



# 1 Introduction

Tail risk and extreme events are important research topics in economics and finance. In many applications, the features of interest are tail properties such as tail index and extreme quantiles. Existing literature has extensively studied the case with fully observed datasets. In comparison, this article explores the case with tail censoring. We argue that it is important to take into account the censoring if the research interest is in the tail, even when the censoring fraction is small. We provide a new method to construct estimators and confidence intervals for tail features.

Suppose one has a random sample from some underlying distribution  $F$ , where the observations larger than some threshold  $T$  are replaced with  $T$  or simply unobserved. In principle, tail features cannot be even identified if they entirely depend on the right tail part of  $F$  that is beyond  $T$ . However, we can back out the tail-related features by extrapolation under two assumptions. They are that (i) the tail of  $F$  can be well approximated by some suitably chosen parametric distribution, and (ii)  $T$  is sufficiently large so that only a small fraction of samples are censored. The first assumption has been thoroughly studied in the statistical literature and is satisfied by many commonly used distributions. The second assumption is also satisfied in many interesting empirical applications, which motivate this article.

Our first motivating example is the Current Population Survey (CPS) dataset, which has been the primary data source used for investigating the distributions of individual earnings and household income in the US. Featured studies using CPS data include Armour, Burkhauser, and Larrimore (2013) and Eika, Mogstad, and Zafar (2019), among many others. In CPS, the individual earnings larger than some threshold  $T$  are typically censored (also called topcoded) and replaced with  $T$  for confidential reasons.<sup>1</sup> In 2019, the censoring threshold is 310000 USD, leading to an approximately 0.58% censoring fraction in the full sample of individuals between 18 and 70 years old. This quantity is also substantially different across the subsamples defined by race and gender but remains small, as seen in Table 1. Using this dataset, we aim to estimate and construct confidence intervals for the tail index that measures the tail heaviness of the income distribution and the extreme quantiles.

---

<sup>1</sup>The topcoding has constantly been changing. Description of the topcoding mechanism is available at [https://cps.ipums.org/cps/topcodes\\_tables](https://cps.ipums.org/cps/topcodes_tables).

Table 1: Fractions and Numbers of Censored Observations in the 2019 CPS Dataset

	<i>n</i>	cen%	cen#	<i>n</i>	cen%	cen#
full sample	115424	0.582	672			
racengender		male			female	
all	55553	0.884	491	59871	0.302	181
white	43371	0.966	419	45424	0.310	141





in the Appendix.

## 2 The Maximum Likelihood Estimator

Consider a random sample  $\{Y_i\}_{i=1}^n$  generated from some cumulative distribut

of one of the three limit laws. The parameter  $\alpha$  is referred to as the tail index, which is uniquely determined by  $F$





and hence researchers typically choose a sufficiently small  $k$  to retain the asymptotic zero mean. This is similar in spirit to the undersmoothing in standard kernel regressions.

In practice, it makes no difference, at least asymptotically, between treating  $T_n$  as constant and the number of censored observations  $m$  as stochastic and the opposite treatment.

inf  $y : 1 - p = F(y)$ . We set  $p = p_n \rightarrow 0$  to capture the extremeness. The estimator can be constructed by inverting (2), that is,

$$\hat{Q}(1 - p_n) = u_n + \frac{1}{\lambda} \hat{d}_n^{-1} ;$$

where  $d_n = (m + k) = (p_n n)$ . To derive a non-trivial asymptotic result, we let  $d_0 = \lim_{n \rightarrow \infty} d_n > 0$  so that the target quantile is of the same or larger magnitude of  $u_n$  (otherwise it can be estimated by the corresponding empirical quantile). The following proposition derives the asymptotic distribution of  $\hat{Q}(1 - p_n)$ .

**Proposition 2** *Suppose Conditions 1-4 hold. If  $d_0 > 0$ , then*

$$k^{1/2} \frac{\hat{Q}(1 - p_n) - Q(1 - p_n)}{q(d_n)} \xrightarrow{d} N(0, 1)$$

where  $q(t) = -t \log t$  and

$$= \left(1 - \frac{d_0}{q(d_0)}\right) M^{-1} \left(1 - \frac{d_0}{q(d_0)}\right) + q(d_0)^{-2} ;$$

Proposition 2 establishes the asymptotic normality of the extreme quantile estimator. Then the confidence intervals for  $u_n$  and  $Q(1 - p_n)$  can be constructed by plugging in the estimators for the asymptotic variance.

### 3 Small Sample Modification under the Fixed- $k$ Asymptotics

The results in Section 2 suggest that the asymptotic normal approximation can be used for inference about the tail features as  $k$  goes to  $\infty$ . In practice, however, the choice of the tail sample size  $k$  is widely accepted as a challenging question even without censoring. This is because a good selection of  $k$  has to balance the tail approximation bias and the variance delicately. Ultimately, the underlying distribution has to be reasonably close to the Pareto distribution in the tail to guarantee a satisfactory finite sample performance.

The asymptotic approximation can be quite accurate for some cases, but it is also easy to find examples where the limiting normal distribution provides a poor approximation. Con-

sider the example that  $F$  is a mixture of the standard normal distribution with probability 0.8 and some Pareto distribution with probability 0.2. Such a mixture structure implies that only the very few largest observations are informative about the true tail. In this case, choosing a large  $k$  means including too many contaminating observations from the mid-sample, while choosing a small  $k$  invalidates the asymptotic Gaussianity. In principle, there is no such a procedure that consistently justifies whether a given  $k$  is appropriate when  $F$  is entirely unknown. See Theorem 5.1 in Müller and Wang (2017) for a discussion on the non-censored case.

Therefore, in this article, we do not focus on the choice of  $k$  but instead treat it as given. In some cases,  $k$  is determined by some economic theory or empirical guidance. For example, in the macroeconomic disaster application, the economic definition of disasters for more than 10% of GDP decline yields the choice of  $k$ . In other cases, we may employ some data-driven algorithms that balance the Pareto approximation bias and the variance. See, for example, Hall (1982), Drees (2001), and Clauset, Shalizi, and Newman (2009).

When  $k$  and  $n=k$  are both sufficiently large, we would expect the MLE in (5) based on the increasing- $k$  asymptotics to work well. Nevertheless,  $k$  is only moderate in some situations, including our macroeconomic disaster application and the Asian male subsample in CPS. This causes a small sample issue that the asymptotic Gaussianity is questionable. To find a better alternative, we resort to the asymptotic embedding that requires  $n$  diverges, but  $k$  remains a fixed constant. Under this fixed- $k$  asymptotic framework, the consistent estimation of the tail index and the extreme quantiles are out of the question since the tail sample size is fixed. Fortunately, inference about these tail features is still implementable, as we discuss in this section.

We first study the tail index. EV theory (the Fisher–Tippett–Gnedenko theorem) suggests that when the underlying distribution is within the maximum domain of attraction (e.g., Chapter 1 of de Haan and Ferreira (2007)), the sample maximum is asymptotically distributed as the EV distribution, which is parametric and entirely characterized by. Specifically, our Condition 2 is sufficient for the maximum domain of attraction assumption. Then, EV theory implies that there exist sequences of constants  $a_n$  and  $b_n$  such that, up to some location and scale normalization,

$$\frac{Y_{(1)} - b_n}{a_n} \xrightarrow{d} X_1; \quad (6)$$

where the CDF of  $X_1$  is given by

$$V(x) = 1 - \exp$$

If the constants  $a_n$  and  $b_n$  were known, the vector

$$Y = Y_{(m+1)}; \dots; Y_{(m+k)}$$

is then approximately distributed as  $X$ , and the limiting problem is reduced to the small sample parametric one: constructing a confidence interval based on one draw  $X$  whose density  $f_{X|}$  is known up to . However,  $a_n$  and  $b_n$  respectively correspond to the scale and the tail location  $u$ . Therefore, they ultimately depend on  $F$  and are challenging to estimate. Consider the standard Pareto distribution, for example. The Pareto exponent is simply  $1 =$  . Then the fact that  $a_n = n$  implies that a small estimation bias in could be amplified by the  $n$ -power and lead to a poor inference.

To avoid the knowledge (and estimation) of  $a_n$  and  $b_n$ , we consider the following self-normalized statistics:

$$\begin{aligned} Y^* &= \frac{Y}{Y_{(m+1)}} \frac{Y_{(m+k)}^k}{Y_{(m+k)}} \\ &= \left( 1; \frac{Y_{(m+2)}}{Y_{(m+1)}} \frac{Y_{(m+k)}}{Y_{(m+k)}}, \dots, \frac{Y_{(m+k-1)}}{Y_{(m+1)}} \frac{Y_{(m+k)}}{Y_{(m+k)}}; 0 \right) \end{aligned} \quad (10)$$

It is easy to establish that  $Y^*$  is maximal invariant with respect to the group of location and scale transformations (cf. Chapter 6 of Lehmann and Romano (2005)). In words, the statistic constructed as a function of  $Y^*$  remains unchanged if data are shifted and multiplied by any non-zero constant. This invariance is also intuitive since the tail shape should preserve no matter how data are linearly transformed.

The continuous mapping theorem and Proposition 3 yield that

$$Y^* \xrightarrow{d} X^* = \left( 1; \frac{X_{m+2}}{X_{m+1}} \frac{X_{m+k}}{X_{m+k}}, \dots, \frac{X_{m+k-1}}{X_{m+1}} \frac{X_{m+k}}{X_{m+k}}; 0 \right) ; \quad (11)$$

which is again invariant to location and scale transformation. By change of variables, the density of  $X^*$  is given by

$$f_{X^*|} (x^*) = \frac{(k+m)!}{m!} \int_0^\infty s^{k-2} \exp \left( - \sum_{i=1}^k \log(1 + x_i^* s) \right) e(x^*; s) ds; \quad (12)$$

where  $e(x^*; s) = \exp \left( - \sum_{i=1}^k \log(1 + x_i^* s) \right)$  and  $x_i^*$  denotes the  $i$ th component of  $x^*$ . See Appendix A.1 for more details.



$P(Q(1 - \rho_n) \leq S(Y)) \geq 1 - \alpha$ , at least as  $n \rightarrow \infty$ , where  $\alpha$  denotes the significance level.

To eliminate  $a_n$  and  $b_n$ , we use the self-normalized vector  $Y^*$  as in (12). Besides, we also impose location and scale equivariance on our confidence interval  $S$ . Specifically, we impose that for any constants  $a > 0$  and  $b$ , our interval  $S$  satisfies that  $S(aY + b) = aS(Y) + b$ , where  $aS(Y) + b = \{y : (y - b)/a \in S(Y)\}$ . Under this equivariance constraint, we can write

$$\begin{aligned} P(Q(1 - \rho_n) \leq S(Y)) &= P\left(\frac{Q(1 - \rho_n) - b_n}{a_n} \leq S\left(\frac{Y - b_n}{a_n}\right)\right) \\ &= P\left(\frac{Q(1 - \rho_n) - Y_{(m+k)}}{Y_{(m+1)} - Y_{(m+k)}} \leq S(Y^*)\right) \\ &\geq P\left(\frac{q(\cdot; h) - X_{m+k}}{X_{m+1} - X_{m+k}} \leq S(X^*)\right); \end{aligned}$$

where the notation  $P$  (and  $E$  below) indicates that the randomness is entirely characterized by  $\mathbf{X}$  asymptotically. The asymptotic problem then is the construction of a location and scale equivariant  $S$  that satisfies

$$P\left(\frac{q(\cdot; h) - X_{m+k}}{X_{m+1} - X_{m+k}} \leq S(X^*)\right) \geq 1 - \alpha \text{ for all } \mathbf{X} \quad (14)$$

since any  $S$  that satisfies Proposition 3 and the equivariance constraint also satisfies

$$\liminf_{n \rightarrow \infty} P(Q(1 - \rho_n) \leq S(Y)) \geq 1 - \alpha:$$

This problem involves a single observation  $\mathbf{X} \in \mathbb{R}^k$  from a parametric distribution indexed only by the scalar parameter  $\theta \in \mathbb{R}$ .

In principle, there could still be many solutions that satisfy the asymptotic size constraint. To obtain the optimal one, we consider the weighted average expected length criterion

$$Z = E[\text{lgth}(S(\mathbf{X}))] dW(\theta), \quad (15)$$

where  $W$  again denotes some weighting measure on  $\mathbb{R}$ , and  $\text{lgth}(\mathbf{A}) = \int_{\mathbb{R}} \mathbb{1}_{\{y \in \mathbf{A}\}} dy$  for any Borel set  $\mathbf{A} \subset \mathbb{R}$ .

To solve the program of minimizing (15) subject to (14) among all equivariant set estimators  $S$ , we introduce

$$Y^*(\cdot) = \frac{q(\cdot; h) - X_{m+k}}{X_{m+1} - X_{m+k}};$$





## 4 Monte Carlo Simulations

This section examines the finite sample performance of the proposed method and compares it with several popular existing methods. We generate random samples from four commonly used distributions: the generalized Pareto distribution with  $\alpha = 0.5$  and  $\beta = 1$  (GPD), the absolute value of the Student-t distribution with 2 degrees of freedom ( $|t(2)|$ ), the F distribution with parameters 4 and 4 ( $F(4,4)$ ), and the double Pareto-lognormal distribution (dPIN), that is,

$$Y = \exp(c_1 + c_2 Z_1 + Z_2 - c_3 Z_3);$$

where  $Z_1, Z_2, Z_3$  are independent and  $Z_1 \sim N(0;1)$ , and  $Z_2; Z_3 \sim \text{Exp}(1)$ . For parameter values, we set  $c_1 = 0$ ,  $c_2 = 0.5$ ,  $\alpha = 0.5$ , and  $c_3 = 1$ , which are typical values for income data as documented in Toda (2012). In particular, the dPIN distribution is the product of independent double Pareto and lognormal variables. It has been documented to fit well to size distributions of economic variables including income (Reed (2003)), city size (Giesen, Zimmermann, and Suedekum (2010)), and consumption (Toda (2017)). In all four DGP's, the true value of the tail index is 0.5. Regarding the tail censoring, we set the censoring threshold  $T$  as the 99% and 99.9% quantiles of the underlying (i)6(n)12(g)-386(i)6(n)12(c)9(o)10(m)18(

Table 2: Small Sample Properties of Estimation and Inference about Tail Index, Ignoring Tail Censoring

cen_p	1%				0.1%			
	Bias		Cov		Bias		Cov	
n=1000	Hill	GI	Hill	GI	Hill	GI	Hill	GI
GPD	-0.18	-0.24	0.03	0.00	-0.04	-0.07	0.88	0.89
t(2)	-0.16	-0.23	0.10	0.00	-0.02	-0.06	0.93	0.93
F(4,4)	-0.12	-0.20	0.39	0.05	0.03	-0.03	0.98	0.98
dPIN	-0.17	-0.24	0.05	0.00	-0.04	-0.04	0.89	0.90
n=2000	Hill	GI	Hill	GI	Hill	GI	Hill	GI
GPD	-0.18	-0.24	0.00	0.00	-0.04	-0.07	0.85	0.81
t(2)	-0.16	-0.23	0.00	0.00	-0.02	-0.06	0.93	0.86
F(4,4)	-0.12	-0.20	0.12	0.00	0.03	-0.03	0.97	0.97
dPIN	-0.17	-0.24	0.00	0.00	-0.04	-0.04	0.88	0.82
n=5000	Hill	GI	Hill	GI	Hill	GI	Hill	GI
GPD	-0.18	-0.24	0.00	0.00	-0.04	-0.08	0.73	0.45
t(2)	-0.16	-0.23	0.00	0.00	-0.02	-0.07	0.90	0.63
F(4,4)	-0.12	-0.20	0.00	0.00	0.03	-0.03	0.93	0.95
dPIN	-0.17	-0.24	0.00	0.00	-0.03	-0.08	0.77	0.47

Note: Entries are the biases and coverage probabilities (Cov) of the 95% confidence intervals based on Hill's estimator (Hill) and Gabaix and Ibragimov (2010)'s estimator (GI). Data are generated from the Pareto(0.5), the absolute value of Student-t(2), the F(4,4), and the dPIN distributions with the censored probability (cen\_p) being 1% or 0.01%. The results are based on 1000 simulations.

Now we implement the new method proposed in Sections 2 and 3. Table 3 depicts the coverage and length of the 95% maximum likelihood confidence intervals (denoted ml) based on Proposition 1 and those of the fixed- $k$  intervals (denoted fk) by inverting (13). Several interesting findings can be made as follows. First, the maximum likelihood confidence intervals are substantially longer than the fixed- $k$  ones when the sample size is not large. Besides, the coverage probability is smaller than the nominal level when the censoring is at the 99.9% quantile. This is because the asymptotic normality cannot perform well when  $k$  is not large. Second, in comparison, the fixed- $k$  ones always deliver the nominal size with shorter length, especially when the sample size is not large. Finally, when  $n$  reaches

5000 (and  $k$  reaches 250), the maximum likelihood intervals are comparable with the fixed- $k$  ones. Hence a simple rule-of-thumb choice of the switching cutoff is  $k \geq 250$ , provided  $n$  is sufficiently large.

Table 3: Small Sample Properties of Inference about Tail Index

cen_p	1%				0.1%			
	Cov		Lgth		Cov		Lgth	
	ml	fk	ml	fk	ml	fk	ml	fk
n=1000								
GPD	0.98	0.93	1.39	0.73	0.91	0.95	0.88	0.70
t(2)	0.98	0.94	1.40	0.73	0.90	0.95	0.87	0.70
F(4,4)	0.99	0.93	1.39	0.73	0.90	0.95	0.87	0.70
dPIN	0.99	0.93	1.30	0.73	0.90	0.94	0.87	0.70
n=2000								
GPD	0.96	0.94	0.99	0.69	0.93	0.94	0.63	0.58
t(2)	0.96	0.93	0.99	0.69	0.92	0.93	0.62	0.58
F(4,4)	0.96	0.94	0.99	0.69	0.93	0.94	0.63	0.58
dPIN	0.96	0.94	0.99	0.69	0.92	0.93	0.62	0.58
n=5000								
GPD	0.97	0.94	0.63	0.54	0.95	0.94	0.40	0.39
t(2)	0.96	0.94	0.62	0.54	0.93	0.92	0.40	0.38
F(4,4)	0.97	0.95	0.63	0.54	0.94	0.93	0.40	0.38
dPIN	0.96	0.93	0.62	0.54	0.94	0.94	0.40	0.38

Note: Entries are the coverage probabilities (Cov) and the averaged length (Lgth) of the maximum likelihood intervals (ml) and the fixed- $k$  intervals (fk) for the tail index. Data are generated from the Pareto(0.5), the absolute value of Student-t(2), the F(4,4), and the dPIN distributions with the censored probability (cen\_p) being 1% or 0.01%. The results are based on 1000 simulations. The level of significance is 5%.

Table 4 depicts the coverage probabilities and lengths of the confidence intervals of the 99% quantile, using either the maximum likelihood method as in Proposition 2 or the fixed- $k$  method (17). Both methods deliver satisfactory size and length properties, although the maximum likelihood intervals suffer from slight undercoverage. However, as we target the more extreme 99.9% quantile as in Table 5, such undercoverage is substantial when  $k$  is less than 250. In contrast, the fixed- $k$  ones always perform excellently. These results reinforce our switching cutoff at  $k = 250$ .

Table 4: Small Sample Properties of Inference about the 0.99 Quantile

cen_p	1%				0:1%			
	Cov		Lgth		Cov		Lgth	
n=1000	ml	fk	ml	fk	ml	fk	ml	fk
GPD	0.95	0.94	6.71	7.09	0.91	0.96	4.83	5.79
t(2)	0.95	0.94	6.73	7.13	0.91	0.94	4.87	5.89
F(4,4)	0.94	0.94	11.67	12.17	0.91	0.95	8.32	9.91
dPIN	0.95	0.94	4.92	5.25	0.91	0.95	3.54	4.37
n=2000	ml	fk	ml	fk	ml	fk	ml	fk
GPD	0.96	0.94	4.60	4.86	0.92	0.96	3.38	3.88
t(2)	0.96	0.94	4.60	4.76	0.93	0.96	3.43	3.89
F(4,4)	0.96	0.95	8.04	8.09	0.92	0.95	5.85	6.72
dPIN	0.96	0.94	3.44	3.50	0.93	0.96	2.52	2.90
n=5000	ml	fk	ml	fk	ml	fk	ml	fk
GPD	0.97	0.96	2.85	2.93	0.93	0.94	2.12	2.38
t(2)	0.97	0.95	2.84	2.91	0.94	0.96	2.15	2.40
F(4,4)	0.97	0.95	4.93	5.05	0.93	0.94	3.68	4.10
dPIN	0.97	0.95	2.08	2.12	0.93	0.96	1.57	1.77

Note: Entries are the coverage probabilities (Cov) and the averaged length (Lgth) of the maximum likelihood intervals (ml) and the fixed- $k$  intervals (fk) for the 99% quantiles. Data are generated from the Pareto(0.5), the absolute value of Student-t(2), the F(4,4), and the dPIN distributions with the censored probability (cen\_p) being 1% or 0.01%. The results are based on 1000 simulations. The level of significance is 5%.

Table 5: Small Sample Properties of Inference about the 0.999 Quantile

cen_p	1%				0.1%			
	Cov		Lgth		Cov		Lgth	
n=1000	ml	fk	ml	fk	ml	fk	ml	fk
GPD	0.86	0.92	128.9	102.6	0.83	0.95	61.68	75.08
t(2)	0.85	0.93	123.0	103.1	0.83	0.94	60.87	72.62
F(4,4)	0.85	0.91	226.8	178.1	0.83	0.95	106.3	130.8
dPIN	0.85	0.93	93.16	78.21	0.82	0.95	45.16	55.13
n=2000	ml	fk	ml	fk	ml	fk	ml	fk
GPD	0.89	0.92	79.20	72.25	0.87	0.94	39.71	44.91
t(2)	0.88	0.94	75.60	71.65	0.88	0.95	39.32	45.96
F(4,4)	0.89	0.92	136.6	126.5	0.88	0.93	69.35	80.28
dPIN	0.88	0.92	56.17	51.48	0.87	0.93	28.99	32.38
n=5000	ml	fk	ml	fk	ml	fk	ml	fk
GPD	0.94	0.95	43.06	40.41	0.92	0.94	23.96	23.96

5(1)11(7)-990(5-2022(m)17(l)-2

## 5.1 US Individual Earnings

Our first application is about the tail features of the individual earnings distribution. Following the convention, we use the variable ERN\_VAL in the March CPS dataset and drop the individuals that are younger than 18 or older than 70 years old. This yields 115,424 observations in the 2019 sample. The censoring threshold is 310000 USD, which leads to a 0.58% censoring fraction in the full sample and various censoring fractions in different subsamples. The first several columns in Table 6 present the sample sizes ( $n$ ) and the numbers (cen#) and the fractions (cen%) of the censored observations, respectively. We use the previously introduced method to construct the 95% confidence intervals of the tail index and the 99% and 99.9% quantiles. Specifically, we follow the simulation study to use the maximum likelihood confidence intervals developed in Section 2 when  $k$  is larger than 250 and switch to the fixed- $k$

6 present the results with  $k = [0: n]$ . The results based on other choices are similar and reported in Appendix A.3.

Several interesting findings can be summarized as follows. First, in Panel A, the tail index is around 0.5 in the full sample, as commonly found in the existing literature. But it is substantially different across subsamples. Second, the tail also exhibits substantial heterogeneity across genders. In particular, the male sample has significantly higher quantiles than the female at both the 99% and 99.9% levels. Third, this difference also exists across races. In particular, the 99.9% quantile of all males is at least twice larger than that of the black males. All such heterogeneity provides new evidence for potential racial and gender discrimination. Finally, Panel B depicts the heterogeneity across ages, with substantially heavier tails showing up in the middle-aged groups.

Table 6: Empirical Results in 2019 March CPS Data

Panel A: 95% confidence intervals in race-and-gender-based subsamples									
race-gender	<i>n</i>	cen#	cen%	tail index		Q(0.99)		Q(0.999)	
full sample	115424	672	0.58	(0.41	0.52)	(24.24	25.32)	(62.63	75.61)
all males	55553	491	0.88	(0.35	0.53)	(28.64	30.68)	(67.71	92.58)
all females	59871	181	0.30	(0.42	0.55)	(18.02	19.03)	(46.34	58.01)
white males	43371	419	0.97	(0.88	1.00)	(29.83	33.56)	(145.2	279.0)
white females	45424	141	0.31	(0.42	0.57)	(18.09	19.28)	(46.56	60.66)
Asian males	3676	50	1.36	(0.00	0.45)	(30.73	37.20)	(48.03	86.76)
Asian females	4099	22	0.54	(0.35	0.94)	(20.77	27.16)	(45.12	145.0)
Hispanic males	44420	445	1.00	(0.71	0.95)	(31.10	34.75)	(119.3	214.0)
Hispanic females	48192	155	0.32	(0.47	0.62)	(18.70	19.90)	(50.17	66.13)
black males	6144	12	0.20	(0.22	0.58)	(15.92	18.22)	(28.73	49.39)
black females	7827	9	0.16	(0.16	0.44)	(13.90	15.64)	(25.25	37.18)

  

Panel B: 95% confidence intervals in age-based subsamples									
age	<i>n</i>	cen#	cen%	tail index		Q(0.99)		Q(0.999)	
18-30	27829	35	0.13	(0.33	0.49)	(12.69	13.60)	(28.16	36.16)
30-40	25213	158	0.63	(0.28	0.50)	(24.12	26.36)	(52.24	75.73)
40-50	23419	213	0.91	(0.83	1.00)	(28.89	33.82)	(119.7	297.2)
50-60	21767	196	0.90	(0.52	0.83)	(28.19	32.21)	(78.05	154.2)
60-65	17196	70	0.41	(0.17	0.41)	(20.95	23.13)	(41.86	59.31)

Note: Entries are the sample size (*n*), the number of censored observations (cen#), the censored fraction in percentage points (cen%), 95% confidence intervals of the tail index and those of the 99% and 99.9% quantiles measured in  $10^4$  USD. The results are based on the variable ERN\_VAL in the CPS dataset and equivalently the variable inlongj from the IPUMS dataset. Data are available at <https://cps.ipums.org/cps>.

## 5.2 Macroeconomic Disasters

This section studies the size distribution of macroeconomic disasters, which is an important research topic in macroeconomics. Barro and Ursúa (2008) and Barro and Jin (2011) construct and analyze the dataset that consists of annual GDP (and consumption) growth rates in 36 countries from 1870 to 2005. The authors sort these observations and define a macroeconomic disaster if the GDP declines by more than 10%. This leads to  $k = 157$  tail

observations. Then the authors fit these data to the (double) Pareto distribution to estimate the Pareto exponent, which is the reciprocal of the tail index, and back out the coefficient of the relative risk aversion by a theoretical model (eq.2 in Barro and Jin (2011)).

However, the largest disasters tend to be missing because some governments collapsed or were fighting wars (p.1581 in Barro and Jin (2011)). Ignoring these missing data in the upper tail could lead to substantial bias, as we show in the Monte Carlo simulations. We revisit this problem by applying our fixed- $k$  method since  $k$  is only moderate. Specifically, the most recent data missing happens in four countries, which are Greece, Malaysia, the Philippines, and Singapore during WWII. Therefore, we set  $m = 4$  and apply the fixed- $k$  method to construct the 95% confidence intervals for the tail index and those for the coefficient of relative risk aversion by solving eq.2 in Barro and Jin (2011). For comparison, we also construct the intervals based on Hill (1975)'s estimator and the bias-reduced estimator (GI) proposed by Gabaix and Ibragimov (2011). Table 7 presents the result.

As shown in the table, the fixed- $k$  intervals contain substantially larger values of the tail



## 6 Concluding Remarks

This paper develops a new approach to estimate and conduct inference about tail features for censored data. The method can be viewed as a hybrid approach that uses the maximum likelihood estimation when the tail sample size is large and switches to a small sample modification otherwise. As shown in Monte Carlo simulations, the new method has excellent small sample performance.

This new approach is empirically relevant in broad areas studying tail features (e.g., tail index and extreme quantiles). We illustrate this with the March CPS data and the macroeconomic disaster data and find considerably different results from the existing literature.

There are theoretical extensions and empirical applications of our method, which we suppress in the current paper due to space limitations. We list a few here. First, our method naturally applies to the no censoring case by setting  $\alpha = 1$  in the MLE and  $m = 0$  in the fixed- $k$  method. Besides, we can follow Müller and Wang (2019) to construct the (quantile) unbiased estimation of the tail features, which could perform better in terms of mean absolute deviation and mean squared error, especially when  $k$  is not large.

Second, many other tail features can be learned by our new method as long as they can be expressed as functions of the tail index. For example, the conditional tail expectation is another important risk measure in finance, which is defined as the expectation conditional on being larger than some high quantile, that is,  $E[X|X > F^{-1}(k)]$

# Appendix

## A.1 Computational Details

The estimators defined in Section 3 require evaluation of some expectations. Define  $\Gamma(a; x) = \int_0^x t^{a-1} e^{-t} dt$  as the incomplete Gamma function and  $e(x^*; \mathbf{s}) = \exp\left(-\sum_{i=1}^k \log(1 + x_i^* s)\right)$ . Change of variables and integration by parts yield that

$$E [X_{m+1} | X_{m+k} = x^*] f_{X|Y}(x^*) = \frac{(k+m)!}{m!} \int_0^{\infty} s^{k-1} \exp\left(-\sum_{i=1}^k \log(1 + x_i^* s)\right) ds$$

and

$$f_{Y|X}(y; x^*) = \frac{1}{m!} \int_0^{\infty} \left(1 + \frac{x_i^*}{y} s\right)^{-m} ds > 0 \text{ for } T.3520T7.ITJ/F2211.9552Tf32.4.3J/F2211.95525TJ/F5311.9$$

The tables of the Lagrange multipliers and the corresponding MATLAB code are provided on our website: <https://sites.google.com/site/yulongwanghome/>.

## A.2 Proof

To prove Proposition 1, we first establish two intermediate results, Lemmas 1 and 2.

$$\begin{aligned}
& + (1 - D_i) \left( \frac{1}{2} \log \frac{1 - Y_i}{1 + Y_i} + \frac{1}{1 + Y_i} - \frac{1}{1 - Y_i} \right); \\
\frac{\partial I_i(\cdot)}{\partial} = & D_i \frac{1}{z^{-1}} \\
& + (1 - D_i) \left( \frac{1}{1 + Y_i} + \frac{1}{1 - Y_i} \right);
\end{aligned}$$

$$\begin{aligned}
& \frac{\partial^2 I_i(\cdot)}{\partial^2} \\
= & D_i \left[ \frac{2}{3} \log z - \frac{T_u^2}{2} z^{-2} - \frac{2T_u}{2} \right]
\end{aligned}$$

$$E_{GPD} \left[ 1 + \frac{Y_i^{-r}}{1 - [Y_i - T] Y_i > u} \right] = \frac{1 - z^{-r-1}}{1 + r} \text{ for any } r > 0 \quad (20)$$

$$E_{GPD} \log \left[ 1 + \frac{Y_i}{1 - [Y_i - T] Y_i > u} \right] = z^{-1} (1 + \log z) \quad (21)$$

Then using (19)-(21) to obtain that

$$E_{GPD} \frac{\partial^2 I_i(\cdot; \cdot)}{\partial^2} \Big|_{T_i > u}$$





where recall  $(u) = u^{-\alpha}$ . Use the change of variable  $z = 1 + \frac{T-u}{u}$  and recall  $\alpha = 1 - \beta$ . Then by Lemma 1, we have  $z = 1 + \frac{T-u}{u} = T-u$  and

$$\begin{aligned}
 & E \frac{dI_i(\cdot)}{d\alpha} \Big|_{Y_i > u} \\
 &= E [D_{ij} Y_i > u] \left( \frac{1}{2} \log z + \frac{1}{2} \frac{1}{z} z^{-1} \right) \\
 &+ E \left( \frac{1}{2} \log \left( 1 + \frac{Y_i}{u} \right) + \frac{1}{2} \frac{1}{1 + \frac{Y_i}{u}} \right) \mathbb{1}[Y_i > T] \Big|_{Y_i > u} \\
 &= (1 - F_u(T-u)) \left( \frac{1}{2} \log(T-u) + \frac{1}{2} \frac{1}{(T-u)^{-1}} \right) \\
 &+ \frac{1}{2} E \left( \log \left( 1 + \frac{Y_i}{u} \right) + \frac{1}{1 + \frac{Y_i}{u}} \right) \mathbb{1}[Y_i > T] \Big|_{Y_i > u} \\
 &+ E \left( \frac{1}{2} \frac{1}{1 + \frac{Y_i}{u}} \right) \mathbb{1}[Y_i > T] \Big|_{Y_i > u} \\
 &= (T-u)^{-\alpha} \left( \frac{1}{2} \log(T-u) + \frac{1}{2} \frac{1}{(T-u)^{-1}} \right) + \frac{1}{2} \frac{1}{(T-u)^{-\alpha}} \\
 &+ (1 + \frac{1}{u}) \frac{1}{2} \frac{1}{(T-u)^{-\alpha}} \frac{1}{(T-u)^{-1}} + O(u) \\
 &= O(u):
 \end{aligned}$$



To this end, it suffices to show that  $E[j^3 I_i(\cdot; \cdot) = \mathbf{u} | Y_i > \mathbf{u}]$ ,  $E[j^2 I_i(\cdot; \cdot) = \mathbf{u} | Y_i > \mathbf{u}]$ ,  $E[j^3 I_i(\cdot; \cdot) = \mathbf{u} | Y_i > \mathbf{u}]$ , and  $E[j^2 I_i(\cdot; \cdot) = \mathbf{u} | Y_i > \mathbf{u}]$  are all uniformly bounded over this neighborhood. This is done by straightforward calculations as we show in Lemma 3. For brevity, we present

have

$$\begin{aligned}
 & E \left[ 1 + \frac{Y_i}{v} \right]^{-r} \mathbb{1}[Y_i \leq T] \mathbb{1}[Y_i > u] \\
 &= E \left[ 1 + \frac{Y_i}{vu} \right]^{-r} \mathbb{1}[Y_i \leq T] \mathbb{1}[Y_i > u] \\
 &= \int_0^{T-u} \left( 1 + \frac{y}{vu} \right)^{-r} d(1 - F_u(y)) \\
 &= \left( 1 + \frac{y}{vu} \right)^{-r} (1 - F_u(y)) \Big|_0^{T-u} + vr \int_0^{T-u} Q_r \oplus
 \end{aligned}$$

We next derive the limits of  $C_{jn}$  for  $j = 1; 2; 3; 4$ . To this end, we define  $U(t) = Q(1 - 1/t)$  and denote  $U'(t) = @U(t) = @t$ . We introduce the second-order tail approximation that

$$\lim_{t \rightarrow \infty} \frac{\frac{U(yt) - U(t)}{a(t)} \frac{y - 1}{A(t)}}{A(t)} = H(y) \quad (26)$$

Proof of Proposition 3 We prove this by induction. By standard EVT, for any fixed positive integer  $I$ ,

$$f_{x_1, \dots, x_I} (x_1; \dots; x_I) = \prod_{i=1}^I v(x_i) = V(x_i); \quad (27)$$

Consider  $m = 1$  first. For any fixed positive integer  $k$ , (27) with  $I = k + 1$  implies that

$$\begin{aligned} f_{x_{m+1}, \dots, x_{m+k}} (x_{m+1}; \dots; x_{m+k}) &= f_{x_2, \dots, x_{k+1}} (x_2; \dots; x_{k+1}) \\ &= \int_{x_2}^{\infty} \frac{v(x_1)}{V(x_1)} dx_1 \prod_{i=2}^{k+1} v(x_i) = V(x_i) \\ &= \log V(x_2) \prod_{i=2}^{k+1} v(x_i) = V(x_i); \end{aligned}$$

which satisfies (9).

Now assume (9) holds for some fixed positive integer  $m \geq 1$ . This implies that for any  $k$ ,

$$\begin{aligned} & \int_{-\infty}^{\infty} f_{x_{m+2}, \dots, x_{m+1+k}} (x_{m+2}; \dots; x_{m+1+k}) \\ &= \int_{-\infty}^{\infty} f_{x_{m+1}, \dots, x_{m+k+1}} (x_{m+1}; \dots; x_{m+k+1}) dx_{m+1} \\ &= \int_{x_{m+2}}^{\infty} \frac{1}{m!} (\log V(x_{m+1}))^m \frac{v(x_{m+1})}{V(x_{m+1})} dx_{m+1} \prod_{i=m+2}^{m+k+1} v(x_i) = V(x_i) \\ &= \int_{\log G(x_{m+2})}^{\infty} (v)^m dv \frac{1}{m!} V(x_{m+k+1}) \prod_{i=m+2}^{m+k+1} v(x_i) = V(x_i) \\ &= \frac{1}{(m+1)!} (\log V(x_{m+2}))^{m+1} \prod_{i=m+2}^{m+k+1} v(x_i) = V(x_i); \end{aligned}$$

which means that (9) holds for  $m + 1$

Table 8: Empirical Results Using 2019 March CPS Data

Panel A: 95% confidence intervals with race-based subsample									
	<i>n</i>	cen#	cen%	tail index		Q(0.99)		Q(0.999)	
Full Sample	115424	672	0.58	(0.37	0.49)	(24.26	25.32)	(59.88	73.00)
Male	55553	491	0.88	(0.39	0.62)	(28.76	30.97)	(71.86	106.3)
Female	59871	181	0.30	(0.30	0.44)	(18.31	19.33)	(42.46	52.03)
Male White	43371	419	0.97	(0.09	0.34)	(29.57	31.84)	(54.54	75.61)
Female White	45424	141	0.31	(0.29	0.45)	(18.40	19.59)	(42.36	53.68)
Male Asian	3676	50	1.36	(0.00	0.55)	(30.73	37.77)	(48.30	109.6)
Female Asian	4099	22	0.54	(0.13	0.76)	(21.27	26.66)	(42.89	104.5)
Male Hispanic	44420	445	1.00	(0.11	0.36)	(29.98	32.27)	(55.56	77.78)
Female Hispanic	48192	155	0.32	(0.38	0.55)	(18.83	19.99)	(45.77	59.36)
Male Black	6144	12	0.20	(0.16	0.53)	(16.06	18.55)	(29.70	47.84)
Female Black	7827	9	0.16	(0.12	0.44)	(13.94	15.72)	(25.08	36.81)

  

Panel B: 95% confidence intervals with age-based subsample									
Age	<i>n</i>	cen#	cen%	tail index		Q(0.99)		Q(0.999)	
18-30	27829	35	0.13	(0.34	0.53)	(12.63	13.54)	(28.42	37.19)
30-40	25213	158	0.63	(0.24	0.51)	(24.15	26.40)	(50.57	75.56)
40-50	23419	213	0.91	(0.00	0.32)	(28.67	31.48)	(48.07	71.41)
50-60	21767	196	0.90	(0.63	1.00)	(28.30	32.89)	(86.35	218.9)
60-65	17196	70	0.41	(0.43	0.76)	(20.29	22.63)	(49.71	88.34)

Note: Entries are the sample size (*n*), the number of censored observations (cen#), the censored fraction in percentage points (cen%), 95% confidence intervals of the tail index and those of the 99% and 99.9% quantiles measured in 10<sup>4</sup> USD. The results are based on  $k = [0.04n]$  and the variable ERN\_VAL in the CPS dataset (and equivalently the variable inclongj from the IPUMS dataset). Data are available at <https://cps.ipums.org/cps>.

Table 9: Empirical Results Using 2019 March CPS Data

Panel A: 95% confidence intervals in race-based subsamples									
	<i>n</i>	cen#	cen%	tail index		Q(0.99)		Q(0.999)	
Full Sample	115424	672	0.58	(0.31	0.40)	(24.33	25.34)	(56.01	65.09)
Male	55553	491	0.88	(0.30	0.45)	(28.59	30.52)	(63.88	82.90)
Female	59871	181	0.30	(0.43	0.54)	(18.03	19.05)	(46.57	57.68)

## References

Aban, I. B., M. M. Meerschaert, and

Chernozhukov, V., and I. Fernández-Val (2011): "Inference for extremal conditional quantile models, with an application to market and birthweight Risks," *The Review of Economic Studies*, 78, 559–589.

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009): "Power-law Distributions in Empirical Data," *SIAM Review*, 51(4), 661–703.

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

Drees, H.



- 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.
- Leadbetter, M. R. (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.
- Mikosch, T., and C. Staric (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 (2017): "Fixed-k Asymptotic Inference about Tail Properties," *the Journal of the American Statistical Association*, 112, 1134–1143.
- (2019): "Nearly Weighted Risk Minimal Unbiased Estimation," *Journal of Econometrics*, 209, 18–34.
- O'Brien, G. (1987): "Extreme Values for Stationary and Markov Sequences," *Annals of Probability*, 15(1), 281–291.
- Pickands, III, J. (1975): "Statistical inference using extreme order statistics," *Annals of Statistics*, 3(1), 119–131.
- Portnoy, S. (2003): "Censored Regression Quantiles," *Journal of the American Statistical Association*, 98(464), 1001–1012.
- Powell, J. L. (1986): "Censored Regression Quantiles," *Journal of Econometrics*, 32(1), 143–155.
- Reed, W. (2003): "The Pareto Law of Incomes-an Explanation and an Extension," *Physica A*, 319(1), 469–486.
- Smith, R. L. (1987): "Estimating Tails of Probability Distributions," *Annals of Statistics*, 15, 1174–1207.
- Toda, A. A. (2012): "The Double Power Law in Income Distribution: Explanations and Evidence," *Journal of Economic Behavior and Organization*, 84(1), 364–381.

——— (2017): "A Note on the Size Distribution of Consumption: More Double Pareto than Lognormal," *Macroeconomic Dynamics*