Extremes
https://doi.org/10.1007/s10687-021-00429-z

Handling missing extremes in tail estimation

Hui Xu[1] · **Richard Davis**[2] · **Gennady Samorodnitsky**[3]

Received: 9 June 2020 / Revised: 7 September 2021 / Accepted: 8 September 2021

Abstract

In some data sets, it may be the case that a portion of the extreme observations is missing. This might arise in cases where the extreme observations are just not available or are imprecisely measured. For example, considering human lifetimes, a topic of recent interest, birth certificates of centenarians may not even exist and many such individuals may not even be included in the data sets that are currently available. In essence, one does not have a clear record of the largest lifetimes of human populations. If there are missing extreme observations, then the assessment of risk can be severely underestimated resulting in rare events occurring more often than originally thought. In concrete terms, this may mean a 500 year flood is in fact a 100 (or even a 20) year flood. In this paper, we present methods for estimating the number of missing extremes together with the tail index associated with tail heaviness of the data. Ignoring one or the other can severely impact the estimation of risk. Our estimates are based on the HEWE (Hill estimate without extremes) of the tail index that adjusts for missing extremes. Based on a functional convergence of this process to a limit process, we consider an asymptotic likelihood-based procedure for estimating both the number of missing extremes and the tail index. We derive the asymptotic distribution of the resulting estimates. By artificially removing segments of extremes in the data, this methodology can be used for assessing the reliability of the underlying assumptions that are imposed on the data.

Keywords Heavy tails · Regular variation · Missing extremes · Tail estimation

Mathematics Subject Classification Primary 62G32 · 60G70

Richard Davis: This research was supported in part by NSF Grant DMS-2015379. Gennady Samorodnitsky: This research was partially supported by the ARO Grant W911NF-18 -10318 at Cornell University.

✉ Richard Davis
rdavis@stat.columbia.edu

Extended author information available on the last page of the article

1 Introduction

For modeling heavy-tailed data, the typical operating assumption is that the tails of the common distribution function F are regularly varying. That is,

$$\frac{\bar{F}(tx)}{\bar{F}(t)} \to x^{-\alpha} \tag{1.1}$$

as $t \to \infty$ for all $x > 0$, where $\alpha > 0$ and $\bar{F}(t) = 1 - F(t)$ is the survival function. The tail index α which governs how heavy the tail is, with smaller α indicating heavier tails, is often the key parameter of interest in applications. The ratio in (1.1) corresponds to the risk probability of $P(X > tx|X > t) \sim x^{-\alpha}$ for large t and $x \geq 1$. In fact, the generalized Pareto distribution (GPD), for heavy-tailed distributions essentially originates from this equation:

$$P(X > t(1 + x)|X > t) \sim (1 + x)^{-1/\gamma} \quad x \geq 0 \tag{1.2}$$

where $\gamma = 1/\alpha > 0$ is known as *the shape parameter*.[1]

The most commonly used estimator of γ is the Hill estimator defined by

$$H_n(k) = \frac{1}{k}\sum_{i=1}^{k} \log X_{(i)} - \log X_{(k+1)},$$

where $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ are the order statistics of an independent and identically distributed (iid) sample $X_1, X_2, \ldots, X_n \sim F$. See Hill (1975) and Drees et al. (2000) for further discussion on this estimator. The Hill estimator is weakly consistent for estimating γ provided the number of order statistics $k = k(n)$ used in estimating γ satisfies, $k \to \infty$ and $k/n \to 0$ as $n \to \infty$ (see for example de Haan and Ferreira 2006.)

The principal goal of this research is to provide estimates of γ and the number of missing extremes in the case when some of the extreme values in the data are missing. As noted in Zou et al. (2019), if some of the extremes in the sample are missing and this is ignored in an estimation procedure, then the tails of the distribution will be underestimated, i.e., they will appear to be lighter than they really are. In terms of risk calculations such as estimating large quantiles, these would be severely underestimated if the estimate of γ is too small. With missing extremes, the plot of the Hill estimate as a function of the number k of upper order statistics tends to be increasing and much smoother than without the missing values.

In Zou et al. (2019), the in-degree distribution from a snapshot on October 19, 2012 of the social network Google+ was examined. The Google+ data, which was owned and operated by Google, consisted of 76,438,791 nodes (registered users) and 1,442,504,499 edges (directed connections). The in-degree of each user is the number of other users following the user. The degree distributions in natural and

[1] The general form of the GPD that combines light- and heavy-tailed cases has a similar form, see de Haan and Ferreira (2006).

social networks are often heavy-tailed (see Newman 2010). Based on the analysis in Zou et al. (2019), it was estimated that around 150 extreme in-degree values were missing, which raises the question of whether these values were excluded from the Google+ data set provided to the researchers.

In addition to detecting possible manipulation of data, being aware of the possibility of missing extremes, and developing tools for modeling and analyzing data in the presence of missing extremes, is important in a variety of fields. This includes analysis of natural disasters such as earthquakes, forest fires and floods for which extreme values might be missing due to difficulty in data collection and in an actuarial science context where claims of extremely large amounts might be covered by a reinsurance company and not included in the claims total (Embrechts et al. 1997; Benchaira et al. 2016). In short, some of the extremes may be just under-reported.

This research builds on the work in Zou et al. (2019), in which an *adjusted* Hill estimator, called the *Hill Estimator Without Extremes* (HEWE) is defined that allows for the possibility of missing extreme observations. The HEWE estimator, $H_n(\cdot)$ is a process on $(0, \infty)$ that, suitably normalized, converges in law to a Gaussian process $G(\cdot)$ whose covariance function depends on the shape parameter γ, and on δ, a parameter related to the number of missing extremes. By approximating the distribution of $(H_n(\theta_1), \dots, H_n(\theta_m))$ for m distinct θ values, by the distribution of the limiting process, i.e., the distribution of $(G(\theta_1), \dots, G(\theta_m))$, one can use an approximate likelihood procedure to estimate γ and δ. The main contribution of this paper is to develop rigorous statistical procedures to estimate these parameters and investigate the properties of the resulting estimators. Only a heuristic estimation procedure was used in Zou et al. (2019), but the ideas from that paper, in particular, the strong approximation results are used in the present paper.

We suggest two implementations of this estimation procedure. The first uses a fixed number of points θ in which the Hill estimator is evaluated. In the second, where we assume that the observations come from a Pareto sample, the number of points increases to infinity with the sample size. For both estimators we show consistency and asymptotic normality. The key to proving our results for the first implementation of the procedure is the strong approximation of the HEWE process $H_n(\cdot)$ derived in Zou et al. (2019).

A limitation in our modeling framework is that we assume that a consecutive block of the largest observations is missing. A more realistic assumption is that the missing extremes are not necessarily consecutive. A graphical method grounded on our theory for estimating the number of missing extremes is given in Sect. 5.2. The basic idea is to artificially remove a number of extremes from the observed data. If enough extremes have been removed, then we will have a consecutive number of the largest observations missing so we can apply our method. The graphical procedure will help us identify the number of extremes to artificially remove from the data in order to perform this second stage estimation.

It is important to note that our work is related to existing literature on modifications of the Hill estimator (including, potentially, removing a part of the upper order statistics from the sample) in order to improve the performance of the estimator; see e.g., the recent optimality result established in this direction by Bhattacharya et al. (2019). The setup and the emphasis in our work are, of course, different. It remains to be investigated whether it is possible to implement the idea of robustifying the Hill estimator in the context of missing extremes.

In another direction, our model formulation shares some commonalities with work (see, e.g., Aban et al. 2006; Beirlant et al. 2016, 2017) in the case the data come from a class of truncated Pareto distributions. The truncation is modeled via an unknown threshold parameter. Maximum likelihood estimators are then derived for the threshold parameter and the tail index. In our case, we do not have a notion of a threshold. A closely related, but slightly different formulation, considers observations that are heavy-tailed but subject to right censoring. In this context, the observed data consists of the minimum of two random variables, the measured variable X and a censoring variable C, which are assumed to be independent. Assuming X has heavy-tails, the goal is to estimate the tail index based on observations $Z_i = \min\{X_i, C_i\}$ and $\delta_i = I_{X_i \le C_i}, i = 1, \dots, n$. See Einmahl et al. (2008); Worms and Worms (2014) and Stupfler (2016) for further details. Unlike our setting, one knows which data values, and hence the number of observations, in the sample that are *incomplete*.

The paper is organized as follows. In Sect. 2, we provide some background on the HEWE process $H_n(\cdot)$. In Sect. 3 we describe an approximate asymptotic maximum likelihood estimation procedure of the parameters γ and δ based on the HEWE process, in the case when the Hill estimator is evaluated at a fixed number of points, and establish its consistency and asymptotic normality. In the case when the observations are drawn the Pareto distribution, we describe, in Sect. 4, another estimation procedure, not directly based on the HEWE process, when the number of points increases to infinity. Section 5 illustrates the methodology via a simulation study. In Sect. 6, we apply our methodology to three data sets: the well-known Danish fire-insurance claim data, the Google+ data, and a NOAA data set consisting of insurance costs associated with climate and natural disasters in the U.S. The proofs of the results in Sects. 3 and 4 are contained in the Appendix.

2 Preliminaries

In this section we formulate a framework for estimating the shape parameter and the number of missing extremes, if any. The setup is somewhat unconventional, so we will explain it in detail. The starting point is the assumption that the observations form a part of a larger sample from which a certain number of largest observations may have been removed. Neither the original sample size nor the number of removed observations (if any) are assumed to be known, but their difference is the observed number of the remaining observations. The absolute number of removed observations may potentially be large, but still a small fraction of the large original sample size. We now describe how we incorporate these unknowns in our setup.

We assume that the original (partially unobserved) sample is a sequence $X_1, \dots, X_n$ of iid random variables with distribution function F satisfying the regular variation condition (1.1). We remind the reader that the overall number n is not known. If $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ denotes the order statistics of the original sample, then a certain number of the largest of them may have been removed. The number of removed observations (if any) is not known (and deciding whether or not observations have been removed and, if so, estimating the number of the removed observations, is of interest to us). We find it convenient to fix the order of magnitude of the potentially removed part of the sample as a sequence of integers $k_n \to \infty$ with $k_n/n \to 0$. This allows us to describe the unknown number of the removed number of the upper order statistics as $\lfloor \delta k_n \rfloor$ for an unknown $\delta \geq 0$, which becomes the object of estimation. Note that we allow $\delta = 0$, which corresponds to the case where no extremes are missing. We view δ as a parameter associated with the number of removed order statistics (in the scale of (k_n)) that we wish to estimate. This leads to a semiparametric model with a two-dimensional parameter (δ, α) to be estimated. We would like to emphasize that the choice of the sequence (k_n) is inevitably arbitrary, but, at least in theory, our setup makes the choice of this sequence less crucial due to the presence of the parameter δ: changing k_n corresponds to a multiplicative change of δ. A second important point is that our chosen notation (k_n) is the same as the common notation for the number of order statistics used in the standard Hill estimator for the exponent of regular variation. However, even though the Hill estimator forms a part of our estimation procedure, the number of order statistics used there is far more involved than simply k_n, as we will see momentarily.

With the notation having been set up as above, the observed sample size becomes $n - \lfloor \delta k_n \rfloor$, and the observations themselves are simply (an unordered version of) the collection $X_{(\lfloor \delta k_n \rfloor + 1)}, X_{(\lfloor \delta k_n \rfloor + 2)}, \dots, X_{(n)}$. This allows us to define, following (Zou et al. 2019), a stochastic process we call "the Hill estimator without extremes process", or the HEWE process, as the functional Hill process based on $\lfloor \theta k_n \rfloor$ of the upper order statistics of the observations. In the notation we have established, the HEWE process is

$$H_n(\theta) = \begin{cases} \frac{1}{\lfloor \theta k_n \rfloor} \sum_{i=1}^{\lfloor \theta k_n \rfloor} \log X_{(\lfloor \delta k_n \rfloor + i)} - \log X_{(\lfloor \delta k_n \rfloor + \lfloor \theta k_n \rfloor + 1)}, & \theta \geq 1/k_n, \\ 0, & \theta < 1/k_n. \end{cases} \tag{2.1}$$

Note that the same sequence (k_n) is used when parameterizing the number of missing extremes and in the definition of the HEWE process. It is clear that k_n is not playing the role of the number of upper order statistics in the Hill estimator. As before, due to the functional nature of the process, the choice of the sequence (k_n) should not matter, at least in theory.

The HEWE process depends, implicitly, on the unknown value of δ (which is the key parameter to be estimated). Occasionally, when we want to emphasize this dependence, we will use the notation $H_n(\theta;\delta)$ instead of $H_n(\theta)$.

In Zou et al. (2019), a strong approximation to $H_n(\cdot)$ was established under a second-order regular variation. This condition, which is given in the Appendix (see (7.1)) quantifies the rate of convergence in (1.1). Pareto distributions with tail index $\alpha > 0$ ($\bar{F}(x) = x^{-\alpha}$ for $x \geq 1$ and zero otherwise) should be viewed as satisfying this condition when the function A in the denominator in the left hand side of (7.1) vanishes. In this case (7.2) is automatically satisfied with $\lambda = 0$.

We now state the key result (Theorem 2.1(b)) in Zou et al. (2019) that is the basis for our procedures to estimate the shape parameter and the number of the missing extremes, i.e. γ and δ.

Theorem 1 *Assume that the second-order condition* (7.1) *holds. Let $k_n \to \infty$ be such that $k_n/n \to 0$ and that* (7.2) *holds for some $\lambda \in \mathbb{R}$. Then*

$$\sqrt{k_n}\Big(H_n(\cdot;\delta) - \gamma g_\delta(\cdot)\Big) - \lambda b_{\delta,\rho}(\cdot) \Rightarrow \gamma G_\delta(\cdot)$$

weakly in $D(0,\infty)$, where

$$g_\delta(\theta) = \begin{cases} 1, & \delta = 0, \\ 1 - (\delta/\theta)\log\big((\theta/\delta) + 1\big), & \delta > 0, \end{cases} \tag{2.2}$$

$$b_{\delta,\rho}(\theta) = \begin{cases} \frac{1}{1-\rho}\frac{1}{\theta^\rho}, & \delta = 0, \\ \frac{1+(\theta/\delta)\rho-(\theta/\delta+1)^\rho}{(\theta/\delta)(1-\rho)\rho}\frac{1}{(\delta+\theta)^\rho}, & \delta > 0, \end{cases} \tag{2.3}$$

and $G_\delta(\cdot)$ is a centered Gaussian process with the following representation. Denoting by W the standard Brownian motion,

$$G_\delta(\theta) = \frac{1}{\theta}\int_\delta^{\delta+\theta}(1-\delta/x)dW(x),\ \theta > 0\,.$$

The process $G_\delta(\cdot)$ has continuous sample paths and a covariance function given by

$$\mathrm{Cov}\big(G_\delta(\theta_1), G_\delta(\theta_2)\big) = \begin{cases} \frac{1}{\theta_1\theta_2}\left[\theta_1 \wedge \theta_2 - 2\delta\log\left(1 + \frac{\theta_1\wedge\theta_2}{\delta}\right) + \frac{\delta(\theta_1\wedge\theta_2)}{\delta+(\theta_1\wedge\theta_2)}\right], & \delta > 0, \\ \frac{1}{\theta_1\vee\theta_2}, & \delta = 0. \end{cases} \tag{2.4}$$

3 An algorithm for estimating γ and δ

Throughout this section we fix points $0 < \theta_1 < \cdots < \theta_m$. According to Theorem 1 the random vector $\mathbf{H}_n = (H_n(\theta_1), \cdots, H_n(\theta_m))'$ has, for large n, an approximately Gaussian likelihood given by

$$\sqrt{\frac{k_n^m}{(2\pi)^m\gamma^{2m}|\Sigma_{m,\delta}|}}\exp\left\{-\frac{k_n}{2\gamma^2}\left(\mathbf{H}_n-\mathbf{g}_\delta\gamma-\frac{\lambda\mathbf{b}_{\delta,\rho}}{\sqrt{k_n}}\right)^T\right.$$
$$\left.\Sigma_{m,\delta}^{-1}\left(\mathbf{H}_n-\mathbf{g}_\delta\gamma-\frac{\lambda\mathbf{b}_{\delta,\rho}}{\sqrt{k_n}}\right)\right\}, \tag{3.1}$$

where $\Sigma_{m,\delta}$ is the covariance matrix of the Gaussian vector $(G_\delta(\theta_1),\cdots,G_\delta(\theta_m))^T$, $\mathbf{g}_\delta=(g_\delta(\theta_1),\dots,g_\delta(\theta_m))'$, and $\mathbf{b}_{\delta,\rho}=(b_{\delta,\rho}(\theta_1),\dots,b_{\delta,\rho}(\theta_m))'$, with $g_\delta(\cdot)$ and $b_{\delta,\rho}$ given in (2.2) and (2.3), respectively. Since we are interested in estimating γ and δ, while ρ and λ are nuisance parameters, we devised a procedure that estimates the parameters of interested while assuming that $\lambda=0$. Note that this eliminates the nuisance parameter ρ as well and leads to a significant simplification in calculations. We will show that the resulting estimators are still consistent and asymptotically normal even if the true value of λ is different from 0. That is, we will maximize the "pseudo-likelihood" function given by

$$\sqrt{\frac{k_n^m}{(2\pi)^M\gamma^{2m}|\Sigma_{m,\delta}|}}\exp\left\{-\frac{k_n}{2\gamma^2}(\mathbf{H}_n-\mathbf{g}_\delta\gamma)^T\Sigma_{m,\delta}^{-1}(\mathbf{H}_n-\mathbf{g}_\delta\gamma)\right\}. \tag{3.2}$$

In order to facilitate analysis of the pseudo-likelihood function, we transform the $\mathbf{H}_n$ and $\mathbf{g}_\delta$. Set $\theta_0=0$, and let

$$T_{ni}=H_n(\theta_i)-(\theta_{i-1}/\theta_i)H_n(\theta_{i-1}),\quad h_{\delta i}=g_{\delta,i}-(\theta_{i-1}/\theta_i)g_{\delta,i-1},\ i=1,\dots,m,$$

with $g_{\delta,i}=g_\delta(\theta_i)$, where $T_{n1}=H_n(\theta_1)$ and $h_{\delta 1}=g_{\delta,1}$. Put $\mathbf{T}_n=(T_{n1},\dots,T_{nm})'$ and $\mathbf{h}_\delta=(h_{\delta 1},\dots,h_{\delta m})'$. After some algebraic manipulations, the -2log(pseudo-likelihood) corresponding to (3.2) can be written in the form

$$\tilde{L}_n(\gamma,\delta):=C+2m\log\gamma-\sum_{i=1}^m\log\omega_{i,\delta}+\frac{k_n}{\gamma^2}\sum_{i=1}^m\omega_{i,\delta}(T_{ni}-\gamma h_{\delta i})^2, \tag{3.3}$$

where

$$\omega_{i,\delta}=\begin{cases}1\Big/\left(1/\theta_i-\theta_{i-1}/\theta_i^2\right) & \delta=0\\ \delta\Big/\left(v(\theta_i/\delta)-(\theta_{i-1}/\theta_i)^2v(\theta_{i-1}/\delta)\right) & \delta>0\end{cases},$$
$$v(x)=\frac{1}{x}-\frac{2\log(1+x)}{x^2}+\frac{1}{x(x+1)}$$

and C is a constant independent of γ and δ; see (11) in Zou et al. (2019).

We now separate the notation for the unknown true parameters γ_0 and δ_0 in the observed sample from the optimization variables which we continue denoting by γ and δ. Let $\mathbf{Y}_n=(Y_{n1},\dots,Y_{nm})=\gamma_0^{-1}\mathbf{T}_n\sqrt{k_n}-\mathbf{h}_{\delta_0}\sqrt{k_n}$. This random vector converges weakly to a Gaussian vector with independent components such that, in the limit,

$$\text{the } i\text{th component is } N\left(\gamma_0^{-1}\lambda b^*_{\delta_0,\rho,i}, 1/\omega_{i,\delta_0}\right), \tag{3.4}$$

where

$$b^*_{\delta_0,\rho,i} = b_{\delta_0,\rho}(\theta_i) - \frac{\theta_{i-1}}{\theta_i} b_{\delta_0,\rho}(\theta_{i-1}),\ i = 1, \ldots, m; \tag{3.5}$$

see Zou et al. (2019). Then the -2log(pseudo-likelihood) becomes

$$\begin{aligned}\tilde{L}_n(\gamma,\delta) =& C + 2m\log\gamma - \sum_{i=1}^m \log\omega_{i,\delta} \\ &+ \frac{1}{\gamma^2}\sum_{i=1}^m \omega_{i,\delta}\left(\gamma_0 Y_{ni} - (\gamma h_{\delta i} - \gamma_0 h_{\delta_0 i})\sqrt{k_n}\right)^2.\end{aligned}$$

Since C is independent of γ and δ, we ignore this term and optimize the function

$$\begin{aligned}L_n(\gamma,\delta) =& 2m\log\gamma - \sum_{i=1}^m \log\omega_{i,\delta} \\ &+ \frac{1}{\gamma^2}\sum_{i=1}^m \omega_{i,\delta}\left(\gamma_0 Y_{ni} - \sqrt{k_n}(\gamma h_{\delta i} - \gamma_0 h_{\delta_0,i})\right)^2.\end{aligned} \tag{3.6}$$

The main result of this section, Theorem 3.1 below, proves the consistency and the asymptotic normality of the estimators obtained by minimizing the function L_n. It applies only in the case when $\delta_0 > 0$, i.e. when some extremes are missing. One can interpret the case when no extremes are missing as corresponding to a small but positive value of δ_0. We will assume that the true value of the parameters, (γ_0, δ_0), belong to the interior Θ^o of a known compact set $\Theta = [m_1, M_1] \times [m_2, M_2] \subset (0, \infty)^2$, and compute our estimator via

$$(\hat{\gamma}, \hat{\delta}) = \arg\min_{(\gamma,\delta)\in\Theta} L_n(\gamma,\delta). \tag{3.7}$$

Note that the observations only enter $L_n(\gamma,\delta)$ through the vector $\mathbf{T}_n$ so that the optimization in (3.7) does not depend explicitly on γ_0 and δ_0. Rather these true parameters only play a role in deriving the limiting properties of $L_n(\gamma,\delta)$.

The following quantities will be used in the statement of our main results. Let

$$b_m = \sum_{i=1}^m \omega_{i,\delta_0} h^2_{\delta_0,i},\quad c_m = \sum_{i=1}^m \omega_{i,\delta_0} (h'_{\delta_0,i})^2,\quad d_m = \sum_{i=1}^m \omega_{i,\delta_0} h_{\delta_0,i} h'_{\delta_0,i}. \tag{3.8}$$

where $h'_{\delta_0,i} = \frac{\mathrm{d}h_{\delta,i}}{\mathrm{d}\delta}|_{\delta=\delta_0}$.

Theorem 3.1 *Assume that the second-order condition (7.1) holds. Let $k_n \to \infty$ be such that $k_n/n \to 0$ and that (7.2) holds for some $\lambda \in \mathbb{R}$. Suppose that $\delta_0 > 0$, that $(\gamma_0, \delta_0) \in \Theta^o$, and let $m \geq 2$ be fixed. Then the optimization problem (3.7) has a* *unique solution* $(\hat\gamma, \hat\delta)$ *with probability increasing to 1 as* $n \to \infty$. *This solution is a weakly consistent estimator of* (γ_0, δ_0), *and*

$$\left(\sqrt{k_n}(\hat\gamma - \gamma_0), \sqrt{k_n}(\hat\delta - \delta_0)\right) \Rightarrow N\left(.5\Gamma_m^{-1}\mathbf{a}, \Gamma_m^{-1}\right) \tag{3.9}$$

as $n \to \infty$, *where in the notation of* (3.8),

$$\Gamma_m = \begin{bmatrix} \frac{b_m}{\gamma_0^2} & \frac{d_m}{\gamma_0} \\ \frac{d_m}{\gamma_0} & c_m \end{bmatrix} \quad \text{and} \quad \Gamma_m^{-1} = \begin{bmatrix} \frac{\gamma_0^2 c_m}{b_m c_m - d_m^2} & -\frac{\gamma_0 d_m}{b_m c_m - d_m^2} \\ -\frac{\gamma_0 d_m}{b_m c_m - d_m^2} & \frac{b_m}{b_m c_m - d_m^2} \end{bmatrix}, \tag{3.10}$$

and $\mathbf{a} = (a_1, a_2)^T$ *with*

$$a_1 = 2\gamma_0^{-2}\lambda \sum_{i=1}^{m} \omega_{i,\delta_0} h_{\delta_0,i} b^*_{\delta_0,\rho,i}, \quad a_2 = 2\gamma_0^{-1}\lambda \sum_{i=1}^{m} \omega_{i,\delta_0} h'_{\delta_0,i} b^*_{\delta_0,\rho,i}. \tag{3.11}$$

A part of the claim of Theorem 3.1 is the non-singularity of the matrix Γ_m, which we will establish below. It is interesting to compare the performance of the estimator $\hat\gamma$ of the shape parameter given in the theorem with the limiting variance of the plain Hill estimator of the shape parameter. Of course, the Hill estimator is only consistent when there are no missing extremes, i.e. when $\delta_0 = 0$, whereas Theorem 3.1 only applies in the case $\delta_0 > 0$. However, the comparison is instructive if one takes the limiting distribution in Theorem 3.1 and considers the situation when $\delta_0 \to 0$. It would not be surprising to expect a loss in efficiency, in the sense that having to estimate both γ and δ may lead to higher bias and/or higher variance of the estimator of the shape parameter from Theorem 3.1 in comparison with the Hill estimator which does not need to estimate δ. However, it turns that there is no loss of efficiency, after all. We have the following proposition.

Proposition 3.1 *Suppose that* $\delta_0 > 0$ *and let* $m \geq 2$ *be fixed. Then the matrix* Γ_m *is invertible and, as* $\delta_0 \downarrow 0$, *the parameters of the Gaussian limit of* $\sqrt{k_n}(\hat\gamma - \gamma_0)$ *satisfy*

$$\begin{aligned} &\text{the mean converges to } \frac{\lambda}{1-\rho}\frac{\theta_m^{1-\rho} - \theta_1^{1-\rho}}{\theta_m - \theta_1}, \\ &\text{the variance converges to } \frac{\gamma_0^2}{\theta_m - \theta_1}. \end{aligned} \tag{3.12}$$

Recall that under the assumptions(7.1) and (7.2) the Hill estimator of γ satisfies

$$\sqrt{k_n}(H_n(k_n) - \gamma_0) \Rightarrow N\left(\lambda/(1-\rho), \gamma_0^2\right); \tag{3.13}$$

see e.g. Theorem 3.2.5 in de Haan and Ferreira (2006).

The significance of (3.12) becomes clear when we let $\theta_m = 1$ and choose θ_0 close to 0, in which case the bias and the variance become the same as those in (3.13).

The proof of Theorem 3.1 (including the invertibility of the matrix Γ_m) will follow from a sequence of lemmas, whose proofs are given in the Appendix. We start with the asymptotic behavior of the gradient of the function L_n evaluated at the true values of the parameters.

Lemma 3.1 *Suppose that $\delta_0 > 0$ and let $m \geq 2$ be fixed. Then*

$$\left(\frac{\partial_1 L_n(\gamma_0, \delta_0)}{\sqrt{k_n}}, \frac{\partial_2 L_n(\gamma_0, \delta_0)}{\sqrt{k_n}}\right) \Rightarrow N\Big(-\mathbf{a}, 4\Gamma_m\Big).$$

Next, we address the asymptotic behavior of the Hessian matrix

$$\mathcal{H}_n(\gamma, \delta) = \begin{bmatrix} \partial_1^2 L_n(\gamma, \delta) & \partial_1\partial_2 L_n(\gamma, \delta) \\ \partial_1\partial_2 L_n(\gamma, \delta) & \partial_2^2 L_n(\gamma, \delta) \end{bmatrix} \tag{3.14}$$

of the function L_n evaluated at any weakly consistent estimator of the true values.

Lemma 3.2 *Suppose that $\delta_0 > 0$ and let $m \geq 2$ be fixed. If $(\tilde{\gamma}, \tilde{\delta}) \xrightarrow{P} (\gamma_0, \delta_0)$ then $k_n^{-1}\mathcal{H}_n(\tilde{\gamma}, \tilde{\delta}) \xrightarrow{P} 2\Gamma_m$.*

The next lemma proves the weak consistency of our estimator.

Lemma 3.3 *Suppose that $\delta_0 > 0$ and let $m \geq 2$ be fixed. Then the optimization problem (3.7) has a unique solution $(\hat{\gamma}, \hat{\delta})$ with probability increasing to 1 as $n \to \infty$ and*

$$(\hat{\gamma}, \hat{\delta}) \xrightarrow{P} (\gamma_0, \delta_0).$$

Proof of Theorem 3.1 The Taylor expansion of the gradient of L_n around the true values of the parameter tells us that

$$\left(\frac{\partial_1 L_n(\hat{\gamma}, \hat{\delta})}{\sqrt{k_n}}, \frac{\partial_2 L_n(\hat{\gamma}, \hat{\delta})}{\sqrt{k_n}}\right)^T = \left(\frac{\partial_1 L_n(\gamma_0, \delta_0)}{\sqrt{k_n}}, \frac{\partial_2 L_n(\gamma_0, \delta_0)}{\sqrt{k_n}}\right)^T + \frac{\mathcal{H}_n(\bar{\gamma}, \bar{\delta})}{\sqrt{k_n}}(\hat{\gamma} - \gamma_0, \hat{\delta} - \delta_0)^T$$

for some $(\bar{\gamma}, \bar{\delta})$ between $(\hat{\gamma}, \hat{\delta})$ and (γ_0, δ_0). By Lemma 3.3, with probability increasing to 1, the infimum in (3.7) is achieved in the interior of the set Θ, and on that event $(\partial_1 L_n(\hat{\gamma}, \hat{\delta}), \partial_2 L_n(\hat{\gamma}, \hat{\delta}))^T = (0, 0)^T$. Therefore, the relation

$$-\left(\frac{\partial_1 L_n(\gamma_0, \delta_0)}{\sqrt{k_n}}, \frac{\partial_2 L_n(\gamma_0, \delta_0)}{\sqrt{k_n}}\right)^T = \frac{\mathcal{H}_n(\bar{\gamma}, \bar{\delta})}{\sqrt{k_n}}(\hat{\gamma} - \gamma_0, \hat{\delta} - \delta_0)^T$$

also holds on an event whose probability increases to 1. Lemma 3.3, also implies that $(\bar{\gamma}, \bar{\delta}) \xrightarrow{P} (\gamma_0, \delta_0)$, so the claim of the theorem follows from Lemmas 3.1 and 3.2. □

4 Estimating γ and δ in a Pareto sample

In this section we assume that the observations $X_1, X_2, \ldots$ follow the Pareto distribution $\bar{F}(x) = x^{-\alpha}$ for $x \geq 1$. In this case the exact distribution of the order statistics is available, so we will not need to rely as much on the asymptotic normality of H_n in Theorem 1. Unlike the algorithm of the previous section, we will now use k_n equally spaced points $\theta_{i,n} = \varepsilon + i/k_n, 1 \leq i \leq k_n$, for some $\varepsilon > 0$.

Let, once again, $X_{(1)} > X_{(2)} > \cdots > X_{(n)}$ be the order statistics from the sample $X_1, \ldots, X_n$. Then

$$(X_{(1)}, \ldots, X_{(n)}) \stackrel{d}{=} \left(\left(\frac{S_{n+1}}{S_1} \right)^{\gamma}, \ldots, \left(\frac{S_{n+1}}{S_n} \right)^{\gamma} \right),$$

see e.g. Corollary 1.6.9 of Reiss (1989). Here $S_i = \sum_{j=1}^{i} E_j$, with (E_j) iid standard exponential random variables. It follows that the HEWE process satisfies

$$\left(H_n(\theta_{i,n}), \, i = 1, \ldots, k_n \right) \stackrel{d}{=} \left(\frac{\gamma}{\lfloor \theta_{i,n} k_n \rfloor} \sum_{i=1}^{\lfloor \theta_{i,n} k_n \rfloor} \log \left(\frac{S_{\lfloor \delta k_n \rfloor + \lfloor \theta_{i,n} k_n \rfloor + 1}}{S_{\lfloor \delta k_n \rfloor + i}} \right), \, i = 1, \ldots, k_n \right).$$

Since $E_i^* =: i \log(S_{i+1}/S_i)$, $i = 1, \ldots, n$, are also iid standard exponential random variables, a bit of algebra shows that

$$\left(H_n(\theta_{i,n}), \, i = 1, \ldots, k_n \right) \stackrel{d}{=} \left(\frac{\gamma}{\lfloor \theta_{i,n} k_n \rfloor} \sum_{i=\lfloor \delta k_n \rfloor + 1}^{\lfloor \delta k_n \rfloor + \lfloor \theta_{i,n} k_n \rfloor} \left(1 - \frac{\lfloor \delta k_n \rfloor}{i} \right) E_i^*, \, i = 1, \ldots, k_n \right).$$

Denote now $\xi_{i,n} = H_n(\theta_{i,n}) - \lfloor \theta_{i-1,n} k_n \rfloor H_n(\theta_{i-1,n}) / \lfloor \theta_{i,n} k_n \rfloor$, $i = 1, \ldots, k_n$ and set $\theta_{0,n} = 0$. Since for any $i \geq 2$ we have $\lfloor \theta_{i,n} k_n \rfloor - \lfloor \theta_{i-1,n} k_n \rfloor = 1$, it follows that

$$\left(\xi_{i,n}, \, i = 2, \ldots, k_n \right) \stackrel{d}{=} \left(\frac{\gamma E^*_{\lfloor \delta k_n \rfloor + \lfloor \theta_{i,n} k_n \rfloor}}{\lfloor \delta k_n \rfloor + \lfloor \theta_{i,n} k_n \rfloor}, \, i = 2, \ldots, k_n \right). \tag{4.1}$$

We wish to perform an MLE procedure based on the joint density of $(\xi_{1,n}, \ldots, \xi_{k_n,n})$. The joint density of $(\xi_{2,n}, \ldots, \xi_{k_n,n})$ can be read off from (4.1). Since these random variables are independent of $\xi_{1,n}$, we will construct a mixed likelihood function by combining the exact joint density of $(\xi_{2,n}, \ldots, \xi_{k_n,n})$ with the asymptotic normal density for $\xi_{1,n}$ given in (3.2) with $m = 1$ and $\theta_1 = \theta_{1,n}$. We simplify the resulting expression by dropping the rounding (replacing $\lfloor \delta k_n \rfloor$ by δk_n and $\lfloor \theta_{i,n} k_n \rfloor$ by $\theta_{i,n} k_n$ as needed). Taking the logarithm of the result multiplied by -2 gives us the following function of γ and δ (a pseudo-log-likelihood function, modulo additive constants not depending on γ or δ), to be optimized:

$$
\begin{aligned}
L_n(\gamma,\delta) = 2\log\gamma - \log\omega_{1,\delta} - 2\sum_{i=2}^{k_n}\log\left(\frac{\delta+\theta_{i,n}}{\gamma}\right) \\
+ \frac{k_n\omega_{1,\delta}}{\gamma^2}(\xi_{1,n}-\gamma g_{\delta,1})^2 + \frac{2k_n}{\gamma}\sum_{i=2}^{k_n}(\delta+\theta_{i,n})\xi_{i,n}\,.
\end{aligned}
\tag{4.2}
$$

As in the previous section, we separate the notation for the unknown true parameters γ_0 and δ_0 from the optimization variables γ and δ. Denote $\eta_n = \gamma_0^{-1}\xi_{1,n}\sqrt{k_n} - g_{\delta_0,1}\sqrt{k_n}$ and $Z_{i,n} =_d k_n(\delta_0+\theta_{i,n})\xi_{i,n}/\gamma_0$, and notice that these independent random variables satisfy $\eta_n \Rightarrow N(0, 1/\omega_{1,\delta_0})$ and each $Z_{i,n}$ converges to a standard exponential random variable. With this notation the function to be minimized becomes

$$
\begin{aligned}
L_n(\gamma,\delta) =& 2k_n\log\gamma - \log\omega_{1,\delta} - 2\sum_{i=2}^{k_n}\log(\delta+\theta_{i,n}) \\
&+ \frac{\omega_{1,\delta}}{\gamma^2}\left(\gamma_0\eta_n - \sqrt{k_n}(\gamma g_{\delta,1} - \gamma_0 g_{\delta_0,1})\right)^2 \\
&+ \frac{2\gamma_0}{\gamma}\sum_{i=2}^{k_n}\frac{(\delta+\theta_{i,n})Z_{i,n}}{\delta_0+\theta_{i,n}}\,.
\end{aligned}
\tag{4.3}
$$

As in the previous section we will assume that the true value of the parameters, (γ_0,δ_0), belong to the interior Θ^o of a known compact set $\Theta = [m_1, M_1]\times[m_2, M_2] \subset (0,\infty)^2$, and compute our estimator via (3.7), this time using L_n in (4.3).

The following is the main result of this section.

Theorem 4.1 *Suppose that the sample $X_1, X_2, \ldots$ is drawn from the Pareto distribution. Let $k_n \to \infty$ be such that $k_n/n \to 0$. Suppose that $\delta_0 > 0$, that $(\gamma_0,\delta_0) \in \Theta^o$, and let $\theta_{i,n} = \varepsilon + i/k_n$, $1 \le i \le k_n$ for some $\varepsilon > 0$. Then the optimization problem (3.7) has a unique solution $(\hat\gamma,\hat\delta)$ with probability increasing to 1 as $n \to \infty$. This solution is a weakly consistent estimator of (γ_0,δ_0), and*

$$
\left(\sqrt{k_n}(\hat\gamma-\gamma_0), \sqrt{k_n}(\hat\delta-\delta_0)\right) \Rightarrow N\left(0, \Gamma_\infty^{-1}\right)
$$

where

$$
\Gamma_\infty = \Gamma_1 + \begin{bmatrix} \gamma_0^{-2} & -\gamma_0^{-1}\log\left(1+1/(\delta_0+\varepsilon)\right) \\ -\gamma_0^{-1}\log\left(1+1/(\delta_0+\varepsilon)\right) & (\delta_0+\varepsilon)^{-1}(\delta_0+\varepsilon+1)^{-1} \end{bmatrix}
\tag{4.4}
$$

and Γ_1 is as in (3.10) with $m = 1$ and $\theta_1 = \varepsilon$. The matrix Γ_∞ is invertible with

$$\Gamma_\infty^{-1} = \Delta^{-1}\begin{bmatrix} \gamma_0^2\big(c_1 + (\delta_0+\varepsilon)^{-1}(\delta_0+\varepsilon+1)^{-1}\big) & \gamma_0\Big(-d_1 + \log\big(1+1/(\delta_0+\varepsilon)\big)\Big) \\ \gamma_0\Big(-d_1 \log\big(1+1/(\delta_0+\varepsilon)\big)\Big) & b_1+1 \end{bmatrix},$$

where $\Delta = (b_1+1)\big(c_1 + (\delta_0+\varepsilon)^{-1}(\delta_0+\varepsilon+1)^{-1}\big) - \big(d_1 - \log\big(1+1/(\delta_0+\varepsilon)\big)\big)^2.$

The structure of proof of Theorem 4.1 is nearly identical to that of Theorem 3.1, with Lemmas 3.1-3.3 replaced by their counterparts, Lemmas 4.1–4.3. Once again, we start with the asymptotic behavior of the gradient of the function L_n evaluated at the true values of the parameters.

Lemma 4.1 *Suppose* $\delta_0 > 0$ *and* $\theta_{i,n} = \varepsilon + i/k_n, 1 \le i \le k_n$. *Then*

$$\left(\frac{\partial_1 L_n(\gamma_0,\delta_0)}{\sqrt{k_n}}, \frac{\partial_2 L_n(\gamma_0,\delta_0)}{\sqrt{k_n}}\right) \Rightarrow N\Big(0, 4\Gamma_\infty\Big).$$

As before, we proceed with the asymptotic behavior of the Hessian matrix (3.14) of the function L_n evaluated at a weakly consistent estimator of the true values.

Lemma 4.2 *Suppose that* $\delta_0 > 0$ *and let* $\theta_{i,n} = \varepsilon + i/k_n$, $i = 1,\dots,k_n$. *If* $(\tilde\gamma,\tilde\delta)\xrightarrow{P}(\gamma_0,\delta_0)$ *then* $k_n^{-1}\mathcal{H}_n(\tilde\gamma,\tilde\delta)\xrightarrow{P}2\Gamma_\infty$.

The final lemma, once again, proves the weak consistency.

Lemma 4.3 *Suppose that* $\delta_0 > 0$ *and let* $\theta_{i,n} = \varepsilon + i/k_n$, $i = 1,\dots,k_n$. *Then the optimization problem* (3.7) *has a unique solution* $(\hat\gamma,\hat\delta)$ *with probability increasing to 1 as* $n \to \infty$ *and*

$$(\hat\gamma,\hat\delta)\xrightarrow{P}(\gamma_0,\delta_0).$$

Proof of Theorem 4.1 One can use an argument identical to that in the proof of Theorem 3.1, hence only the invertibility of Γ_∞ needs to be shown. Since Γ_1 is nonnegative definite, we only have to check that the second matrix in Γ_∞ has a positive determinant. However, by Jensen's inequality,

$$(\delta_0+\varepsilon)^{-1}(\delta_0+\varepsilon+1)^{-1} - \left(\log\left(1+\frac{1}{\delta_0+\varepsilon}\right)\right)^2$$
$$= \int_\varepsilon^{1+\varepsilon} \frac{1}{(\delta_0+x)^2}dx - \left(\int_\varepsilon^{1+\varepsilon} \frac{1}{\delta_0+x}dx\right)^2 > 0,$$

as required. □

Remark 4.1 It is elementary to check that the entry in the upper left corner of the matrix Γ_∞^{-1} converges, as $\delta_0 \to 0$, to γ_0^2. This is the same somewhat surprising lack of efficiency lost we have seen in Proposition 3.1.

Table 1 Pareto distribution, $n = 5000, k_n = 200$

δ_0	$\hat{\delta}_a$ Mean (sd)	$\hat{\gamma}_a$ Mean (sd)	$\rho_{\hat{\delta}_a,\hat{\gamma}_a}$ Corr (asy)	$\hat{\delta}_b$ Mean (sd)	$\hat{\gamma}_b$ Mean (sd)	$\rho_{\hat{\delta}_b,\hat{\gamma}_b}$ Corr (asy)
0.1	0.113 (0.057)	1.015 (0.143)	0.858 (0.829)	0.104 (0.049)	1.006 (0.129)	0.841 (0.796)
0.2	0.222 (0.104)	1.025 (0.187)	0.915 (0.894)	0.207 (0.096)	1.010 (0.177)	0.915 (0.878)
0.5	0.547 (0.285)	1.040 (0.309)	0.965 (0.956)	0.515 (0.254)	1.014 (0.282)	0.962 (0.951)

Table 2 Fréchet distribution, $n = 5000, k_n = 200$

δ_0	$\hat{\delta}_a$ Mean (sd)	$\hat{\gamma}_a$ Mean (sd)	$\rho_{\hat{\delta}_a,\hat{\gamma}_a}$ Corr (asy)	$\hat{\delta}_b$ Mean (sd)	$\hat{\gamma}_b$ Mean (sd)	$\rho_{\hat{\delta}_b,\hat{\gamma}_b}$ Corr (asy)
0.1	0.106 (0.050)	0.992 (0.130)	0.829 (0.829)	0.101 (0.045)	0.988 (0.122)	0.826 (0.796)
0.2	0.208 (0.094)	0.993 (0.176)	0.906 (0.894)	0.196 (0.085)	0.981 (0.165)	0.904 (0.878)
0.5	0.535 (0.287)	1.011 (0.300)	0.961 (0.956)	0.502 (0.252)	0.985 (0.274)	0.961 (0.951)

5 Simulation results

5.1 Estimation of δ and γ

In this section, we compare the performance of the estimation procedures described in Sects. 3 and 4 on simulated data. As a test data set, we generate $n = 5000$ observations from a Pareto distribution ($F(x) = 1 - 1/x$, $x \geq 1$) and from a standard Fréchet distribution ($F(x) = \exp\{-x^{-1}\}, x \geq 0$). For each distribution, we considered three cases corresponding to removing the largest 20, 40, and 100 extremes, respectively, from the data. We chose $k_n = 200$ in all cases so that the resulting parameter of interest is then $\delta = 0.1, 0.2, 0.5$ corresponding to the three scenarios. In theory, the choice of k_n does not matter. In practice we have tried a variety of difference choices of k_n and they generally worked well in terms of estimating the number of missing extremes except for the situation when k_n is not sufficiently larger than the number of missing extremes. Fine tuning the choice of k_n in practice is the subject of future research. For the estimation method of Sect. 3, we chose $m = 10$ distinct θ's with $\theta_i = i/10, i = 1, \ldots, 10$, and minimized the pseudo-likelihood function given in (3.3) with respect to δ and γ. This was repeated 1000 times and the summary statistics (means and standard deviations) are given in Tables 1 (Pareto) and 2 (Fréchet) corresponding to the columns labeled $\hat{\delta}_a$ and $\hat{\gamma}_a$. Notice that both the bias and standard deviation of $\hat{\delta}_a$ increase with δ_0 where the latter increases at a rate that is roughly proportionally to δ_0.

We also used the estimation procedure of Sect. 4 ($m = k_n, \epsilon = 1/200$) in which the objective function in (4.2) was minimized. This procedure was applied to the same Pareto and Fréchet generated data as before, even though, in theory, the method was introduced only for Pareto samples. The results are also summarized in Tables 1 and 2 using the labels $\hat{\delta}_b$ and $\hat{\gamma}_b$. The bias for $\hat{\delta}_b$ is considerably smaller than that for $\hat{\delta}_a$ in most cases (even in the Fréchet samples). The standard deviations were also a bit smaller in all cases. On the other hand the biases for $\hat{\gamma}_b$ were similar to those for $\hat{\gamma}_a$, but in all cases the standard deviation was a touch smaller. This may not be too surprising since these estimates are based on more θ_i. The asymptotic standard deviations for the two estimates using $m = 10$ and $m = k_n$ can be computed using the formulae in (3.10) and (4.4), respectively, and are all smaller than their finite sample counterparts. For example, in the Pareto case for $\hat{\delta}_a$ the asymptotic standard errors for $\delta_0 = 0.1, 0.2, 0.5$ were 0.047, 0.083, 0.219, respectively.

It is worth emphasizing again that the two estimation procedures based on $m = 10$ and $m = k_n$ generally performed well for the Fréchet case even though our theory for $m = k_n$ is not directly applicable to this case. Interestingly, the biases and standard deviations were generally smaller in the Fréchet case in comparison with the Pareto case, across the range of parameter values and the two estimation procedures. The histograms of the estimates leading to $\hat{\delta}_a$ in Table 2 corresponding to $\delta_0 = 0.1$ and 0.5 are displayed in Fig. 1. The asymptotic normal density function (in red) is overlayed on the histograms. These are a very good approximation in the $\delta_0 = 0.1$ case, while there is a slight bias in the $\delta = .05$ case. Notice the long right tails of the histograms, which are absent in the normal densities. This is more pronounced in the $\delta_0 = 0.5$ case, which is due in part to having fewer of the most extreme observations to estimate tail parameters.

Tables 1 and 2 also contain two additional columns showing correlations (both sample and asymptotic based on (3.10) and (4.4)) between $\hat{\delta}_a$ and $\hat{\gamma}_a$ and between $\hat{\delta}_b$ and $\hat{\gamma}_b$. There is good agreement between the sample and asymptotic correlations and all are large (close to 1). That means that in the two-dimensional optimization likelihood procedure moderate errors in estimating γ_0 lead to significant errors in estimating δ_0 because of the large standard errors of the estimates of δ_0. It turns out, however, that fixing γ, the one-dimensional likelihood optimization procedures for δ has less variability and is, moreover, fairly robust to a mild misspecification of γ. We exploit this fact in the sequel.

We repeated the same simulation analysis for two other distributions: the folded Cauchy ($\gamma = 1$) and the standard Lévy distribution ($\gamma = 2$). The results were quite similar to those for the Pareto and Fréchet distributions.

5.2 Graphical methods for estimating the number of missing extremes

Although our estimation procedure assumes that $\delta_0 k_n$ of the largest extremes are missing, in practice, it might be more likely that missing extremes, if any exist, do not occur consecutively from the largest. Our method can still be used to estimate the total number of missing extremes. For example if there are 10 missing extremes scattered among the largest 50, we can artificially remove the largest 40 extremes from the data set. The altered data set can then be viewed as having 50 consecutive missing extremes and our estimation procedure for estimating the number of missing extremes is applicable. If the resulting estimate is near 50, as it should be, then since we know that 40 have been artificially removed, we would be able to recover

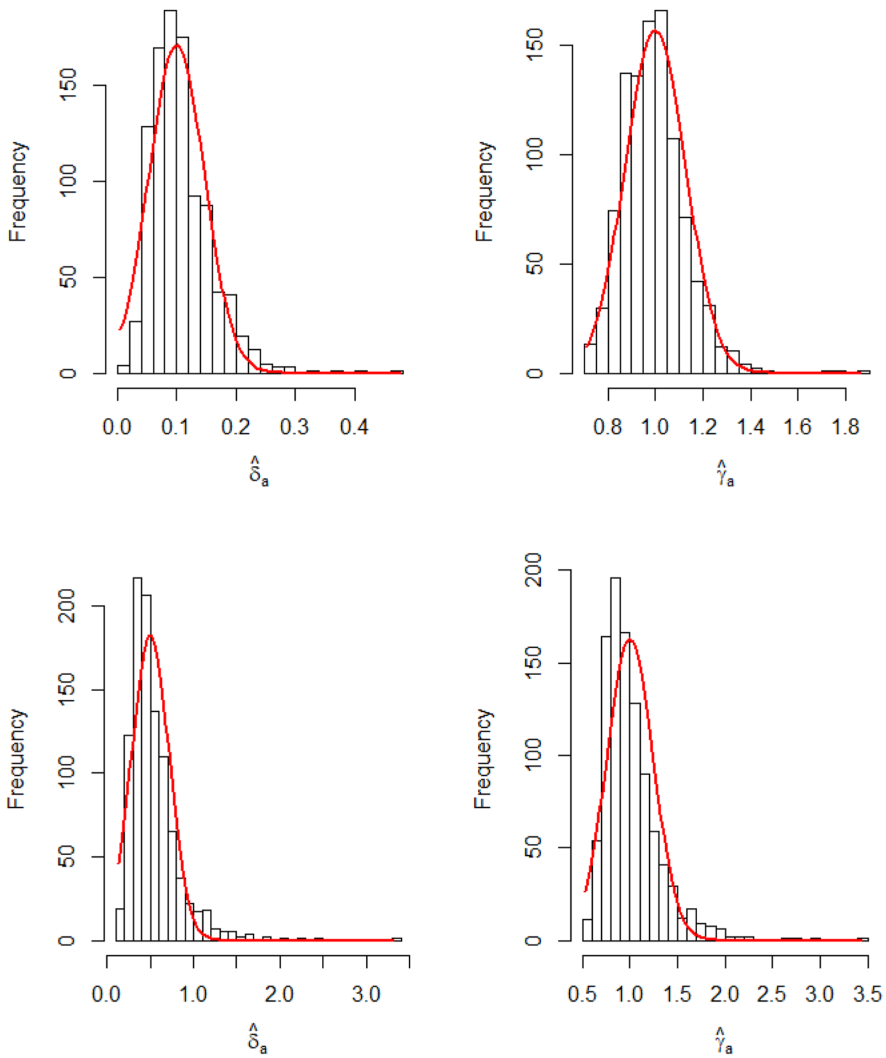

Fig. 1 Histogram for estimates in Table 2 for the Fréchet simulation and its related asymptotic normal density (in red) corresponding to $\delta_0 = 0.1$ (top) and $\delta_0 = 0.5$ (bottom), colour figure online

an estimate of the number of original missing extremes, even when non-consecutive. We use a graphical procedure to give an idea of how this works.

For a given data set with $\delta_0 k_n$ missing observations among the largest $(\delta_0 + \delta^\dagger)k_n$, we remove for each $i = 1, 2, \dots, k_n$, the i largest observations (this corresponds to $\delta = i/k_n$) from the observed data and produce estimates $\hat{\delta}$. Now if the missing extremes are all consecutive, i.e., $\delta^\dagger = 0$, then the plot of $\hat{\delta}$ vs. δ should look approximately linear for $\delta > 0$. On the other hand, if the missing extremes are not consecutive, then by removing $\delta^\dagger k_n$ extremes from the data, we obtain a data set with

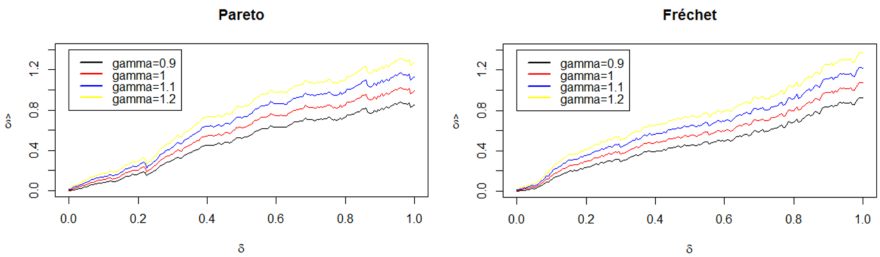

Fig. 2 Estimated number of missing extremes for samples from Pareto and Fréchet distributions with $n = 5000$, $k_n = 200$, $\gamma_0 = 1$, colour figure online

$(\delta_0 + \delta^\dagger)k_n$ consecutive missing extremes. So the plot of $\hat{\delta}$ vs. δ should look approximately linear with slope 1 for $\delta > \delta^\dagger$. In particular for $\delta < \delta^\dagger$ the plot may look very nonlinear depending on the configuration of the $\delta_0 k_n$ missing extremes. Unfortunately, since $\delta^\dagger$ is unknown, it needs to be estimated. The shape of the plots of $\hat{\delta}$ vs. δ provides a clue on how to estimate $\delta^\dagger$. Namely, we pick off the threshold for which the plot of $\hat{\delta}$ vs. δ becomes approximately linear for δ larger than that threshold and the slope of the linear piece should be near 1. This value is then identified as $\delta^\dagger$ and the difference $\hat{\delta} - \delta^\dagger$ is then an estimate of δ_0. Once we have estimated $\delta^\dagger$, if desired, we can remove the largest $\delta^\dagger k_n$ observations and re-apply our estimation procedure to provide updated estimates of δ_0 and γ_0. This approach involves a reasonable amount of user judgment. Automation of this approach is the subject of future research.

We illustrate this procedure with several simulation examples under three different scenarios: (i) no missing extremes, (ii) the upper $\delta_0 k_n$ are missing and (iii) the $\delta_0 k_n$ missing extremes are not consecutive upper extremes. The setup for this simulation is similar to that of Sect. 5.1. Samples of size $n = 5000$ are generated from both Pareto and Fréchet distributions with index $\alpha = 1.0$. In all cases, we take $k_n = 200$ and begin by maximizing the bivariate likelihood to obtain initial estimates of δ_0 and, more importantly, γ_0. The method of Sect. 3 is used throughout.

(i) no missing extremes. In this case, no extremes have been removed from the simulated data so that $\delta_0 = 0$. The estimates of δ_0 and γ_0 using the method of Sect. 3 with $m = 10$ are near 0 and 0.912 in the Pareto case, and 0.001 and 0.904 in the Fréchet case. In order to test our estimation procedure, for each $\delta_i = i/k_n$, $i = 1, \dots, k_n$, we remove the upper i extremes of the simulated data set and then compute $\hat{\delta}_i$, by minimizing the objective function (3.6) for fixed values of $\gamma = .9, 1.0, 1.1, 1.2$ using the altered data, i.e., with the appropriate number of extremes removed. (As mentioned above, we avoid here maximizing the bivariate likelihood and, at the same time, check the robustness of the univariate likelihood maximization to a mild misspecification of γ.) In Fig. 2 we plot $\hat{\delta}$ vs. δ for each of the the four choices of γ for the Pareto sample (left panel) and the Fréchet sample (right panel). Notice that each of the four curves are approximately linear with intercept 0, strongly suggesting that $\delta_0 = 0$. The red line (corresponding to the true $\gamma = 1$) has the slope closest to 1.

(ii) the upper $\delta_0 k_n$ extremes are missing. For this simulation, we take $\delta_0 = .25$ so that 50 largest observations are removed from the samples described in (i).

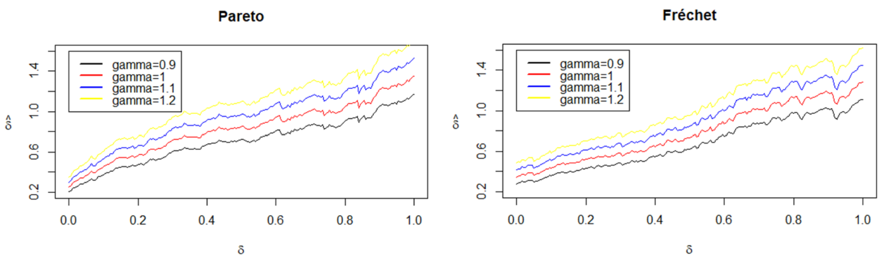

Fig. 3 Estimated number of missing extremes for samples from Pareto and Fréchet distributions with the 50 largest observations removed. $n = 5000$, $k_n = 200$, $\gamma_0 = 1$

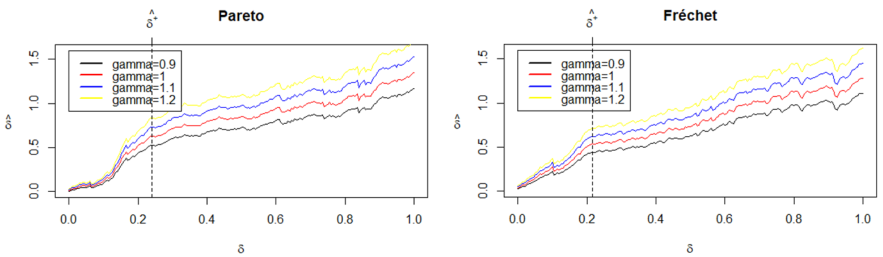

Fig. 4 Estimated number of missing extremes and tail index for samples from Pareto and Fréchet distributions, with 50 observations among the largest 100, removed. $n = 5000$, $k_n = 200$, $\gamma_0 = 1$, colour figure online

The estimates of δ_0 and γ_0 are 0.145 and 0.769 for the Pareto case and 0.431 and 1.122 for the Fréchet case. The same style plots as those in Fig. 2 are displayed in Fig. 3. The reader should keep in mind that now the horizontal axis for δ corresponds to δk_n extreme observations missing from the *observed data* (in addition to the 50 largest removed from the originally generated data). The plots of $\hat{\delta}$ vs δ are again nearly linear for the four values of γ, with the red line ($\gamma = 1$) having slope closest to 1, and the corresponding intercept value close to the true $\delta_0 = .25$.

(iii) the $\delta_0 k_n$ missing extremes are not consecutive. For this simulation, we again take $\delta_0 = .25$, but this time the 50 missing extremes are randomly selected from among the 100 largest observations. The estimate of δ_0 is near 0 and the estimate of γ_0 is 0.774 in the Pareto case, while the estimates of δ_0 and γ_0 are 0.004 and 0.789 in the Fréchet case. For this scenario $\delta^\dagger = 50/200 = .25$ so that after removing another 50 extremes from the *observed* data, we have all 100 of the top extremes removed. The corresponding plots displayed in Fig. 4 now have a different look. They are essentially connected segments with nodes around δ just over .1 and just over .2 for both the Pareto and Fréchet cases, respectively. Notice that locations of these nodes are robust to the choice of γ; each of the 4 curves have nodes at approximately the same horizontal location. On the Pareto plot we would estimate $\delta^\dagger$ to be around .24, which is near the *true* $\delta^\dagger$ of .25. The third segment of the red curve (corresponding to the true $\gamma = 1$), from .24 to 1, has the slope closest to 1. For this curve the value at .24 is .63. This gives us an estimate of δ_0 as $.63 - .24 = .39$. A similar analysis for the Fréchet case gives us an estimate of $\delta^\dagger = .22$ with corresponding value on the red curve of .54. The estimate of δ_0 is

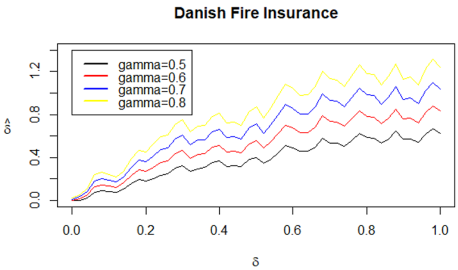

Fig. 5 Estimated number of missing extremes for Danish fire insurance. $n = 2492$, $k_n = 50$

then $.54 - .22 = .32$. So in both cases, we retrieve reasonable (though not perfect) estimates of $\delta_0 = .25$.

6 Applications

In this section, we apply the methodology described in Sects. 5.1 and 5.2 to several real data sets. The goal here is to estimate the shape parameter γ_0 in addition to δ_0 and $\delta^\dagger$, where $\delta_0 k_n$ is the number of missing extremes among the largest $(\delta^\dagger + \delta_0)k_n$ extremes. Of course if our estimate is $\hat{\delta}^\dagger = 0$, we declare that missing extremes, if any, are consecutive. Once again, the estimation method of Sect. 3 is used throughout.

6.1 Danish fire insurance

The Danish Fire Insurance data set is a standard example used in extreme value theory. It is a part of the R-statistics package and consists of 2492 large Danish fire insurance claims from January 1, 1980 to December 31, 1990. Using $k_n = 50$ and $m = 10$, the estimate of δ_0 is near 0 and the estimate of γ_0 is 0.565. Next we explore the possibility of some missing (not necessarily consecutive) extremes by applying the methodology in Sect. 5.2 with 4 values of γ based on the initial estimate. The resulting plots of $\hat{\delta}$ vs δ for four different values of γ are displayed in Fig. 5. All of these plots look roughly linear without any obvious nodes, with the intercepts close to 0. Hence we estimate $\hat{\delta}^\dagger = 0 = \hat{\delta}_0$. The blue curve appears to have the slope closest to 1, so we estimate $\hat{\gamma}_0$ to be around .7. This corresponds to an estimate of α of $1/.7 = 1.43$. This is in the range of other estimates in the literature for α for this data set (see Resnick 1997, 2007).

6.2 Google+

The second example consists of the in-degrees values from 76,438,791 nodes in a snapshot of the Google+ social network that was explored in Zou et al. (2019). They concluded that around 150 consecutive largest extremes were missing. Using $k_n = 500$, the methodology of Sect. 3 gives initial estimates of δ_0 and γ_0 as .327 and 1.418, respectively, so that the number of missing extremes would be .327*500=163. This

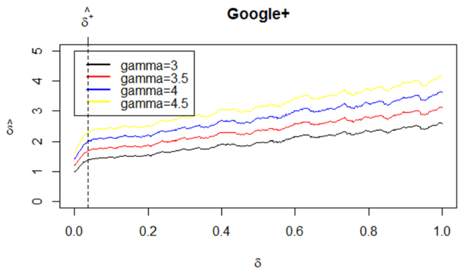

Fig. 6 Estimated number of missing extremes for Google+ with $k_n = 500$, with $\hat{\delta}^\dagger = 0.036$ marked, colour figure online

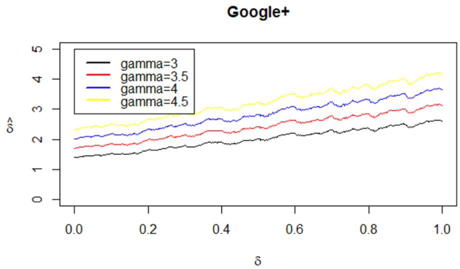

Fig. 7 Estimated number of missing extremes for Google+ with the 18 largest observations removed, $k_n = 500$, colour figure online

estimate is consistent with the number of missing found in Zou et al. (2019). Unfortunately, our estimate of γ_0 does not produce reasonable plots as described in Sect. 5.2, so upon further experimentation we settled on a different range of γ and construct plots of $\hat{\delta}$ by removing the largest observations for four values of $\gamma = 3, 3.5, 4, 4.5$. The resulting plots are displayed in Fig. 6. Using the methodology of Sect. 5.2, we would estimate $\hat{\delta}^\dagger$ to be near 0.036, and the red curve appears to have slope closest to 1. We therefore estimate γ_0 to be close to 3.5. The value of the red curve corresponding to $\hat{\delta}^\dagger$ is 1.38, so we estimate δ_0 as $1.38 - .036 = 1.344$. That corresponds to $1.344 * 500 = 672$ missing extremes among 1.38*500= 690 largest extremes. We now remove the additional $.036 * 500 = 18$ largest values in the data set and re-plot in Fig. 7, the curves corresponding to the 4 values of γ above. Note that all curves are roughly linear and that the red curve ($\gamma = 3.5$) has the slope closest to 1.

6.3 Natural and climate disasters in the U.S. from 1980–2019

This data, which can be accessed from http://ncdc.noaa.gov/billions/, was assembled by the National Oceanic and Atmospheric Administration (NOAA). They identify 258 costly natural and climatic events such as wild fires, hurricanes, flooding, earthquakes, droughts, tornadoes, and severe storms during the period from January 1980 through December 2019. This data set represents the financial costs in billions of 2019 US dollars associated with these events. More details about the data set can be found in Smith and Katz (2013). While one should always exercise caution in applying extreme value theory to small data sets, we nevertheless apply our methods

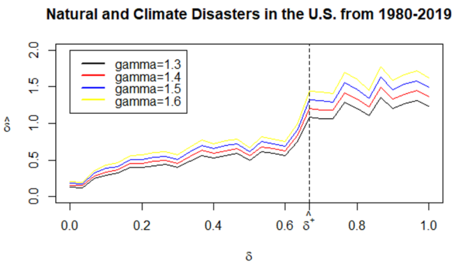

Fig. 8 Weather and climate disasters from 1980 to 2019. $n = 258$, $k_n = 30$, colour figure online

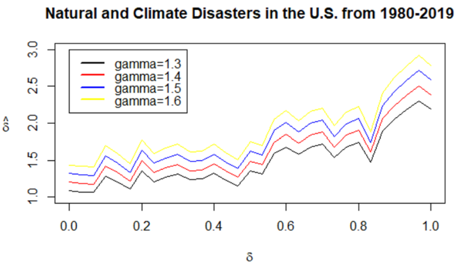

Fig. 9 Weather and climate disasters from 1980 to 2019 with 20 largest observations removed. $n = 258$, $k_n = 30$, colour figure online

in this case with $k_n = 30$. The initial estimates of δ_0 and γ_0 are .08 and 1.349, respectively. This would lead to an estimate of $0.08 * 30 = 2.4$ missing extremes. Since there are no truly *missing extremes* (every disaster event has a recorded value), we interpret missingness as being reflective of some extremes being underreported. As done previously, we explore the possibility that there are non-consecutive missing extremes among of a fraction of the largest observations. To this end, we construct plots of $\hat{\delta}$ by removing largest observations for four values of $\gamma = 1.3, 1.4, 1.5, 1.6$. The resulting plots are in Fig. 8. As in Sect. 5.2, the estimate $\hat{\delta}^{\dagger}$ is near 0.667, and the red curve has the slope closest to 1 in the last part of the plot. We therefore estimate γ_0 to be near 1.4. The value of the red curve at $\hat{\delta}^{\dagger}$ is 1.193, so we estimate δ_0 as $1.193 - 0.667 = 0.526$, corresponding to 16 missing extremes. Now re-estimating δ_0 and γ_0 for the observed data with the additional $\hat{\delta}^{\dagger} k_n = 0.667 * 30 = 20$ extreme observations removed, we obtain $\hat{\delta}_0 = 1.193$ and $\hat{\gamma}_0 = 1.718$. So our final estimate of δ_0 would be $1.193 - 0.667 = 0.526$. This corresponds to $0.526 * 30 = 16$ missing observations among the $1.193 * 30 = 36$ largest extremes. After having removed $36 - 16 = 20$ largest values from the data set and re-plotting the curves corresponding to the four value of γ above, we see the results in Fig. 9. The curves are roughly linear, with the red curve having the slope closest to 1.

Appendix

Second-order regular variation

Second-order regular variation can be thought of as a way to quantify the vanishing difference between the left hand side and the right hand side of (1.1). It assumes that there is $\rho \le 0$ and a positive or negative function A that is regularly varying with exponent ρ and $\lim_{t\to\infty} A(t) = 0$, such that for $x > 0$,

$$\lim_{t\to\infty} \frac{\log U(tx) - \log U(t) - \gamma \log x}{A(t)} = \begin{cases} \frac{x^\rho - 1}{\rho} & \rho < 0, \\ \log x & \rho = 0, \end{cases} \tag{7.1}$$

where $U(t) = F^{\leftarrow}(1 - 1/t)$ and $F^{\leftarrow}$ is the generalized inverse of F; see e.g. de Haan and Ferreira (2006).

The results of this paper assume that the sequence (k_n) used to define our estimators satisfies

$$\lim_{n\to\infty} \sqrt{k_n} A(n/k_n) = \lambda \tag{7.2}$$

for some $\lambda \in \mathbb{R}$. Since $k_n \to \infty$, condition (7.2) implies that $n/k_n \to \infty$.

Distributions that satisfy the second-order condition include the Student's t_ν, stable, and Fréchet distributions; see, e.g. Drees (1998) and Drees et al. (2000). In fact, any distribution with $\bar{F}(x) = c_1 x^{-\alpha} + c_2 x^{-\alpha+\alpha\rho}(1 + o(1))$ as $x \to \infty$, where $c_1 > 0$, $c_2 \ne 0$, $\alpha > 0$ and $\rho < 0$, satisfies the second-order condition with the indicated values of α and ρ (de Haan and Ferreira 2006).

Proofs

In this section we present the proofs of the results in the earlier parts of the paper.

Proof of Lemma 3.1 Since

$$\partial_1 L_n(\gamma_0, \delta_0) = \frac{2m}{\gamma_0} - \frac{2}{\gamma_0} \sum_{i=1}^m \omega_{i,\delta_0} Y_{ni}^2 - \frac{2\sqrt{k_n}}{\gamma_0} \sum_{i=1}^m \omega_{i,\delta_0} h_{\delta_0,i} Y_{ni}$$

and

$$\partial_2 L_n(\gamma_0, \delta_0) = -\sum_{i=1}^m \frac{\omega'_{i,\delta_0}}{\omega_{i,\delta_0}} + \sum_{i=1}^m \omega'_{i,\delta_0} Y_{ni}^2 - 2\sqrt{k_n} \sum_{i=1}^m \omega_{i,\delta_0} h'_{\delta_0,i} Y_{ni},$$

the claim of the lemma follows from (3.4). □

Proof of Lemma 3.2 We proceed as in the proof of Lemma 3.1, except now one needs to take second derivatives. For example, elementary calculations give us

$$\frac{\partial_1^2 L_n(\tilde\gamma,\tilde\delta)}{k_n} = -\frac{2k}{\tilde\gamma^2 k_n} + \frac{6\gamma_0^2}{\tilde\gamma^4 k_n}\sum_{i=1}^m \omega_{i,\tilde\delta} Y_{ni}^2 - \frac{12\gamma_0}{\tilde\gamma^4\sqrt{k_n}}\sum_{i=1}^m \omega_{i,\tilde\delta}(\tilde\gamma h_{\tilde\delta,i} - \gamma_0 h_{\delta_0,i})Y_{ni}$$
$$+ \frac{6}{\tilde\gamma^4}\sum_{i=1}^m \omega_{i,\tilde\delta}(\tilde\gamma h_{\tilde\delta,i} - \gamma_0 h_{\delta_0,i})^2 + \frac{8\gamma_0}{\tilde\gamma^3\sqrt{k_n}}\sum_{i=1}^m \omega_{i,\tilde\delta} h_{\tilde\delta,i} Y_{ni}$$
$$- \frac{8}{\tilde\gamma^3}\sum_{i=1}^m \omega_{i,\tilde\delta} h_{\tilde\delta,i}(\tilde\gamma h_{\tilde\delta,i} - \gamma_0 h_{\delta_0,i}) + \frac{2}{\tilde\gamma^2}\sum_{i=1}^m \omega_{i,\tilde\delta} h_{\tilde\delta,i}^2.$$

Using (3.4) and the fact that $(\tilde\gamma,\tilde\delta)\xrightarrow{P}(\gamma_0,\delta_0)$ we see that

$$\frac{\partial_1^2 L_n(\tilde\gamma,\tilde\delta)}{k_n} \xrightarrow{P} \frac{2}{\gamma_0^2}\sum_{i=1}^m \omega_{i,\delta_0} h_{\delta_0,i}^2 = \frac{2b_m}{\gamma_0^2}.$$

The other terms of the Hessian matrix can be handled in a similar manner. □

Proof of Lemma 3.3 Denote

$$L(\gamma,\delta) = \gamma^{-2}\sum_{i=1}^m \omega_{i,\delta}(\gamma h_{\delta,i} - \gamma_0 h_{\delta_0,i})^2,\ (\gamma,\delta)\in\Theta.$$

Since we can write

$$L_n(\gamma,\delta)/k_n = \frac{2m\log\gamma}{k_n} - \frac{1}{k_n}\sum_{i=1}^m \log\omega_{i,\delta} + \frac{\gamma_0^2}{\gamma^2 k_n}\sum_{i=1}^m \omega_{i,\delta} Y_{ni}^2$$
$$- \frac{2\gamma_0}{\gamma^2\sqrt{k_n}}\sum_{i=1}^m \omega_{i,\delta}(\gamma h_{\delta,i} - \gamma_0 h_{\delta_0,i})Y_{ni} + \frac{1}{\gamma^2}\sum_{i=1}^m \omega_{i,\delta}(\gamma h_{\delta,i} - \gamma_0 h_{\delta_0,i})^2,$$

we have

$$\sup_{(\gamma,\delta)\in\Theta}\left|\frac{L_n(\gamma,\delta)}{k_n} - L(\gamma,\delta)\right|$$
$$\le \sup_{(\gamma,\delta)\in\Theta}\left|\frac{2m\log\gamma}{k_n} - \frac{1}{k_n}\sum_{i=1}^m \log\omega_{i,\delta}\right| + \sup_{(\gamma,\delta)\in\Theta}\left|\frac{\gamma_0^2}{\gamma^2 k_n}\sum_{i=1}^m \omega_{i,\delta} Y_{ni}^2\right|$$
$$+ \sup_{(\gamma,\delta)\in\Theta}\left|\frac{2\gamma_0}{\gamma^2\sqrt{k_n}}\sum_{i=1}^m \omega_{i,\delta}(\gamma h_{\delta,i} - \gamma_0 h_{\delta_0,i})Y_{ni}\right| \xrightarrow{P} 0,\ \ n\to\infty,$$

by (3.4), since we know that, by assumption, $\gamma, \omega_{i,\delta}$ and $h_{\delta,i}$ are bounded away from 0 and infinity on Θ.

Clearly, the point (γ_0,δ_0) is a minimizer of the function $\gamma^2 L$. Furthermore, it is elementary to check that the Hessian matrix of $\gamma^2 L$ at that point is equal to $2\gamma_0^2\Gamma_m$. We will see in the proof of Proposition 3.1 below that the matrix Γ_m is invertible, hence the point (γ_0,δ_0) is the unique minimizer of the function $\gamma^2 L$, hence also of the function L. The uniform convergence in probability of the function L_n/k_n to the function L implies that any minimizer of the former function converges in probability to the unique minimizer of the limit function. Hence the statement of the lemma. □

Proof of Proposition 3.1 Introduce functions of $x > 0$

$$l_\delta(x) = x^2/(x+\delta), \;\; m_\delta(x) = x^2 v(x/\delta)/\delta = x - 2\delta\log(1+x/\delta) + \delta x/(x+\delta)\,,$$

so that

$$\begin{aligned} \omega_{i,\delta_0} &= \theta_i^2/\big(m_{\delta_0}(\theta_i) - m_{\delta_0}(\theta_{i-1})\big),\; g_{\delta_0,i} = \big(m_{\delta_0}(\theta_i) + l_{\delta_0}(\theta_i)\big)/2\theta_i, \\ g'_{\delta_0,i} &= \big(m_{\delta_0}(\theta_i) - l_{\delta_0}(\theta_i)\big)/2\delta_0\theta_i,\;\; i = 1,\dots,m\,. \end{aligned}$$

Therefore we can write

$$\begin{aligned} b_m =& \frac{m_{\delta_0}(\theta_m) + 2l_{\delta_0}(\theta_m)}{4} + \frac{1}{4}\sum_{i=1}^m \frac{\big(l_{\delta_0}(\theta_i) - l_{\delta_0}(\theta_{i-1})\big)^2}{m_{\delta_0}(\theta_i) - m_{\delta_0}(\theta_{i-1})}, \\ c_m =& \frac{m_{\delta_0}(\theta_m) - 2l_{\delta_0}(\theta_m)}{4\delta_0^2} + \frac{1}{4\delta_0^2}\sum_{i=1}^m \frac{\big(l_{\delta_0}(\theta_i) - l_{\delta_0}(\theta_{i-1})\big)^2}{m_{\delta_0}(\theta_i) - m_{\delta_0}(\theta_{i-1})}, \\ d_m =& \frac{m_{\delta_0}(\theta_m)}{4\delta_0} - \frac{1}{4\delta_0}\sum_{i=1}^m \frac{\big(l_{\delta_0}(\theta_i) - l_{\delta_0}(\theta_{i-1})\big)^2}{m_{\delta_0}(\theta_i) - m_{\delta_0}(\theta_{i-1})}. \end{aligned}$$

We now show that the matrix Γ_m is invertible. A direct computation shows that

$$4\delta_0^2(b_m c_m - d_m^2) = m_{\delta_0}(\theta_m)\sum_{i=1}^m \frac{\big(l_{\delta_0}(\theta_i) - l_{\delta_0}(\theta_{i-1})\big)^2}{m_{\delta_0}(\theta_i) - m_{\delta_0}(\theta_{i-1})} - l_{\delta_0}^2(\theta_m)\,.$$

It is easy to check that the functions l_{δ_0} and m_{δ_0} are increasing on $(0,\infty)$, so that for any $i \ge 1$, $l_{\delta_0}(\theta_i) - l_{\delta_0}(\theta_{i-1}) > 0$ and $m_{\delta_0}(\theta_i) - m_{\delta_0}(\theta_{i-1}) > 0$. Further, by the Cauchy-Schwarz inequality,

$$\begin{aligned} m_{\delta_0}(\theta_m)\sum_{i=1}^m \frac{\big(l_{\delta_0}(\theta_i) - l_{\delta_0}(\theta_{i-1})\big)^2}{m_{\delta_0}(\theta_i) - m_{\delta_0}(\theta_{i-1})} &= \sum_{i=1}^m \big(m_{\delta_0}(\theta_i) - m_{\delta_0}(\theta_{i-1})\big)\sum_{i=1}^m \frac{\big(l_{\delta_0}(\theta_i) - l_{\delta_0}(\theta_{i-1})\big)^2}{m_{\delta_0}(\theta_i) - m_{\delta_0}(\theta_{i-1})} \\ &\ge \left(\sum_{i=1}^m \big(l_{\delta_0}(\theta_i) - l_{\delta_0}(\theta_{i-1})\big)\right)^2 = l_{\delta_0}^2(\theta_m), \end{aligned}$$

and the equality holds if and only if

$$\frac{l_{\delta_0}(\theta_1)}{m_{\delta_0}(\theta_1)} = \frac{l_{\delta_0}(\theta_2) - l_{\delta_0}(\theta_1)}{m_{\delta_0}(\theta_2) - m_{\delta_0}(\theta_1)} = \cdots = \frac{l_{\delta_0}(\theta_m) - l_{\delta_0}(\theta_{m-1})}{m_{\delta_0}(\theta_m) - m_{\delta_0}(\theta_{m-1})}\,.$$

The latter requirement is equivalent to

$$\frac{l_{\delta_0}(\theta_1)}{m_{\delta_0}(\theta_1)} = \frac{l_{\delta_0}(\theta_2)}{m_{\delta_0}(\theta_2)} = \cdots = \frac{l_{\delta_0}(\theta_m)}{m_{\delta_0}(\theta_m)}, \tag{7.3}$$

so invertibility of Γ_m will follow once we show that (7.3) cannot hold. If we put

$$Q(x) =: m_{\delta_0}(x) - \frac{l_{\delta_0}(x)m'_{\delta_0}(x)}{l'_{\delta_0}(x)} = m_{\delta_0}(x) - \frac{x^3}{(x+\delta_0)(x+2\delta_0)},$$

then $Q(0) = 0$ and

$$Q'(x) = -\frac{2\delta_0 x^2}{(x+\delta_0)(x+2\delta_0)^2} < 0, \quad x > 0,$$

which implies that

$$Q(x) = \frac{l'_{\delta_0}(x)}{m^2_{\delta_0}(x)}\left(\frac{l_{\delta_0}(x)}{m_{\delta_0}(x)}\right)' < 0$$

for any $x > 0$. Since $l'_{\delta_0}(x) > 0$, we conclude that the function $l_{\delta_0}(x)/m_{\delta_0}(x)$ is strictly decreasing on the positive half line, and so (7.3) cannot hold. Hence the matrix Γ_m is invertible.

It is elementary to check that, as $\delta_0 \to 0$,

$$b_m \to \theta_m, \quad c_m \sim \theta_1^{-1}(\log \delta_0)^2, \quad d_m \to \log \delta_0 . \tag{7.4}$$

Substituting this into (3.10) shows convergence of the variance in (3.12).

Similarly, it is elementary to check that, as $\delta_0 \to 0$,

$$a_1 \to \frac{2\lambda}{1-\rho}\gamma_0^{-2}\theta_m^{1-\rho}, \quad a_2 \sim \frac{2\lambda}{1-\rho}\gamma_0^{-1}\theta_1^{-\rho}\log \delta_0 . \tag{7.5}$$

Substituting (7.4) and (7.5) into (3.10) and (3.11) proves convergence of the mean in (3.12). □

Proof of Lemma 4.1 By (4.3),

$$\partial_1 L_n(\gamma_0, \delta_0) = \frac{2k_n}{\gamma_0} - \frac{2\omega_{1,\delta_0}\eta_n^2}{\gamma_0} - \frac{2\sqrt{k_n}}{\gamma_0}\omega_{1,\delta_0} g_{\delta_0,1}\eta_n - \frac{2}{\gamma_0}\sum_{i=2}^{k_n} Z_{i,n}$$

and

$$\partial_2 L_n(\gamma_0, \delta_0) = -\frac{\omega'_{1,\delta_0}}{\omega_{1,\delta_0}} - \sum_{i=2}^{k_n}\frac{2}{\delta_0+\theta_{i,n}} + \omega'_{1,\delta_0}\eta_n^2$$
$$- 2\sqrt{k_n}\omega_{1,\delta_0} g'_{\delta_0,1}\eta_n + 2\sum_{i=2}^{k_n}\frac{Z_{i,n}}{\delta_0+\theta_{i,n}}.$$

Since

$$\begin{pmatrix} -\frac{2}{\gamma_0}\omega_{1,\delta_0} g_{\delta_0,1}\eta_n \\ -2\omega_{1,\delta_0} g'_{\delta_0,1}\eta_n \end{pmatrix} \Rightarrow N\big(0, 4\Gamma_1\big)\,,$$

the claim of the lemma will follow once we show that

$$\begin{pmatrix} -k_n^{-1/2}\gamma_0^{-1}\sum_{i=2}^{k_n}\big(Z_{i,n}-1\big) \\ k_n^{-1/2}\sum_{i=2}^{k_n}\frac{Z_{i,n}-1}{\delta_0+\theta_{i,n}} \end{pmatrix} \Rightarrow N\big(0, \Gamma_0\big)\,,$$

where Γ_0 is the second matrix in the right hand side of (4.4). By (4.1) we only need to prove that

$$\begin{pmatrix} -k_n^{-1/2}\gamma_0^{-1}\sum_{i=2}^{k_n}\frac{k_n(\delta_0+\theta_{i,n})}{[\delta_0 k_n]+[\theta_{i,n}k_n]}\big(E^*_{[\delta_0 k_n]+[\theta_{i,n}k_n]}-1\big) \\ k_n^{-1/2}\sum_{i=2}^{k_n}\frac{k_n}{[\delta_0 k_n]+[\theta_{i,n}k_n]}\big(E^*_{[\delta_0 k_n]+[\theta_{i,n}k_n]}-1\big) \end{pmatrix} \Rightarrow N\big(0, \Gamma_0\big)\,. \tag{7.6}$$

Since the covariance matrix of the random vector in the left hand side of (7.6) converges to Γ_0, only the Lyapunov condition needs to be checked for an application of the central limit theorem. The latter can be performed component-wise and is elementary when taking, for instance, the 4th powers of the terms. □

Proof of Lemma 4.2 Once again, computing the second derivatives, we obtain, for example,

$$\begin{aligned} \frac{\partial_1^2 L_n(\tilde\gamma,\tilde\delta)}{k_n} = &-\frac{2}{\tilde\gamma^2} + \frac{6\omega_{1,\tilde\delta}}{\tilde\gamma^4 k_n}\Big(\gamma_0\eta_n - \sqrt{k_n}(\tilde\gamma g_{\tilde\delta,1} - \gamma_0 g_{\delta_0,1})\Big)^2 \\ &+ \frac{8\omega_{1,\tilde\delta} g_{\tilde\delta,1}}{\tilde\gamma^3\sqrt{k_n}}\Big(\gamma_0\eta_n - \sqrt{k_n}(\tilde\gamma g_{\tilde\delta,1} - \gamma_0 g_{\delta_0,1})\Big) + \frac{2}{\tilde\gamma^2}\omega_{1,\tilde\delta} g^2_{\tilde\delta,1} \\ &+ \frac{4\gamma_0}{\tilde\gamma^3 k_n}\sum_{i=2}^{k_n}\frac{(\tilde\delta+\theta_{i,n})Z_{i,n}}{\delta_0+\theta_{i,n}}\,. \end{aligned}$$

Clearly, the second and the third terms in the right hand side are $o_p(1)$ as $n \to \infty$. Furthermore,

$$-\frac{2}{\tilde\gamma^2} \to -\frac{2}{\gamma_0^2},\quad \frac{2}{\tilde\gamma^2}\omega_{1,\tilde\delta} g^2_{\tilde\delta,1} \to \frac{2}{{\gamma_0}^2}\omega_{1,\delta_0} g^2_{\delta_0,1} \text{ in probability,}$$

and by computing the mean and the variance we see that

$$\frac{4\gamma_0}{\tilde\gamma^3 k_n}\sum_{i=2}^{k_n}\frac{(\tilde\delta+\theta_{i,n})Z_{i,n}}{\delta_0+\theta_{i,n}} \to \frac{4}{\gamma_0^2}$$

in probability. Therefore,

$$\frac{\partial_1^2 L_n(\tilde{\gamma}, \tilde{\delta})}{k_n} \to \frac{2}{\gamma_0^2} + \frac{2}{\gamma_0{}^2}\omega_{1,\delta_0} g_{\delta_0,1}^2$$

in probability, and the limit is the appropriate entry in the matrix $2\Gamma_\infty$. The other terms of the Hessian matrix can be handled in a similar manner. □

Proof of Lemma 4.3 We proceed as in the proof of Lemma 3.3. Denote now

$$\begin{aligned} L(\gamma,\delta) =& 2\log\gamma + \frac{\tilde{\omega}_{1,\delta}}{\gamma^2}(\gamma \tilde{g}_{\delta,1} - \gamma_0 \tilde{g}_{\delta_0,1})^2 \\ & - 2\int_\varepsilon^{1+\varepsilon} \log(\delta + x)\,dx + \frac{2\gamma_0}{\gamma}\int_\varepsilon^{1+\varepsilon} \frac{\delta + x}{\delta_0 + x}\,dx\,, \end{aligned}$$

where $\tilde{\omega}_{1,\delta}$ is defined as $\omega_{1,\delta}$ and $\tilde{g}_{\delta,1}$ is defined as $g_{\delta,1}$, both with $\theta_1 = \varepsilon$. Since we can write

$$\begin{aligned} L_n(\gamma,\delta)/k_n =& 2\log\gamma - \frac{1}{k_n}\log\omega_{1,\delta} - \frac{2}{k_n}\sum_{i=2}^{k_n}\log(\delta + \theta_{i,n}) + \frac{\gamma_0^2\omega_{1,\delta}}{\gamma^2 k_n}\eta_n^2 \\ & - \frac{2\gamma_0\omega_{1,\delta}}{\gamma^2\sqrt{k_n}}(\gamma g_{\delta,1} - \gamma_0 g_{\delta_0,1})\eta_n + \frac{\omega_{1,\delta}}{\gamma^2}(\gamma g_{\delta,1} - \gamma_0 g_{\delta_0,1})^2 \\ & + \frac{2\gamma_0}{\gamma k_n}\sum_{i=2}^{k_n}\frac{(\delta + \theta_{i,n})Z_{i,n}}{\delta_0 + \theta_{i,n}}, \end{aligned}$$

it follows that

$$\begin{aligned} & \sup_{(\gamma,\delta)\in\Theta}\left|\frac{L_n(\gamma,\delta)}{k_n} - L(\gamma,\delta)\right| \\ & \le \sup_{(\gamma,\delta)\in\Theta}\left|\frac{\omega_{1,\delta}}{\gamma^2}(\gamma g_{\delta,1} - \gamma_0 g_{\delta_0,1})^2 - \frac{\tilde{\omega}_{1,\delta}}{\gamma^2}(\gamma \tilde{g}_{\delta,1} - \gamma_0 \tilde{g}_{\delta_0,1})^2\right| \\ & \quad + \sup_{(\gamma,\delta)\in\Theta}\left|-\frac{1}{k_n}\log\omega_{1,\delta} + \frac{\gamma_0^2\omega_{1,\delta}}{\gamma^2 k_n}\eta_n^2 - \frac{2\gamma_0\omega_{1,\delta}}{\gamma^2\sqrt{k_n}}(\gamma g_{\delta,1} - \gamma_0 g_{\delta_0,1})\eta_n\right| \\ & \quad + \sup_{(\gamma,\delta)\in\Theta}\left|\frac{2}{k_n}\sum_{i=2}^{k_n}\log(\delta + \theta_{i,n}) - 2\int_\varepsilon^{1+\varepsilon}\log(\delta + x)\,dx\right| \\ & \quad + \sup_{(\gamma,\delta)\in\Theta}\left|\frac{2\gamma_0}{\gamma k_n}\sum_{i=2}^{k_n}\frac{(\delta + \theta_{i,n})Z_{i,n}}{\delta_0 + \theta_{i,n}} - \frac{2\gamma_0}{\gamma}\int_\varepsilon^{1+\varepsilon}\frac{\delta + x}{\delta_0 + x}dx\right|. \end{aligned}$$

It is clear that the first three terms in the right hand side vanish as $n \to \infty$. The same is true for the last term in the right hand side because we can bound the latter by

$$\sup_{(\gamma,\delta)\in\Theta}\left|\frac{2\gamma_0\delta}{\gamma}\right|\cdot\left|\frac{1}{k_n}\sum_{i=2}^{k_n}\frac{Z_{i,n}}{\delta_0+\theta_{i,n}}-\int_\varepsilon^{1+\varepsilon}\frac{1}{\delta_0+x}dx\right|$$
$$+\sup_{(\gamma,\delta)\in\Theta}\left|\frac{2\gamma_0}{\gamma}\right|\cdot\left|\frac{1}{k_n}\sum_{i=2}^{k_n}\frac{\theta_{i,n}Z_{i,n}}{\delta_0+\theta_{i,n}}-\int_\varepsilon^{1+\varepsilon}\frac{x}{\delta_0+x}dx\right|.$$

It is clear that both suprema are finite, while by computing once again the means and the variances we see that the two differences converge to zero in probability.

Clearly, the point (γ_0, δ_0) is a minimizer of the function $\tilde{\omega}_{1,\delta}\gamma^{-2}(\gamma\tilde{g}_{\delta,0} - \gamma_0\tilde{g}_{\delta_0,1})^2$. Let us denote the remaining part of the function $L(\gamma, \delta)$ by $L_1(\gamma, \delta)$. To check that the point (γ_0, δ_0) is a unique minimizer of the latter function, note that for a fixed value of δ the unique minimizer of $L_1(\cdot, \delta)$ is the point

$$\gamma(\delta) = \gamma_0\int_\varepsilon^{1+\varepsilon}\frac{\delta+x}{\delta_0+x}\,dx.$$

Since, up to δ-independent terms,

$$L_1(\gamma(\delta),\delta) = \log\left(\int_\varepsilon^{1+\varepsilon}\frac{\delta+x}{\delta_0+x}\,dx\right) - \int_\varepsilon^{1+\varepsilon}\log\left(\frac{\delta+x}{\delta_0+x}\right)dx\,,$$

which vanishes for $\delta = \delta_0$ and is strictly positive by Jensen's inequality for $\delta \neq \delta_0$, we see that $\delta = \delta_0$ and $\gamma = \gamma(\delta_0 = \gamma_0$ is the unique minimizer of L_1 and, hence, also of L.

As before, the uniform convergence of L_n/k_n to L implies now that any minimizer of the former function convergence in probability to (γ_0, δ_0). Lemma 4.2 and the fact that Γ_∞ is invertible mean that, with probability converging to 1, the minimizer of L_n is unique. □

Acknowledgements We would like to thank the two anonymous referees for their useful comments that forced us to make changes that were truly necessary.

Data availibility The Danish fire data and the Natural and Climate Disasters in the U.S. data are publicly available as indicated in the text. The Google+ data are not publicly available.

References

Aban, I., Meerschaert, M., Panorska, A.: Parameter estimation for the truncated Pareto distribution. J. Am. Stat. Assoc. **101**, 270–277 (2006)

Beirlant, J., Alves, I., Gomes, I.: Tail fitting for truncated and non-truncated Pareto-type distributions. Extremes **19**, 429–462 (2016)

Beirlant, J., Fraga Alves, I., Reynkens, T.: Fitting tails affected by truncation. Electron. J. Stat. **11**, 2026–2065 (2017)

Benchaira, S., Meraghmi, D., Necir, A.: Tail product-limit process for truncated data with application to extreme value index estimation. Extremes **19**, 219–251 (2016)

Bhattacharya, S., Kallitsis, M., Stoev, S.: Data-adaptive trimming of the Hill estimator and detection of outliers in the extremes of heavy-tailed data. Electronic J. Stat. **13**, 1872–1925 (2019)
de Haan, L., Ferreira, A.: Extreme Value Theory: An Introduction. Springer, New York (2006)
Drees, H.: On smooth statistical tail functionals. Scand. J. Stat. **25**, 187–210 (1998)
Drees, H., de Haan, L., Resnick, S.: How to make a Hill plot. Ann. Stat. **28**, 254–274 (2000)
Einmahl, J.H., Fils-Villetard, A., Guillou, A., et al.: Statistics of extremes under random censoring. Bernoulli **14**, 207–227 (2008)
Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events for Insurance and Finance. Springer-Verlag, Berlin (1997)
Hill, B.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**, 1163–1174 (1975)
Newman, M.: Networks: An Introduction. Oxford University Press, Oxford (2010)
Reiss, R.-D.: Asymptotic Distribution of Order Statistics. Springer, New York (1989)
Resnick, S.: Discussion of the Danish data on large fire insurance losses. Astin Bull. **27**, 139–151 (1997)
Resnick, S.: Heavy-Tail Phenomena: Probabilistic and Statistical Modeling. Springer, New York (2007)
Smith, A., Katz, R.: US billion-dollar weather and climate disasters: data sources, trends, accuracy and biases. Nat. Hazards **67**, 387–410 (2013)
Stupfler, G.: Estimating the conditional extreme-value index under random right-censoring. J. Multivar. Anal. **144**, 1–24 (2016)
Worms, J., Worms, R.: New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. Extremes **17**, 337–358 (2014)
Zou, J., Davis, R., Samorodnitsky, G.: Extreme value analysis without the largest values: what can be done? Probab. Eng. Inf. Sci. (2019). https://doi.org/10.1017/S0269964818000542:1-21

Authors and Affiliations

Hui Xu[1] · **Richard Davis**[2] · **Gennady Samorodnitsky**[3]

Hui Xu
hx223@cornell.edu

Gennady Samorodnitsky
gs18@cornell.edu

1 Center for Applied Mathematics, Cornell University, Ithaca, New York, USA

2 Department of Statistics, Columbia University, New York, New York, USA

3 School of Operations Research and Information Engineering, Cornell University, Ithaca, New York, USA