasdaq	0.06	3.29	2.15	0.92	0.02	0.17	2.47	0.34

Note: Entries are quantile biases and mean absolute deviations of estimators in a sample of size $n = 250$ about the $1 - h/n$ quantile of the underlying distribution F, with the largest 5 observations censored. See the main text for a description of the two types of estimators. Based on 5,000 Monte Carlo simulations.

Table 4: Small Sample Properties for the $1 - h/n$ Quantile with Truncated Data

Quantile	Fixed-k						TP-WH					
							$h = 0.5$					
Truncation	$\tilde{h} = 0$		$\tilde{h} = 1$		$\tilde{h} = 2$		$\tilde{h} = 0$		$\tilde{h} = 1$		$\tilde{h} = 2$	
	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
Pareto	0.05	9.98	-0.01	5.85	0.01	3.27	0.31	1.03	0.93	1.11	2.07	1.58
Normal	0.07	5.19	-0.01	2.64	0.03	1.70	1.36	0.44	3.07	0.49	4.62	0.73
Lognormal	0.03	48.3	-0.00	28.6	-0.00	16.1	0.62	6.51	1.91	5.36	3.47	7.70
Student-t	0.07	20.5	0.02	12.3	0.03	6.57	0.38	3.11	1.14	2.54	2.40	3.61
SP500	0.07	6.39	0.00	3.19	0.07	2.02	1.01	0.54	2.39	0.59	4.40	0.88
Nasdaq	0.10	5.72	0.03	2.85	0.04	1.89	0.84	0.48	2.70	0.50	4.21	0.73
							$h = 5$					
Pareto	0.06	1.73	-0.01	1.48	0.05	1.09	-0.12	0.28	0.60	0.19	1.77	0.20
Normal	0.06	1.02	-0.05	0.84	0.05	0.69	0.87	0.22	2.66	0.21	4.25	0.26
Lognormal	0.00	8.35	-0.02	7.24	0.03	5.28	0.19	1.44	1.56	0.94	3.15	1.01
Student-t	0.13	3.49	0.12	3.06	0.15	2.23	-0.09	0.64	0.76	0.41	2.05	0.45
SP500	0.05	1.18	-0.11	0.96	-0.01	0.78	0.54	0.22	2.03	0.19	4.06	0.27
Nasdaq	0.04	1.06	-0.01	0.87	0.06	0.69	0.38	0.19	2.33	0.18	3.86	0.23

Note: Entries are quantile biases and mean absolute deviations of the fixed-k estimator in a sample of size $n = 250$ about the $1 - h/n$ quantile of the underlying distribution F, with data generated from the truncated F at $Q(F, 1 - \tilde{h}/n)$. Based on 5,000 Monte Carlo simulations.

index estimator. The corresponding confidence intervals are readily constructed as $1.96 \pm \hat{\sigma}_{\hat{\xi}} k^{-1/2}$ for their corresponding asymptotic variance estimator $\hat{\sigma}_{\hat{\xi}}$. There are other location and scale invariant estimators proposed in the literature such as Pickands (1975) and Aban and Meerschaert (2001). These methods are stictly dominted in our small sample experiments and hence the results are not reported. For a coherent comparison, we restrict the parameter space to be $[0, 1]$ in all experiments so that any estimator or confidence interval outside this range are censored. To compare estimators, we report the median bias $P\left(\hat{\xi} > \xi\right) - 1/2$ and the root mean squared error. For confidence intervals, we report the probabilities of covering the null value and the length.

In Table 5, we investigate the effect of data shifting by generating 250 i.i.d draws from standard Pareto distribution with $\xi = 0.5$ and shifting them by d multiplies of the interquantile range of Pareto (0.5) for $d = 0, 1, 2$. We implement all four methods with different choices of k. Panels A and B depict the finite sample performance of the estimators. Clearly the shift variant estimators (Hill and GI) are nearly median unbiased under no data shifting, but exhibit severe biases if data are shifted. Such bias increases as more tail observations are taken into account. The Smith estimator which is invariant to shift also suffers substantive bias since the sample size not large enough for the asymptotics to perform well. In contrast, our fixed-k estimator is nearly median unbiased in all setups with the cost of a higher mean squared error, reflecting the fact that the location is very informative for studying the tail. These findings are coherent with Panels C and D, where results of different confidence intervals are collected. In particular, the three competing methods lead to confidence intervals with coverages substantially different from the 95% target, espectially when data are shifted, while the fixed-k approach has very precise size properties.

Table 6 examines the robustness to different parametric assumptions. We generate 250 i.i.d draws from three distributions: Student-t with degree of freedom 2 (denoted as t(2)), F(4,4), and a mixture distribution (denoted as mixP) with 20% draws from Pareto (0.5) and 80% draws from Pareto (0.1). All three distributions share the same tail index $\xi = 0.5$. The mixture distribution is essentially motivated from the double-power law proposed by Barro and Jin (2011), reflecting that only a very limited number of observations are relevant for the true tail. As seen from the columns associated with mixP, choosing a large k leads to large bias and incorrect coverage. In Panels A and B, we observe that the Hill estimator and the GI estimator are senstive to deviation from the Pareto model in terms of bias, and the Smith estimator is relatively robust and still exhibits large bias in small samples (see also Embrechts, Klupperberg, and Mikosch (1997) pp. 197 and 337 and Huisman, Koedijk, Kool, and Palm (2001)). In constrat, our fixed-k estimator performs well in controlling bias in all data generating processes. Regarding confidence intervals, Panels C and D provide similar findings: the fixed-k method dominates the Smith method in both coverage and length while the Hill and the GI approaches are subject to substantive undercoverage. It is worth mentioning that choosing $k \geq 50$ in the Student-t distribution situation amounts to use about half

Table 5: Small Sample Properties for Pareto Distribution Draws with Different Shift

Panel A: Median Bias of Estimators												
	k=10			k=30			k=50			k=70		
d	0	1	2	0	1	2	0	1	2	0	1	2
Smith	-0.19	-0.19	-0.19	-0.12	-0.12	-0.12	-0.09	-0.09	-0.09	-0.07	-0.07	-0.07
Hill	-0.05	-0.18	0.27	-0.03	-0.34	-0.45	-0.02	-0.44	-0.50	-0.01	-0.48	-0.50
GI	-0.04	-0.12	0.18	-0.00	-0.19	-0.30	0.00	-0.26	-0.37	0.01	-0.32	-0.42
fixed-k	-0.00	-0.00	-0.00	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
Panel B: RMSE of Estimators												
d	0	1	2	0	1	2	0	1	2	0	1	2
Smith	0.40	0.40	0.40	0.30	0.30	0.30	0.24	0.24	0.24	0.20	0.20	0.20
Hill	0.17	0.16	0.17	0.09	0.11	0.16	0.07	0.12	0.17	0.06	0.12	0.18
GI	0.22	0.22	0.22	0.13	0.14	0.15	0.10	0.12	0.15	0.09	0.11	0.15
fixed-k	0.40	0.40	0.40	0.28	0.28	0.28	0.22	0.22	0.22	0.19	0.19	0.19
Panel C: Coverage Probability of Confidence Intervals												
d	0	1	2	0	1	2	0	1	2	0	1	2
Smith	1.00	1.00	1.00	0.81	0.81	0.81	0.86	0.86	0.86	0.89	0.89	0.89
Hill	0.88	0.81	0.74	0.93	0.78	0.44	0.94	0.54	0.17	0.94	0.35	0.04
GI	0.93	0.88	0.83	0.95	0.85	0.72	0.95	0.79	0.57	0.96	0.72	0.43
fixed-k	0.96	0.96	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Panel D: Length of Confidence Intervals												
d	0	1	2	0	1	2	0	1	2	0	1	2
Smith	1.00	1.00	1.00	0.88	0.88	0.88	0.75	0.75	0.75	0.66	0.66	0.66
Hill	0.62	0.55	0.51	0.36	0.30	0.26	0.28	0.22	0.19	0.23	0.18	0.15
GI	0.92	0.85	0.79	0.52	0.46	0.42	0.40	0.35	0.31	0.34	0.28	0.25
fixed-k	0.73	0.73	0.73	0.74	0.74	0.74	0.69	0.69	0.69	0.64	0.64	0.64

Note: Entries are median biases and root mean squared errors of estimators and coverage and length of confidence intervals in a sample of size n=250 i.i.d. draws from Pareto (0.5) shifted by d times the interquantile range. Based on the largest k order statistics. See the main text for a description of the four types of estimators. Based on 5,000 Monte Carlo simulations.

of the (positive) sample to approximate the Pareto tail. This clearly results in poor finite sample approximation of the extreme value theorem and thus the large bias and undercoverage of the fixed-k approach.

5 Empirical Applications

5.1 Macroeconomic Disaster

This session applies the new approach to estimate the tail index of macroeconomic disasters and the corresponding coefficient of relative risk aversion. Barro and Ursua (2008) define the macroeconomic

Table 6: Small Sample Properties for iid draws from Different Distributions with No Shift

Panel A: Median Bias of Estimators												
		k=10			k=30			k=50			k=70	
Dist.	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP
Smith	-0.19	-0.19	-0.18	-0.18	-0.12	-0.11	-0.25	-0.11	-0.12	-0.34	-0.10	-0.14
Hill	0.04	0.08	-0.24	0.36	0.37	-0.45	0.50	0.48	-0.49	0.47	0.50	-0.50
GI	0.03	0.07	-0.15	0.21	0.36	-0.28	0.41	0.40	-0.36	0.33	0.47	-0.41
fixed-k	-0.01	0.00	0.01	-0.08	-0.02	-0.00	-0.19	-0.04	-0.06	-0.30	-0.05	-0.08
Panel B: RMSE of Estimators												
Dist.	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP
Smith	0.41	0.40	0.40	0.31	0.29	0.29	0.27	0.23	0.23	0.26	0.20	0.20
Hill	0.18	0.19	0.17	0.15	0.16	0.15	0.24	0.19	0.16	0.40	0.23	0.17
GI	0.23	0.23	0.22	0.15	0.16	0.15	0.16	0.16	0.14	0.21	0.17	0.14
fixed-k	0.40	0.40	0.40	0.28	0.28	0.28	0.24	0.22	0.22	0.23	0.19	0.19
Panel C: Coverage Probability of Confidence Intervals												
Dist.	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP
Smith	1.00	1.00	1.00	0.75	0.83	0.84	0.75	0.88	0.87	0.71	0.89	0.87
Hill	0.93	0.93	0.75	0.93	0.92	0.47	0.39	0.62	0.22	0.01	0.19	0.07
GI	0.95	0.95	0.84	0.99	0.99	0.73	0.96	0.97	0.61	0.75	0.88	0.48
fixed-k	0.96	0.96	0.95	0.95	0.96	0.96	0.92	0.95	0.94	0.86	0.95	0.94
Panel D: Length of Confidence Intervals												
Dist.	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP	t(2)	F(4,4)	mixP
Smith	1.00	1.00	1.00	0.82	0.88	0.89	0.67	0.75	0.75	0.56	0.66	0.65
Hill	0.66	0.69	0.52	0.44	0.44	0.26	0.40	0.37	0.19	0.41	0.34	0.16
GI	0.97	0.99	0.81	0.59	0.60	0.42	0.49	0.49	0.31	0.46	0.43	0.25
fixed-k	0.73	0.73	0.73	0.73	0.73	0.73	0.67	0.69	0.69	0.59	0.63	0.63

Note: Entries are median biases and root mean squared errors of estimators and coverage and length of confidence intervals in a sample of size n=250 i.i.d. draws from different distributions without shifting. Based on the largest k order statistics. See the main text for a description of the four types of estimators. Based on 5,000 Monte Carlo simulations.

disaster if the GDP (or consumption) declines by more than 10% and collect $n = 157$ observations from 36 countries from 1870 to 2005. Barro and Jin (2011) further apply these data to estimate the tail index of the disaster size distribution and then back out the coefficient of the relative risk aversion by a theoretical model (eq. 2). The authors find that a double power law fits the data well and conclude that the very large observations mainly determine the risk premium. This suggests that the effective observations for learning tail consist of only limited number of largest observations. Columns 4, 7, 10, and 13 in Table 6 also confirm that incorrectly incorporating non-tail observations could lead to substantively biased estimation of the tail index. Thus to main robustness, we implement the fixed-k approach and the other three popular methods investigated in the previous section with the $k = 21$ largest observations, which are treated by Barro and Jin (2011) as stemming from the true tail. The point estimate and confidence interval are reported in Table 7.

Several interesting findings are obtained by comparing different approaches. First, the estimators are substantially different across methods. In particular, the Smith estimator is close to zero, indicating a possibly underestimate as suggested by the Monte Carlo. In the contrary, the Hill and the GI estimators might overestimate the tail index, the latter of which is used in Barro and Jin (2011). Second, we back out the coefficient of relative risk aversion by solving the risk premium condition proposed in Barro and Jin (2011) (eq. 6). Since we consider the tail contains only the largest 21 observations, the disaster probability and the cutoff value (p and b in their notation) are accordingly replaced with 0.0046 and 0.32, respectively. The Hill and the GI estimators lead to estimate very close to 4, similarly as in the original paper, while the fixed-k approach leads to 6.34, a substantively larger coefficient of risk aversion. Finally, standard bootstrap confidence interval may deliver substantively incorrect coverage in small samples, as demonstrated by Fukuchi (1994) and Zelterman (1993).

Barro and Jin (2011) also point out that the largest disasters tend to be missing due to goverment collapse or fighting wars. The Monte Carlo results clearly suggest that the largest statistics are very informative about learning tails and hence we apply the fixed-k approach for truncated data to estimate the tail index and high quantiles. In the dataset, Turkey enters the record after 1923 and since then the only missing data in the sampled 40 countries come from Greece in 1944, Malaysia in 1943-46, Phillipine 1941-45, and Singapore 1940-49. In principle, all these 20 observations can be large disasters but assuming all of them to be larger than the disaster cutoff 0.32, could be too conservative. To have a rough estimate of the missing range, we first fill in the missing spots by linear interpretation of the closest observated years. Assuming the constraction size is monotone in the missing years, there are only 3 possible years that can exihibit contractions larger than 32% (1 in Malaysia and 2 in Phillipine). Thus, we implement the fixed-k approach with $\bar{h} = 3$, that is, the truncation point is at least the 1-3/157 = 98% percentile of the underlying disaster

Table 7: Tail Index Estimates of Macroeconomic Disasters

Method	Smith		Hill		GI		fixed-k	
	Tail Index							
Est.	0.04		0.29		0.27		0.17	
CI	0.00	0.48	0.17	0.41	0.11	0.44	0.00	0.71

Note: Data source: https://scholar.harvard.edu/barro/data_sets. See the main text for a description of the four methods.

distribution. The estimated ξ is then 0.12, which is smaller than in the complete data senario, and the coefficient of risk aversion is backed out as 8.59 with the same disaster probability[4].

5.2 Earthquake Fatality

Catastrophic earthquak incurs very large loss and fatality, so a precise estimate of its (right) tail behavior is important from the perspective of insurance and macroeconomic policy. However, the observations might be truncated due to complicated reasons such as physical limitation and measurement inaccuracy (Burroughs and Tebbens (2001, 2002) and Clark (2013)). In this section, we apply our method for estimating high quantiles with possibly truncated data to the dataset of earthquake fatalities provided by the U.S. Geological Survey, which was also investigated in Beirlant, Alves, and Gomes (2016) and Zou, Davis, and Samorodnitsky (2017). The dataset contains the fatality information of 125 earthquakes causing 1,000 or more deaths from 1900 to 2014.

Given the small sample size, our Monte Carlo suggests that our fixed-k asymptotics works well for k less than or equal to 20 for the DGP examined in the previous section and hence we choose $k = 20$. Beirlant, Alves, and Gomes (2016) estimate the truncation value to be slightly above the sample maximum, so we consider at most the top $1 - 1/n$ proportion of the underlying distribution is truncated ($\bar{h} = 1$). Table 8 depicts the estimate of $1 - h/n$ quantiles of earthquake fatalily with $h = 0.5$ and 5, corresponding to 99.6% and 96% percentiles, respectively.

The fixed-k approach delivers substantially different estimates than the Truncated Pareto approach (TP-WH) proposed by Beirlant, Alves, and Gomes (2016). Two possible explanations are as follows. First, Clark (2013) suggests that the tail index can be 0 or negative, indicating a bounded support. This is ruled out by the TP-WH approach, indicating a possible source of bias. Second, the TP-WH method tends to underestimate the quantile as suggested in Table 4 in the Monte Carlo section. Since we are in the very extreme quantile, a small underestimate in terms of h leads to a substantive underestimate in the level.

[4] Allowing the possible truncation leads to very small change in the disaster probability.

Table 8: Estimates of 1-h/n Quantile of the Earthquake Fataliy

	$h = 0.5$		$h = 5$	
Method	fixed-k	TP-WH	fixed-k	TP-WH
Estimate	1267	316	243	67.1

Note: Data source: http://earthquake.usgs.gov/earthquakes/world/world_deaths.php. See the main text for a description of the two methods.

6 Concluding Remarks

This paper develops a fixed-k approach to estimate tail properties including the tail index, extreme quantile, and tail conditional expectation. The new approach is specially designed for the small sample senario in which existing methods may exhibit substantive bias. The asymptotic validity of the new approach relies on the widely assumed regular variation condition and the extreme value theorem. In particular, this assumption implies that only a fixed number of the largest observations are assumed to stem from the tail, which is more robust and suitable for very small samples.

Furthermore, the fixed-k estimator is constructed to have several attractive properties: (quantile) unbiasedness, invariance/equivariance to location and scale, robustness to deviation from Pareto distribution, and optimality in a well-defined sense. This cost of achieving these advantages is a higher risk, measured by the mean standard error or the mean absolute deviation, indicating the difficulty of obtaining a precise estimate with few tail observations.

A final remark about the fixed-k approach is on the determination of k, which has been widely considered as a challenging question. In principle, there cannot exist a procedure that consistently determines if a certain k is appropriate. As long as we believe the upper k order statistics are approximately drawn from the tail instead of the central part, capturing the dependence structure among these large order statistics by the joint extreme value distribution works better than treating the exceedances as independent Pareto draws. In practice, we may present the results with varying choices of k or combine with other algorithms to choose k based on higher order assumptions (see, for example, Hall (1982)).

Appendix

A.1 Computational Details

The estimators defined in (9), (10), (11), and (14) , and require evaluation of some expectations. Define $\Gamma(\cdot)$ as the Gamma function, $\Gamma(a, z) = \int_z^\infty t^{a-1}e^{-t}dt$ as the incomplete Gamma function, and $b_0(\xi) = -1/\xi$ for $\xi < 0$, and $b(\xi) = \infty$ otherwise. Also define $e(\mathbf{X}^s, s) =$

$\exp\left(-(1+1/\xi)\sum_{i=1}^{k}\log(1+\xi x_i^s s)\right)$. Use the expression for $f_{\mathbf{X}|\xi}$ below (3). Then for a positive $\hat{Q}(\mathbf{X}^s)$, some calculation yields the following expressions.

1. For the complete data case, the density of $\mathbf{X}^s$ is

$$f_{\mathbf{X}^s|\xi}(\mathbf{X}^s) = \Gamma(k)\int_0^{b_0(\xi)} s^{k-2} e(\mathbf{X}^s, s)\,ds.$$

The asymptotic bias terms read

$$\begin{aligned} & E_\xi[\left|(X_1 - X_k)\hat{Q}(\mathbf{X}^s) + X_k - q(\xi,h)\right| |\mathbf{X}^s] f_{\mathbf{X}^s|\xi}(\mathbf{X}^s) \\ = \;& \Gamma(k+1)\int_0^{b_0(\xi)} \left(1+\xi s\hat{Q}(\mathbf{X}^s)\right)^{-1/\xi} s^{k-2} e(\mathbf{X}^s, s)\,ds \end{aligned}$$

and

$$\begin{aligned} & E_\xi\left[(X_1 - X_k)\hat{T}(\mathbf{X}^s) + X_k - \tau(\xi,h)\,|\mathbf{X}^s\right] f_{\mathbf{X}^s|\xi}(\mathbf{X}^s) \\ = \;& \hat{T}(\mathbf{X}^s)\,\Gamma(k-\xi)\int_0^{b_0(\xi)} s^{k-1} e(\mathbf{X}^s, s)\,ds \\ & + \left(\frac{\Gamma(k-\xi)-\Gamma(k)}{\xi} - \tau(\xi,h)\,\Gamma(k)\right)\int_0^{b_0(\xi)} s^{k-2} e(\mathbf{X}^s, s)\,ds. \end{aligned}$$

The risk terms read

$$\begin{aligned} & E_\xi[\left|(X_1 - X_k)\hat{Q}(\mathbf{X}^s) + X_k - q(\xi,h)\right| |\mathbf{X}^s] f_{\mathbf{X}^s|\xi}(\mathbf{X}^s) \\ = \;& |\xi|^{-1}\int_0^{b_0(\xi)} g(s)\, s^{k-2} e(\mathbf{X}^s, s)\,ds \end{aligned}$$

where for $a(s) = 1 + s\xi\hat{Q}(\mathbf{X}^s)$

$$g(s) = \begin{cases} \begin{pmatrix} -h^{-\xi}(\Gamma[k] - 2\Gamma[k, a(s)^{1/\xi} h]) \\ +a(s)\,(\Gamma[k-\xi] - 2\Gamma[k-\xi, a(s)^{1/\xi} h]) \end{pmatrix} & \text{if } a(s) > 0 \\ \left(h^{-\xi}\Gamma[k] - a(s)\,\Gamma[k-\xi]\right) & \text{otherwise,} \end{cases}$$

and

$$\begin{aligned} & E_\xi\left[\left|(X_1 - X_k)\hat{T}(\mathbf{X}^s) + X_k - \tau(\xi,h)\right| |\mathbf{X}^s\right] f_{\mathbf{X}^s} \\ = \;& |\xi|^{-1}\int_0^{b_0(\xi)} \tilde{g}(s)\, s^{k-2} e(\mathbf{X}^s, s)\,ds \end{aligned}$$

where for $a\left(s\right)=1+s\xi\hat{T}\left(\mathbf{X}^{s}\right)$

$$\tilde{g}\left(s\right)=\begin{cases}\begin{pmatrix}-\frac{h^{-\xi}}{1-\xi}(\Gamma[k]-2\Gamma[k,a\left(s\right)^{1/\xi}h\left(1-\xi\right)^{1/\xi}])\\+a\left(s\right)\left(\Gamma[k-\xi]-2\Gamma[k-\xi,a\left(s\right)^{1/\xi}h\left(1-\xi\right)^{1/\xi}]\right)\end{pmatrix} & \text{if } a\left(s\right)>0\\\left(\frac{h^{-\xi}}{1-\xi}\Gamma[k]-a\left(s\right)\Gamma[k-\xi]\right) & \text{otherwise.}\end{cases}$$

2. For the censored data case, define $e\left(\mathbf{X}_{m}^{s},s\right)$ in same way as $e\left(\mathbf{X}^{s},s\right)$. The density of $\mathbf{X}_{m}^{s}$ is

$$f_{\mathbf{X}_{m}^{s}|\xi}\left(\mathbf{X}_{m}^{s}\right)=\frac{\Gamma\left(k+m\right)}{m!}\int^{b_0(\xi)}s^{k-2}\exp\left(-\frac{m}{\xi}\log\left(1+\xi s\right)\right)e\left(\mathbf{X}_{m}^{s},s\right)ds.$$

The asymptotic bias term reads

$$\begin{aligned}&E_{\xi}\left[\left(1+\xi\left(\left(X_{m+1}-X_{m+k}\right)\hat{Q}\left(\mathbf{X}_{m}^{s}\right)+X_{m+k}\right)\right)^{-1/\xi}|\mathbf{X}_{m}^{s}\right]f_{\mathbf{X}_{m}^{s}|\xi}\left(\mathbf{X}_{m}^{s}\right)\\=\;&\frac{\Gamma\left(m+k+1\right)}{m!}\int_{0}^{b_0(\xi)}\left(1+\xi s\hat{Q}\left(\mathbf{X}^{s}\right)\right)^{-1/\xi}s^{k-2}\exp\left(-\frac{m}{\xi}\log\left(1+\xi s\right)\right)e\left(\mathbf{X}_{m}^{s},s\right)ds\end{aligned}$$

and the asymptotic risk term reads

$$\begin{aligned}&E_{\xi}[\left|\left(X_{m+1}-X_{m+k}\right)\hat{Q}\left(\mathbf{X}_{m}^{s}\right)+X_{m+k}-q\left(\xi,h\right)\right||\mathbf{X}_{m}^{s}]f_{\mathbf{X}_{m}^{s}|\xi}\left(\mathbf{X}_{m}^{s}\right)\\=\;&\frac{1}{m!\left|\xi\right|}\int_{0}^{b_0(\xi)}g\left(k,h,\xi,s\right)\exp\left(-\frac{m}{\xi}\log\left(1+\xi s\right)\right)e\left(\mathbf{X}_{m}^{s},s\right)s^{k-2}ds\end{aligned}$$

where for $a\left(s\right)=\left(1+s\xi\hat{Q}\left(\mathbf{x}^{s}\right)\right)$,

$$g\left(s\right)=\begin{cases}\begin{pmatrix}-h^{-\xi}(\Gamma[k+m]-2\Gamma[k+m,a\left(s\right)^{1/\xi}h])\\+a\left(s\right)\left(\Gamma[k+m-\xi]-2\Gamma[k+m-\xi,a\left(s\right)^{1/\xi}h]\right)\end{pmatrix} & \text{if } a\left(s\right)>0\\\left(h^{-\xi}\Gamma[k+m]-a\left(s\right)\Gamma[k+m-\xi]\right) & \text{otherwise.}\end{cases}$$

3. For the truncated data case, define $e\left(\widetilde{\mathbf{X}}^{s},s\right)$ in the same way as $e\left(\mathbf{X}^{s},s\right)$. The density of $\widetilde{\mathbf{X}}^{s}$ is

$$f_{\widetilde{\mathbf{X}}^{s}|\xi,\tilde{h}}\left(\widetilde{\mathbf{X}}^{s}\right)=\exp\left(\tilde{h}\right)\int_{0}^{b_0(\xi)}\Gamma\left(k,\tilde{h}\left(1+\xi s\right)^{1/\xi}\right)s^{k-2}e\left(\widetilde{\mathbf{X}}^{s},s\right)ds.$$

The asymptotic bias term reads

$$\begin{aligned}&E_{\xi,\tilde{h}}\left[\left(1+\xi\left(\left(\tilde{X}_{1}-\tilde{X}_{k}\right)\hat{Q}\left(\widetilde{\mathbf{X}}^{s}\right)+\tilde{X}_{k}\right)\right)^{-1/\xi}|\widetilde{\mathbf{X}}^{s}\right]f_{\widetilde{\mathbf{X}}^{s}|\xi,\tilde{h}}\left(\widetilde{\mathbf{X}}^{s}\right)\\=\;&\exp\left(\tilde{h}\right)\int_{0}^{b_0(\xi)}\Gamma\left(k+1,\tilde{h}\left(1+\xi s\right)^{1/\xi}\right)\left(1+\xi s\hat{Q}\left(\widetilde{\mathbf{X}}^{s}\right)\right)^{-1/\xi}s^{k-2}e\left(\widetilde{\mathbf{X}}^{s},s\right)ds\end{aligned}$$

and the asymptotic risk term reads

$$\begin{aligned} & E_{\xi,\tilde{h}}[\left|\left(\tilde{X}_1-\tilde{X}_k\right)\hat{Q}\left(\widetilde{\mathbf{X}}^s\right)+\tilde{X}_k-q\left(\xi,h\right)\right||\widetilde{\mathbf{X}}^s]f_{\widetilde{\mathbf{X}}^s|\xi,\tilde{h}}\left(\widetilde{\mathbf{X}}^s\right) \\ = \quad & \exp\left(\tilde{h}\right)|\xi|^{-2}\int_0^{b_0(\xi)}g\left(k,h,\xi,s\right)e\left(\widetilde{\mathbf{X}}^s,s\right)s^{k-2}ds \end{aligned}$$

where for $a\left(s\right)=1+s\xi\hat{Q}\left(\widetilde{\mathbf{X}}^s\right)$

$$\begin{aligned} & g\left(k,h,\xi,s\right) \\ = & \left\{\begin{array}{c} \xi\left(-h^{-\xi}(\Gamma[k,\tilde{h}\left(1+\xi s\right)^{1/\xi}]-2\Gamma[k,a\left(s\right)^{1/\xi}h])+a\left(s\right)(\Gamma[k-\xi,\tilde{h}\left(1+\xi s\right)^{1/\xi}]-2\Gamma[k-\xi,a\left(s\right)^{1/\xi}h])\right) \\ \text{if } \xi>0, a\left(s\right)>h^{-\xi}\left(1+\xi s\right)/\tilde{h}^{-\xi} \text{ or } \xi<0, 0<a\left(s\right)<h^{-\xi}\left(1+\xi s\right)/\tilde{h}^{-\xi} \\ \xi\left(h^{-\xi}\Gamma[k,\tilde{h}\left(1+\xi s\right)^{1/\xi}]-a\left(s\right)(\Gamma[k-\xi,\tilde{h}\left(1+\xi s\right)^{1/\xi}])\right) \\ \text{if } \xi>0, 0<a\left(s\right)<h^{-\xi}\left(1+\xi s\right)/\tilde{h}^{-\xi} \text{ or } \xi<0, a\left(s\right)<0 \text{ or } \xi<0, a\left(s\right)>h^{-\xi}\left(1+\xi s\right)/\tilde{h}^{-\xi}. \end{array}\right. \end{aligned}$$

We evaluate these by numerical quadrature.

To determine the suitable Lagrangian multipliers λ λ_m and $\tilde{\lambda}$ for estimating extreme quantile and TCE, we follow the algorithm suggested by Müller and Wang (2015). For the complete and censored data case, we restrict λ to be discrete distributions with support on $\Xi=\{-1/2,-1/2+1/19,\ldots,1/2\}$, and determine the 20 point masses by fixed-point iterations based on importance sample Monte Carlo estimates of bias. In particular, we simulate the biases with 5,000 i.i.d. draws from a proposal with ξ randomly drawn from Ξ, and iteratively increase or decrease the 20 point masses on Ξ as a function of whether the (estimated) bias given that value of ξ is larger or smaller than zero. Stop this iteration until the bias for all values ξ is smaller than a prespecified tolerance ε, and then the resulting discrete distribution is a candidate for the Lagrangian multiplier. The tolerance is set to be 0.03. A smaller tolerance can be used at the cost of more Monte Carlo draws and a longer computation time. Regarding the data truncation model, we take $\Xi\times H=\{-1/2,-1/2+1/9,\ldots,1/2\}\times\{0,0.5,1.0,1.5,2.0\}$ and compute the Lagrangian multipliers on this 10×5 grid with the tolerance set to be 0.05 for $h=0.5$ and 0.15 for $h=5$. For the weighting function W, we simply use a uniform weight on Ξ in the complete and censoring models, and use the uniform weight on Ξ product a weight proportional to $\exp\left(\tilde{h}\right)$ on H for the truncated data model. Note that the choice of $\xi\leq 1/2$ covers all the distributions with a finite second moment, and that our approach can be easily extended to cover larger range of ξ. See Müller and Wang (2015) for more details about the properties and implementation of this algorithm.

For any given k, h, and m, the Lagrangian multipliers only need to be determined once. Conditional on them, the estimator is readily computed from (9), (10), and (11). The tables of the Lagrangian multipliers and the corresponding Matlab code are provided on our website: https://sites.google.com/site/yulongwanghome/.

To estimate ξ, given most empirical applications involve data with infinite support and finite expecation, we take Ξ to be $[0,1]$, which can be easily extended to any subset of $\mathbb{R}$. Regarding $W(\cdot)$, we again choose a uniform weight which can be treated as a flat prior from a Bayesian perspective. The numerical solution is based on descritizing Ξ into $\{1/100, 2/100, ..., 1\}$. By continuity, the estimator constructed in (14) is asymptotically median unbiased up to only Monte Carlo accuracy.

A.2 Data with Known Censoring or Truncation Value

In the censored data case, if the censoring value is also observed, we may still consider the maximal invariant statistic introduced in Section 2.2. Given the censoring value, denoted by c, is between $Y_{(m)}$ and $Y_{(m+1)}$, we can model it as $\left(Y_{(m+1)} - Y_{(m+k)}\right) t + Y_{(m+k)}$ for $t > 1$, and derive

$$\left(\frac{Y_{(1)} - Y_{(m+k)}}{Y_{(m+1)} - Y_{(m+k)}}, ..., \frac{Y_{(m+1)} - Y_{(m+k)}}{Y_{(m+1)} - Y_{(m+k)}} \right) \Rightarrow \dot{\mathbf{X}}^s$$

where $\dot{X}_i^s \geq t = \frac{c - Y_{(m+k)}}{Y_{(m+1)} - Y_{(m+k)}} > 1$ for $i \leq m$ and $\dot{X}_i^s \in [0,1]$ for $i > m$. The density of $\dot{\mathbf{X}}_m^s = \left(\dot{X}_{m+1}^s, ..., \dot{X}_{m+k}^s\right)$ can be derived as follows

$$\begin{aligned} f_{\dot{\mathbf{X}}_m^s|\xi}\left(\dot{\mathbf{X}}_m^s\right) &= \int ... \int_{\dot{X}_1^s \geq \dot{X}_2^s \geq ... \dot{X}_m^s \geq t} f_{\dot{\mathbf{X}}^s|\xi}\left(\dot{\mathbf{X}}^s\right) \left(d\dot{X}_1^s ... d\dot{X}_m^s\right) \\ &= \Gamma(m+k) \int_0^{b_0(\xi)} \frac{s^{k-2}}{m!} (1 + \xi t s)^{-m/\xi} e\left(\dot{\mathbf{X}}_m^s, s\right) ds, \end{aligned}$$

where $e\left(\dot{\mathbf{X}}_m^s, s\right)$ is defined in the same way as $e(\mathbf{X}^s, s)$ in the previous section, and then the asymptotic bias and risk can be derived in a similar fashion.

In the truncated data case with a known truncation value c, we can model $c = Q\left(F, 1 - \tilde{h}/n\right)$ for some unknown $\tilde{h}$ since the quantile function $Q(\cdot)$ is unknown (and not easy to estimate). Consider

$$\begin{aligned} & \left(\frac{Y_{(1)} - Y_{(k)}}{c - Y_{(k)}}, ..., \frac{Y_{(k)} - Y_{(k)}}{c - Y_{(k)}} \right) \\ \Rightarrow \quad & \widetilde{\mathbf{X}}^s(c) = \left(\frac{\tilde{X}_1 - \tilde{X}_k}{q\left(\xi, \tilde{h}\right) - \tilde{X}_k}, ..., \frac{\tilde{X}_k - \tilde{X}_k}{q\left(\xi, \tilde{h}\right) - \tilde{X}_k} \right). \end{aligned}$$

Given the density of $\widetilde{\mathbf{X}}$ remains the same as stated in Lemma 1, the density of $\widetilde{\mathbf{X}}^s(c)$ can be derived by change of variable as

$$f_{\widetilde{\mathbf{X}}^{s}(c)|\xi,\tilde{h}}\left(\widetilde{\mathbf{X}}^{s}\left(c\right)\right)=\exp\left(\tilde{h}\right)\int_{0}^{b_{1}(\xi)}\Gamma\left(k,\tilde{h}\left(1+\xi s\right)^{1/\xi}\right)s^{k-2}e\left(\widetilde{\mathbf{X}}^{s},s\right)ds$$

where $b_1\left(\xi\right)=\infty$ if $\xi>0$ and $-1/\left(\xi\tilde{X}_1^s\left(c\right)\right)$ if $\xi<0$, and similarly for the asymptotic bias and risk.

A.3 Proof

Proof of Lemma 1. Given the CDF (3), it is equivalent to show that

$$\begin{aligned}&\left(\frac{Y_{(1)}-b_n}{a_n},...,\frac{Y_{(k)}-b_n}{a_n}\right)\\ \Rightarrow\ &\left(\frac{\left(\tilde{h}+E_1^*\right)^{-\xi}-1}{\xi},\frac{\left(\tilde{h}+E_1^*+E_2^*\right)^{-\xi}-1}{\xi},...,\frac{\left(\tilde{h}+E_1^*+E_2^*+\cdots+E_k^*\right)^{-\xi}-1}{\xi}\right)\end{aligned}$$

where $E_1^*,...,E_k^*$ are i.i.d. standard exponentials, and $(x^{-\xi}-1)/\xi$ is intepreted as $-\log\left(x\right)$ if $\xi=0$. To show this, denote $F^c\left(\cdot\right)$ as the truncated CDF by $Q\left(F,1-\tilde{h}/n\right)$, that is, $F^c\left(x\right)=F\left(x\right)/\left(1-\tilde{h}/n\right)$ for $x\leq Q\left(F,1-\tilde{h}/n\right)$. Define $U\left(t\right)=F^{-1}\left(1-1/t\right)$ and similarly for $U^c\left(t\right)$. Then

$$\left(Y_{(1)},...,Y_{(k)}\right)\overset{d}{=}\left(U^c\left(\frac{1}{1-e^{-E_{1,n}}}\right),U^c\left(\frac{1}{1-e^{-E_{2,n}}}\right),...,U^c\left(\frac{1}{1-e^{-E_{k,n}}}\right)\right)$$

where $E_{1,n}...E_{k,n}$ are order statistics of n i.i.d. standard exponentials. Note that $U^c\left(\frac{1}{1-e^{-E_{1,n}}}\right)=U\left(\frac{1}{1-\left(1-\tilde{h}/n\right)e^{-E_{1,n}}}\right)$. Then the proof follows from the same argument of Theorem 2.1.1 of de Haan and Ferreira (2007) and the fact that $n\left(1-\left(1-\tilde{h}/n\right)\exp\left(-x/n\right)\right)\to x+\tilde{h}$. ■

Proof of Theorem 1. For notational ease, we prove the theorem without the autoregression part, i.e., assuming $\phi_1=...=\phi_{\tilde{p}}=\bar{\mu}=0$. The proof with it follows the same logic with more tedious algebra, and is hence omitted. We start with the simplest GARCH(1,1) case, i.e., $\sigma_t^2=\alpha_0+\alpha_1 y_{t-1}^2+\beta\sigma_{t-1}^2$. By iteration, we have $\sigma_t^2=\sum_{l=0}^{t-1}\beta^l(\alpha_0+\alpha_1 y_{t-1}^2)$ and plugging in the PML estimator, denoted as $\left(\hat{\alpha}_0,\hat{\alpha}_1,\hat{\beta}\right)$, of the coefficients leads to an estimator of $\hat{\sigma}_t^2$, that is, $\hat{\sigma}_t^2=\sum_{l=0}^{t-1}\hat{\beta}^l(\hat{\alpha}_0+\hat{\alpha}_1 y_{t-1}^2)$.

Note that

$$\sup_{0\leq w}|\frac{\hat{a}+w\hat{b}}{a+wb}-1|\ \leq\ \frac{\max(|a-\hat{a}|,|b-\hat{b}|)}{\min(a,b)}$$

$$\begin{aligned}\sup_{0\leq w}|\frac{\hat{a}\hat{c}+w\hat{b}}{ac+wb}-1| &\leq \frac{\max(|ac-\hat{a}\hat{c}|,|b-\hat{b}|)}{\min(ac,b)}\\ &\leq \frac{\max(a|c-\hat{c}|,c|a-\hat{a}|,|c-\hat{c}|\cdot|a-\hat{a}|,|b-\hat{b}|)}{\min(ac,b)}\end{aligned}$$

since

$$\begin{aligned}ac-\hat{a}\hat{c} &= a(c-\hat{c}+\hat{c})-\hat{a}\hat{c}\\ &= a(c-\hat{c})+\hat{c}(a-\hat{a})\\ &= a(c-\hat{c})+(\hat{c}-c+c)(a-\hat{a})\\ &= a(c-\hat{c})+c(a-\hat{a})+(\hat{c}-c)(a-\hat{a}).\end{aligned}$$

Thus, by repeated applications of these inequalities, we have

$$\sup_{y_{t-1}^2}|\frac{\hat{\sigma}_t^2}{\sigma_t^2}-1| \leq \frac{\max(\beta|\alpha_0-\hat{\alpha}_0|,\beta|\alpha_1-\hat{\alpha}_1|,\max(\alpha_1,\alpha_0)\sup_l|\beta^l-\hat{\beta}^l|)}{\alpha_0}$$

which converges to zero in probability by consistency of the PML estimator and $\alpha_0 > 0$.

Thus, $\sup_t|\hat{\sigma}_t^2/\sigma_t^2-1| \xrightarrow{p} 0$, and also $\sup_t|\hat{\sigma}_t/\sigma_t-1| \xrightarrow{p} 0$. Let $Y_t = Z_t/\sigma_t$ and $\hat{Y}_t = Z_t/\hat{\sigma}_t$, so that $\hat{Y}_t = Y_t\sigma_t/\hat{\sigma}_t$. Then these results also imply $\sup_t|\hat{Y}_t/Y_t-1| \xrightarrow{p} 0$. Now suppose Y_t satisfies (4), that is,

$$\left(\frac{Y_{(1)}-b_n}{a_n},...,\frac{Y_{(k)}-b_n}{a_n}\right) \Rightarrow \mathbf{X}$$

where $\mathbf{X}$ is jointly extreme value distributed as below (4). Let $I = (I_1,\ldots,I_k) \in \{1,\ldots,T\}^k$ be the k random indices such that $Y_{n:n-j+1} = Y_{I_j}$, $j = 1,\ldots,k$, and let $\hat{I}$ be the corresponding indices such that $\hat{Y}_{n:n-j+1} = \hat{Y}_{\hat{I}_j}$. We claim that $I - \hat{I} \xrightarrow{p} 0$. Suppose otherwise, then (4) implies that $\sup_t|\hat{Y}_t/Y_t-1|$ is not $o_p(a_n)$. This contradicts $\sup_t|\hat{Y}_t/Y_t-1| \xrightarrow{p} 0$ (since $a_n \to \infty$).
Thus,

$$\begin{aligned}&\left(\frac{\hat{Y}_{\hat{I}_1}-b_n}{a_n},...,\frac{\hat{Y}_{\hat{I}_k}-b_n}{a_n}\right)\\ =\ &\left(\frac{\hat{Y}_{I_1}-b_n}{a_n},...,\frac{\hat{Y}_{I_k}-b_n}{a_n}\right)+o_p(1)\\ =\ &\operatorname{diag}(\frac{\sigma_{I_1}}{\hat{\sigma}_{I_1}},\ldots,\frac{\sigma_{I_k}}{\hat{\sigma}_{I_k}})\left(\frac{Y_{I_1}\sigma_{I_1}/\hat{\sigma}_{I_1}-b_n}{a_n},...,\frac{Y_{I_k}\sigma_{I_k}/\hat{\sigma}_{I_k}-b_n}{a_n}\right)'+o_p(1)\\ \Rightarrow\ &\mathbf{X}\end{aligned}$$

by the Slutzky's theorem.

Now for GARCH(p,q) model, we have $\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i y_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2$ with $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_i \geq 0$ and $\sum_{i=1}^{p} \beta_i < 1$.

Let $B(x) = 1 - \beta_1 x - \beta_2 x^2 -, ..., -\beta_p x^p$ and $A(x) = \alpha_1 x + \alpha_2 x^2 +, ..., +\alpha_q x^q$, then we have $B(L)\sigma_t^2 = \alpha_0 + A(L)y_t^2$

$$
\begin{aligned}
\sup_{y_{t-1}^2,\ldots,y_{t-q}^2} \left| \frac{\hat{\sigma}_t^2}{\sigma_t^2} - 1 \right| &\leq \sup_{y_{t-1}^2,\ldots,y_{t-q}^2} \frac{1}{a_0} \left| \hat{\alpha}_0 \hat{B}^{-1}(1) - \alpha_0 B^{-1}(1) + \left(\hat{A}(L)\hat{B}^{-1}(L) - A(L)B^{-1}(L) \right) y_t^2 \right| \\
&\leq \frac{1}{a_0} \left| \hat{\alpha}_0 \hat{B}^{-1}(1) - \alpha_0 B^{-1}(1) \right| + \sup_{y_{t-1}^2,\ldots,y_{t-q}^2} \frac{1}{a_0} \left| \left(\hat{A}(L)\hat{B}^{-1}(L) - A(L)B^{-1}(L) \right) y_t^2 \right| \\
&\leq \left| \hat{B}^{-1}(1) - B^{-1}(1) \right| + \frac{\hat{B}^{-1}(1)}{\alpha_0} \left| \hat{\alpha}_0 - \alpha_0 \right| \\
&\quad + \sup_{y_{t-1}^2,\ldots,y_{t-q}^2} \frac{1}{a_0} \left| \left(\hat{A}(L)\hat{B}^{-1}(L) - A(L)B^{-1}(L) \right) y_t^2 \right| \\
&\leq o_p(1) + \frac{(\max_i \hat{\alpha}_i)}{a_0} \sup_{y_{t-1}^2,\ldots,y_{t-q}^2} \left| \left(\hat{B}^{-1}(L) - B^{-1}(L) \right) y_t^2 \right| \\
&\quad + \frac{\left(\max B^{-1}(L) \right)}{a_0} \sup_{y_{t-1}^2,\ldots,y_{t-q}^2} \left| \left(\hat{A}(L) - A(L) \right) y_t^2 \right| \\
&= o_p(1)
\end{aligned}
$$

where $B^{-1}(L) = \frac{1}{B(L)} = \sum_{j=1}^{\infty} b_j L^j$ with coefficients b_j decaying exponentially fast and $\max B^{-1}(L)$ denotes the maximum of $\{b_1, b_2, ...\}$. In the last inequality, we implicitly use the fact that the consistency of $\hat{B}$ implies the consistency of $\hat{B}^{-1}$. Then the rest of proof is the same as in the GARCH(1,1) case. ■

References

Aban, I. B., and M. M. Meerschaert (2001): "Shifted Hill's estimator for heavy tails," *Communications in Statistics-Simulation and Computation*, 30(4), 949–962.

Aban, I. B., M. M. Meerschaert, and A. K. Panorska (2006): "Parameter Estimation for the Truncated Pareto Distribution," *Journal of the American Statistical Association*, 101(473), 270–277.

Alves, I. F., M. I. Gomes, L. de Haan, and C. Neves (2009): "Mixed Moment Estimator and Location Invariant Alternatives," *Extremes*, 12, 149–185.

Anbarci, N., M. Escaleras, and C. A. Register (2005): "Earthquake fatalities: the interaction of nature and political economy," *Journal of Public Economics*, 89, 1907–1933.

Ancona-Navarrete, M. A., and J. A. Tawn (2000): "A comparsion of methods for estimating theExtremal Index," *Extremes*, 3(1), 5–38.

Andrews, D. W. K. (1993): "Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models," *Econometrica*, 61, 139–165.

Andrews, D. W. K., and W. Ploberger (1994): "Optimal Tests When a Nuisance Parameter Is Present Only under the Alternative," *Econometrica*, 62, 1383–1414.

Arnold, B. C., N. Balakrishnan, and H. H. N. Nagaraja (1992): *A First Course in Order Statistics.* Siam.

Barro, R. J., and T. Jin (2011): "On the Size Distribution of Macroeconomic Disasters," *Econometrica*, 79(5), 1567–1589.

Barro, R. J., and J. F. Ursua (2008): "Macroeconomic Crisis Since 1870," *Brookings Papers on Economic Activity*, (255-335).

Basel Committee on Banking Supervision (2013): "Fundamental Review of the Trading Book: A Revised Market Risk Framework," *Consultative Document.*

Beirlant, J., I. F. Alves, and M. I. Gomes (2016): "Tail Fitting for Truncated and non-truncated Pareto-type Distributions," *Extremes*, 19(3), 429–462.

Beirlant, J., F. Caeiro, and M. I. Gomes (2012): “An Overview and Open Reseach Topics in Statistics of Univariate Extremes,” *Revstat*, 10(1), 1–31.

Bollerslev, T., and V. Todorov (2011a): “Estimation of Jump Tails,” *Econometrica*, 79(6), 1727–1783.

——— (2011b): “Tails, fears, and risk premia,” *The Journal of Finance*, 66(6), 2165–2211.

Burroughs, S. M., and S. F. Tebbens (2001): “Upper-Truncated Power Laws in Natural Systems,” *Pure and Applied Geophysics*, 158(4), 741–757.

——— (2002): “The Upper-Truncated Power Law Applied to Earthquake Cumulative Frequency-Magnitude Distributions: Evidence for a Time-Independent Scaling Parameter.,” *Bulletin of the Seismological Society of America*, 92(8), 2983Ű2993.

Chernick, M. R., T. Hsing, and W. P. McCormick (1991): “Calculating the EExtrema Index for a Class of Stationary Sequences,” *Advances in Applied Probability*, 23(4), 835–850.

Chernobai, A. S., S. T. Rachev, and F. J. Fabozzi (2012): “Testing for the Goodness of Fit,” *Operational Risk: A Guide to Basel II Capital Requirements, Models, and Analysis, John Wiley & Sons, Inc., Hoboken, NJ, USA.*

Clark, D. R. (2013): “A Note on the Upper-Truncated Pareto Distribution,” *Casualty Actuarial Society E-Forum.*

Davis, R. (1985): “Reviewed Work: Extremes and Related Properties of Random Sequences and Processes.,” *Journal of the American Statistical Association*, 80(389), 251.

de Haan, L., and A. Ferreira (2007): *Extreme Value Theory: An Introduction.* Springer Science and Business Media, New York.

de Haan, L., and L. Peng (1998): “Comparison of Tail Endex Estimators,” *Statistica Neerlandica*, 52, 60–70.

de Haan, L., and H. Rootzén (1993): “On the estimation of high quantiles,” *Journal of StatisticalPlanning and Inference*, 35, 1–13.

Dekkers, A. L. M., and L. de Haan (1989): “On the estimation of the extreme-value index and large quantile estimation,” *Annals of Statistics*, 17, 1795–1832.

Diebold, F. X., T. Schuermann, and J. D. Stroughair (2000): "Pitfalls and opportunities in the use of extreme value theory in risk management," *Journal of Risk Finance*, 1(2), 30–35.

Elliott, G., U. K. Müller, and M. W. Watson (2015): "Nearly Optimal Tests When a Nuisance Parameter is Present Under the Null Hypothesis," *Econometrica*, 83, 771–811.

Embrechts, P., C. Klupperberg, and T. Mikosch (1997): *Modelling extremal events for insurance and finance*. Springer, New York.

Engle, R. F., and S. Manganelli (2004): "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles," *Journal of Business & Economic Statistics*, 22(4), 367–381.

Fissler, T., and J. F. Ziegel (2016): "Higher order elicitability and Osbandís principle," *Annuals of Statistics.*

Fukuchi, J. (1994): "Bootstrapping extremes of random variables," Ph.D. thesis, Iowa State University.

Gabaix, X., and R. Ibragimov (2011): "Rank-1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents," *Journal of Business & Economic Statistics*, 29(1), 24–39.

Gomes, M. I., and A. Guillou (2015): "Extreme Value Theory and Statistics of Univariate Extreme: a Review," *International Economic Review*, 83, 263–292.

Groisman, P. Y., R. W. Knight, T. R. Karl, D. R. Easterling, B. Sun, and J. H. Lawrimore (2004): "Contemporary Changes of the Hydrological Cycle Over the Contiguous United States: Trends Derived From in Situ Observations," *Journal of Hydrometeorology*, 5, 64–85.

Hall, P. (1982): "On Some Simple Estimates of an Exponent of Regular Variation," *Journal of Royal Statistic Society, Series B.*

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

Hsing, T. (1993): "Extremal Index Estimation for a Weakly Dependent Stationary Sequence," *Annuals of Statistics*, 21(4), 2043–2071.

Huisman, R., K. G. Koedijk, C. J. M. Kool, and F. Palm (2001): "Tail-index Estimates in Small Samples," *Journal of Business & Economic Statistics*, 19(1), 208–216.

Jenkins, S. P., R. V. Burkhauser, S. Feng, and J. Larrimore (2010): "Measuring Inequality Using Censored Data: A Multiple-imputation Approach to Estimation and Inference," *Journal of the Royal Statistical Society Series A*, 174(1), 63–81.

Jorion, P. (2007): *Value at Risk - The New Benchmark for Managing Financial Risk.* McGraw-Hill.

Kahn, M. E. (2005): "The Death Toll from Natural Disasters: The Role of Income, Geography, and Institutions," *Review of Economics and Statistics*, 87(2), 271–284.

Kearns, P., and A. Pagan (1997): "Estimating the Density Tail Index for Financial Time Series," *Review of Economics and Statistics*, 79, 171–175.

Kuester, K., S. Mittnik, and M. S. Paolella (2006): "Value-at-Risk Prediction: A Comparison of Alternative Strategies," *Journal of Financial Econometrics*, 4(1), 53–89.

Leadbetter, M. R. (1974): "On Extreme Values in Stationary Sequences," *Probability Theory and Related Fields*, 28(4), 289–303.

——— (1983): "Extremes and local dependence in stationary sequences," *Probability Theory and Related Fields*, 65(2), 291–306.

Lehmann, E. L., and J. P. Romano (2005): *Testing Statistical Hypothesis.* Springer, New York.

Malamud, B. D., G. Morein, and D. L. Turcotte (1998): "Forest Fire: An Example of Self-Organized Critical Behavior," *Science*, 281, 1840–1842.

McNeil, A., and R. Frey (2000): "Estimation of Tail-related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach," *Journal of Emprical Finance*, 7, 271–300.

Mikosch, T., and C. Starica (2000): "Limit theory for the sample autocorrelations and extremes of a GARCH(1,1) process," *Annals of Statistics*, 24, 1427–1451.

Müller, U. K., and Y. Wang (2015): "Nearly Weighted Risk Minimal Unbiased Estimation," *Working Paper.*

——— (2017): "Fixed-k Asymptotic Inference about Tail Properties," *the Journal of the American Statistical Association*, 112, 1134–1143.

O'Brien, G. (1987): "Extreme Values for Stationary and Markov Sequences," *Annals of Probability*, 15(1), 281–291.

Patton, A. J., J. Ziegel, and R. Chen (2017): "Dynamic Semiparametric Models for Expected Shortfall (and Value-at-Risk)," working paper.

Pickands, III, J. (1975): "Statistical inference using extreme order statistics," *Annals of Statistics*, 3(1), 119–131.

Reiss, R. D., and M. Thomas (2007): *Statistical Analyis of Extreme Values*. Birkhäuser Verlag, Basel.

Resnick, S. (2007): *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media.

Smith, R. L. (1987): "Estimating Tails of Probability Distributions," *Annals of Statistics*, 15, 1174–1207.

Stock, J. H., and M. W. Watson (1998): "Median Unbiased Estimation of Coefficient Variance in a Time-Varying Parameter Model," *Journal of the American Statistical Association*, 93, 349–358.

Süveges, M. (2007): "Likelihood Estimation of the EExtrema Index," *Extremes*, 10(1), 41–55.

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

Zelterman, D. (1993): "A Semiparametric Bootstrap Technique for Simulating Extreme Order Statistics," *Journal of the American Statistical Association*, 88(422), 477–485.

Zou, J., R. Davis, and G. Samorodnitsky (2017): "Extreme Value Analyis without the Largest Values: What Can Be Done?," *Working Paper*.