Electronic Journal of Statistics
Vol. 17 (2023) 3728–3761
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2191

Estimation of tail parameters with missing largest observations

Jan Beirlant

Department of Mathematics, KU Leuven, Belgium
Department of Statistics and Actuarial Science, University of the Free State, South Africa
e-mail: jan.beirlant@kuleuven.be

Martin Bladt

Department of Mathematical Sciences, University of Copenhagen, Denmark
e-mail: martinbladt@math.ku.dk

Gao Maribe

Department of Statistics, University of Pretoria, South Africa
e-mail: gao.isc@gmail.com

Andrehette Verster

Department of Statistics and Actuarial Science, University of the Free State, South Africa
e-mail: VersterA@ufs.ac.za

Abstract: The setting where an unknown number m of the largest data is missing from an underlying Pareto-type distribution is considered. Solutions are provided for estimating the extreme value index, the number of missing data and extreme quantiles. Asymptotic results of the parameter estimators and an adaptive selection method for the number of top data used in the estimation are proposed for the case where all missing data are beyond the observed data. An estimator of the number of missing extremes spread over the largest observed data is also proposed. To this purpose, a key component is a likelihood solution based on exponential representations of spacings between the largest observations. An effective and fast optimization procedure is established using regularization, and simulation experiments are provided. The methodology is illustrated with a dataset from the diamond mining industry, where large-carat diamonds are expected to be missing.

Keywords and phrases: Extreme value index, high quantiles, missing observations, regularization.

Received April 2023.

Contents

1 Introduction 3729
2 Estimation of γ and the number of missing extremes under the Pareto-type model 3730
2.1 Missing data situated above the largest observation 3730

2.2 Asymptotic results 3733
2.3 Detecting missing extremes below the largest observation . . . 3735
2.4 Regularized fixed-point optimization 3736
2.5 Sample fraction selection 3738
3 Simulation results and diamond case study 3739
3.1 Simulation results 3739
3.2 Diamond case study 3741
4 Conclusion 3743
A Proof of Theorems 3744
B Simulation plots 3748
References 3760

1. Introduction

Extreme value methodology receives growing attention in order to model the occurrence of rare events with high impact in various fields of application. Estimation of the extreme value index (EVI) is then a crucial topic in extreme value methodology, assuming that the underlying distribution satisfies the max-domain of attraction condition, i.e. assuming that the maximum of independent and identically distributed observations $X_1, X_2, \ldots, X_n$ can be approximated by the generalized extreme value distribution: as $n \to \infty$

$$\mathbb{P}\Big(a_n^{-1}\Big(\max_{i=1,\ldots,n} X_i - b_n\Big) \le y\Big) \quad \to \quad G_\gamma(y) = \exp\big(-(1+\gamma y)^{-1/\gamma}\big), \tag{1.1}$$

for $1 + \gamma\, y > 0$, where $b_n \in \mathbb{R}$, $a_n > 0$ and $\gamma \in \mathbb{R}$ are the location, scale and shape parameters, respectively. The EVI γ is a measure of the tail-heaviness of the distribution of X with a larger value of γ implying a heavier tail of F.

We consider the specific case of Pareto-type distributions with a positive EVI, which induces a restriction to right tail functions (RTF) given by

$$\bar{F}(x) = 1 - F(x) = P(X > x) = x^{-\frac{1}{\gamma}}\,\ell(x) \tag{1.2}$$

with $\gamma > 0$ and ℓ a slowly varying function at infinity, that is

$$\lim_{t\to\infty} \frac{\ell(ty)}{\ell(t)} = 1 \quad \text{for every } y > 1. \tag{1.3}$$

We then assume that from the original ordered data set $X_{1,N} \le X_{2,N} \le \ldots \le X_{N,N}$ an unknown number of top data $X_{N-m+1,N} \le X_{N-m+2,N} \le \ldots \le X_{N,N}$ are missing. So we observe $n = N - m$ observations.

The estimation of m, $\gamma > 0$ and extreme quantiles $Q(1-p)$ with small p based on $X_{1,N} \le X_{2,N} \le \ldots \le X_{n,N}$ under (1.2) is then the main problem tackled in this paper. Recent discussions of this problem are given in [11] and [10] based on a Hill Estimator without Extremes (HEWE) process. Similar to their technique, we adopt a likelihood-based approach. However, the method is conceptually different since our starting point is, visually, the Pareto QQ-plot, and then mathematically the log-spacings and their Rényi representation, which allows for transparent, automatic selection of the sample fraction used in the estimation.

Specifically, it is well known that when $\gamma > 0$ the EVI can be estimated from the slope at an ultimate linear part of the Pareto QQ-plot:

$$\left(\log \frac{n+1}{j}, \log X_{n-j+1,N} \right), \ j = 1, \dots, n,$$

(see for instance Chapter 4 in Beirlant et al. 2004). The celebrated [8] estimator of γ can then be considered as an estimator of the slope of this QQ-plot using the top k available data:

$$H_{k,n} = \frac{1}{k} \sum_{j=1}^{k} \log X_{n-j+1,N} - \log X_{n-k,N},$$

for some appropriate $k \in \{1, \dots, n\}$. In this paper we propose an adaptation of the Hill estimator for the case that missing data are or could be present at the top data.

A motivating practical example can be found in [9] from the diamond mining industry. The nature of metallurgical recovery processes in diamond mining causes the under-recovery of large diamonds. Because of the potentially large monetary value of even a small number of large diamonds, the number of diamonds that are not recovered is an important problem in the diamond mining industry. We illustrate the proposed methods with the same sample of carat sizes as used in [9]. Note that here the missing carat data are, though likely, not necessarily all larger than the largest observed carat size.

The remainder of the paper is organized as follows. In Section 2 we propose and analyze the estimation of (γ, m) and extreme quantiles. We propose asymptotic results for the estimators and an adaptive method for selecting k in case the missing observations are situated above the largest observation. A graphical method is given to detect the number of missing data in case some observations are also missing below the largest non-missing observation. Finally, we provide an efficient way of numerically optimizing a regularized version of the likelihood through a contraction operator. In Section 3 we present and discuss simulation results and revisit the diamond data set. Section 4 concludes.

2. Estimation of γ and the number of missing extremes under the Pareto-type model

2.1. Missing data situated above the largest observation

To describe the influence of deleted data from a graphical point of view, in Figures 1 we illustrate the influence of missing top data on Pareto QQ-plots of strict Pareto (with $\ell(x) = 1$, $x > 1$) and Fréchet samples (with $\bar{F}(x) = \exp(-x^{-1/\gamma})$, $x > 1$) both with $\gamma = 0.5$. We consider $N = 200$ and $m = 20$

in both cases. From the Pareto QQ-plots (see label '$m = 0$' in Figure 1) we observe that the introduction of 20 missing data leads to concavity. This leads to underestimating extreme quantiles when using the classical linear extrapolation methods from extreme value analysis. An additional complication arises in the case of a non-constant slowly varying function ℓ, since then the missingness implies non-linearity in the top portion of the data, and hence the linear part in the QQ-plot is shortened. In case of slow convergence in (1.3) and/or significant missingness, linearity could be completely lost, and tail estimation then becomes problematic.

Thus, we propose to correct the QQ-plot by adjusting the inverse ranks j to $j + \hat{m}$ in the construction of the QQ-plot with some appropriate value $\hat{m}$, leading to a Pareto QQ-plot adapted for missingness:

$$\left(\log \frac{n + \hat{m} + 1}{j + \hat{m}}, \log X_{n-j+1,N}\right),\ j = 1, \ldots, n. \tag{2.1}$$

In Figure 1 such corrected QQ-plots are given using different values of $\hat{m}$.

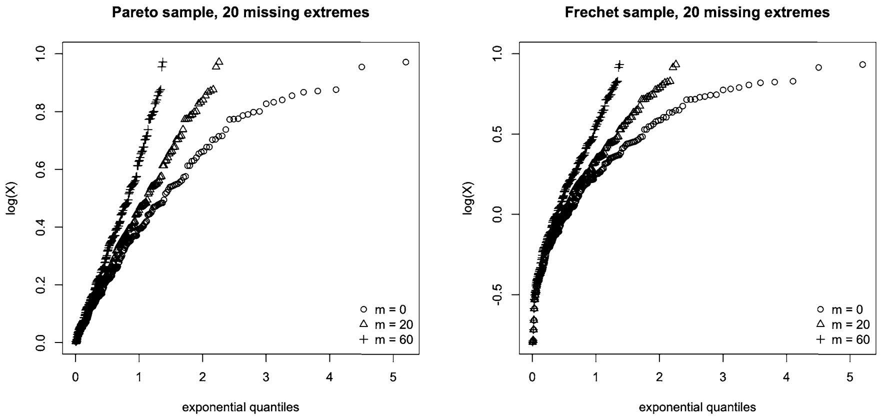

FIG 1. *QQ-plots (2.1) for a sample of size $N = 200$ and $\gamma = 0.5$ from the strict Pareto (left) and Fréchet (right) distributions with $m = 20$ missing data and $\hat{m}$ chosen as $0, 20, 60$.*

When $\hat{m} < 20$ the concavity in the top of the QQ-plot remains to some extent, while linearity is obtained at $\hat{m} = 20$ (at least for upper quantiles), and convex corrected QQ-plots appear when $\hat{m} > 20$. The possibility of estimating m and γ when adjusting the rank numbers is the key idea behind the graphical motivation.

More precisely, when the observed (non-missing) maximum $X_{n,N}$ is situated below the smallest missing data point $X_{N-m+1,N}$ we can exploit the exponential representations of scaled log-spacings as discussed in [2] and [6] which generalize the Rényi representation to Pareto-type distributions: under (1.2) with $V_{j,n} := \log \frac{X_{n-j+1,N}}{X_{n-j,N}}$ we have that

$$(j + m)V_{j,n} = (j + m) \log \frac{X_{N-j-m+1,N}}{X_{N-j-m,N}},\ j = 1, \ldots, k$$

for small enough k can be well approximated by γE_j with $\{E_j,\ j = 1, \ldots, k\}$ independent standard exponential random variables. Hence the pseudo-likelihood based on a set $\{V_{j,n},\ j = 1, \ldots, k\}$ equals

$$L(\gamma, m) = \prod_{j=1}^{k} \frac{j+m}{\gamma} \exp\left(-\frac{j+m}{\gamma} V_{j,n}\right),$$

leading to the log-likelihood function

$$l(\gamma, m) = -k \log \gamma + \sum_{j=1}^{k} \log(j+m) - \gamma^{-1} \sum_{j=1}^{k} j V_{j,n} - \frac{m}{\gamma} \log \frac{X_{n,N}}{X_{n-k,N}},$$

where we use the fact that $\sum_{j=1}^{k} V_{j,n} = \log \frac{X_{n,N}}{X_{n-k,N}}$. In the derivative of l with respect to m we next approximate $\sum_{j=1}^{k}(j+m)^{-1} = \sum_{i=m+1}^{k+m} i^{-1}$ by $\log \frac{m+k}{m}$, which is motivated by Euler's formula when $m, k \to \infty$. The maximum likelihood estimators $(\hat{\gamma}_k, \hat{m}_k)$ are then given by

$$\begin{cases} \hat{\gamma}_k = \frac{1}{k} \sum_{j=1}^{k} j V_{j,n} + \frac{\hat{m}_k}{k} \log \frac{X_{n,N}}{X_{n-k,N}} = H_{k,n} + \frac{\hat{m}_k}{k} \log \frac{X_{n,N}}{X_{n-k,N}}, \\ \frac{\hat{m}_k}{\hat{m}_k + k} = \left(\frac{X_{n,N}}{X_{n-k,N}}\right)^{-\frac{1}{\hat{\gamma}_k}}. \end{cases} \tag{2.2}$$

Note that $\hat{\gamma}_k$ is a simple adaptation of the Hill (1975) estimator $H_{k,n}$ which is induced by the missing observations. In case of a strict Pareto distribution, $X^{1/\gamma}$ is standard Pareto distributed, so that $(\frac{X_{n,N}}{X_{n-k,N}})^{-1/\gamma} =_d U_{m,m+k}$, the m-th smallest order statistics from a uniform (0,1) random sample of size $m+k$. The estimator of m given an estimate of γ hence tries to match the expected value $\frac{m}{m+k}$ of $U_{m,m+k}$.

For any given choice of k, an estimator for extreme quantiles $Q(1-p)$ can now be presented based on estimators of γ and m using the classical Weissman-type extrapolation on the corrected QQ-plot:

$$\log \hat{Q}_k(1-p) = \log X_{n-k,N} + \hat{\gamma}_k \left(\log \frac{1}{p} - \log \frac{\hat{m}_k + n}{\hat{m}_k + k}\right)$$

or

$$\hat{Q}_k(1-p) = X_{n-k,N} \left(\frac{\hat{m}_k + k}{(\hat{m}_k + n)p}\right)^{\hat{\gamma}_k}. \tag{2.3}$$

One of the main practical drawbacks of the methodology in [10] is the extremely sensitive log-likelihood function in terms of $(\gamma, m/k)$. Although the asymptotic theory is sound, the estimator has to be calibrated sequentially and with a good initial guess to produce accurate results. Moreover, the complicated definition of the estimators therein makes it difficult to mathematically analyse these step-wise procedures, and was not pursued in their paper.

In our specification, the likelihood function still empirically suffers the sensitivity issue, especially for small values of k, though to a much lesser degree. To further robustify our estimator we impose a regularization term on the number of missing observations (we refrain from also penalizing the tail index since the two quantities are heavily correlated anyway):

$$l_\lambda(\gamma, m) = l(\gamma, m) - \lambda m.$$

The maximum penalized likelihood estimators are denoted by $(\hat\gamma_k^{(\lambda)}, \hat m_k^{(\lambda)})$, where one observes that

$$\hat m_k^{(\lambda)} = H_{k,n} + \frac{\hat m_k^{(\lambda)}}{k} \log \frac{X_{n,N}}{X_{n-k,N}}, \tag{2.4}$$

$$\hat m_k^{(\lambda)} = k\left\{ e^\lambda \left(\frac{X_{n,N}}{X_{n-k,N}}\right)^{\frac{1}{\hat\gamma_k^{(\lambda)}}} - 1\right\}^{-1}. \tag{2.5}$$

2.2. Asymptotic results

For a mathematical analysis of the estimators $(\hat m_k^{(\lambda)}, \hat\gamma_k^{(\lambda)})$ we consider two cases. First we consider $m/k \to 0$ as $m, k, N \to \infty$, which refers to cases with small m values. Next, we assume that $m/k \to \delta \in (0,1)$ as $k, N \to \infty$ and $k/N \to 0$. Proofs are deferred to Appendix A.

To this end we assume the classical second-order assumption (refer to [7]) on $U(x) = Q(1-x^{-1})$ with Q the quantile function of the underlying Pareto-type distribution:

$$\frac{U(ux)}{U(x)} = u^\gamma\big(1 + h_{-\beta}(u) b(x)\big(1+o(1)\big)\big), \tag{2.6}$$

with $h_{-\beta}(u) = (1-u^{-\beta})/\beta$, $\beta > 0$ and b a regularly varying function at infinity with index $-\beta$. Then from Theorem 4.1 in [3] we have the exponential representations

$$M_{i,k,N} := \left(\gamma + b_{N,k}\left(\frac{i}{k+1}\right)^\beta\right) E_i$$

for $Z_{i,N} = i \log \frac{X_{N-i+1,N}}{X_{N-i,N}}$ with $i = 1, \dots, k$, where $E_i, i \geq 1$ is a sequence of i.i.d. standard exponential random variables and $b_{N,k} = b(N/k)$. More precisely,

$$\sup_{1\leq j\leq k} \big| Z_{j,N} - (M_{j,k,N} + R_{j,N})\big| = o_p(b_{N,k})$$

where $\sup_{1\leq i\leq k} |\sum_{j=i}^k R_{j,N}/j| / \max(\log \frac{k+1}{i}, 1) = o_p(b_{N,k})$.

When the number of missing observations m is limited, expressed by the assumption $m/k \to 0$ as $k, N \to \infty$ and $N/k \to \infty$, any conventional estimator $\hat\gamma_k$ of γ could well be adequate as a substitute for $\hat\gamma_k^{(\lambda)}$ in (2.5). This is confirmed in a first asymptotic result. In this $\zeta_{k,n}$ denotes the bias and Z the zero centered asymptotic normal distribution of $\hat\gamma_k$.

Theorem 1. *Assume the second-order condition (2.6). Then as N, $k \to \infty$, k/N, m/k, $\lambda \to 0$, we have, when imputing an estimator $\hat\gamma_k$ of γ in (2.5) satisfying*

$$\hat\gamma_k - \gamma = \zeta_{k,n} + k^{-1/2} Z_k$$

where $\zeta_{k,n}$ is a deterministic bias sequence satisfying $\sqrt{k}\zeta_{k,n} = O(1)$ and $Z_k := \sqrt{k}(\hat{\gamma}_k - \gamma - \zeta_{k,n})$ is a sequence of rv's converging weakly to a centered normal distribution, that

$$\hat{m}_k^{(\lambda)} =_d \Gamma_m \times \Big\{ 1 - \big[\lambda + b_{N,k}(\gamma\beta)^{-1}\big]\big(1 + o_p(1)\big) \\ + \big[\zeta_{k,n} + k^{-1/2} Z_k\big]\gamma^{-1} \log \frac{k+m}{m}\big(1 + o_p(1)\big)\Big\},$$

where the rv Γ_m has density

$$f_{k,m}(v) = \frac{\Gamma(k+m+1)}{\Gamma(m)\Gamma(k+1)k^m} v^{m-1}\left(1 + \frac{v}{k}\right)^{-(k+2m)}, \; v > 0.$$

Assuming further that $m^2/k \to 0$ we have that

$$f_{k,m}(v) = \frac{1}{\Gamma(m)} v^{m-1} e^{-v}\big(1 + o(1)\big), \; v > 0.$$

From the above result it follows that the penalization parameter λ can be used to reduce the bias in estimating m induced by the bias of $\hat{\gamma}_{k,n}$ especially for small values of k. A positive penalization parameter then works for estimators $\hat{\gamma}_{k,n}$ with a positive bias $\zeta_{k,n}$. Estimation of $b_{N,k}$, β and $\zeta_{k,n}$ in the present setting is far from straightforward. In the practical realizations below we choose $\lambda = 0.01$, as a first attempt to reduce the bias. In the simulations we will consider the finite sample behaviour of $\hat{m}_k^{(\lambda)}$ when using different well established estimators of γ next to the estimator $\hat{\gamma}_k^{(\lambda)}$ following from optimizing $l_\lambda(\gamma, m)$.

From Theorem 1 it appears appropriate to use the $\Gamma(m,1)$ distribution as the sampling distribution for $\hat{m}_k$ at fixed k. The goodness-of-fit of this model will be discussed in the simulation study.

Next we consider the case $m/k \to \delta \in (0,1)$ as $k, N \to \infty$ and $k/N \to 0$. We propose an asymptotic result for the estimators $(\hat{\gamma}_k, \hat{\delta}_k)$ solving (2.2) in case the missing data were all deleted at the top of the original data set. We here take $\lambda = 0$. Below we use the notation $\zeta(\delta, a) = ((1+\delta)^a - \delta^a)/a$, $a > 0$.

Theorem 2. *Assume the second-order condition (2.6). Then as $N, k \to \infty$, $k/N \to 0$, $m/k \to \delta > 0$ and $\sqrt{k} b_{N,k} \to \nu \geq 0$ we have that*

$$\sqrt{k}\big((\hat{\gamma}_k, \hat{\delta}_k) - (\gamma, \delta)\big) \to \mathcal{N}_2(\nu A, \Sigma),$$

with

$$A = \frac{\delta(1+\delta)}{\gamma[1 - \delta(1+\delta)\log^2(1+\delta^{-1}]} \begin{pmatrix} \frac{\gamma}{\delta(1+\delta)} & -\gamma\log(1+\delta^{-1}) \\ \log(1+\delta^{-1}) & -1 \end{pmatrix} \begin{pmatrix} \zeta_{\delta,1+\beta} \\ \zeta_{\delta,\beta} \end{pmatrix}$$

and

$$\Sigma = \big[1 - \delta(1+\delta)\log^2\big(1+\delta^{-1}\big)\big]^{-1} \\ \times \begin{pmatrix} \gamma^2 & \gamma\delta(1+\delta)\log(1+\delta^{-1}) \\ \gamma\delta(1+\delta)\log(1+\delta^{-1}) & \delta(1+\delta) \end{pmatrix}.$$

Note that when $\delta \to 0$ the result for $\hat{\gamma}_k$ naturally corresponds with the classical limit result for the Hill estimator when no data are missing, i.e. a normal limit distribution with asymptotic variance γ^2.

2.3. Detecting missing extremes below the largest observation

So far we have made the assumption that all the missing data is above the largest observation. However, if the data has missing values in large quantiles but not necessarily above the largest observation, we may adapt our methodology to this scenario. In essence, the idea is to sequentially keep removing the largest datapoint from the sample until we reduce to the canonical scenario, that is where all missing observations are above the largest datapoint in the sample. We provide details below.

Given an appropriate value of k, the search for missing extreme values that are situated within the top extreme data (and not necessarily above the largest observation) can be performed by trimming the likelihood from the preceding section until we obtain stabilization of the estimators. Trimming has shown to provide stability in other settings (cf. [4] for an application in outlier detection and [5] for general threshold selection).

Thus, we consider the deletion of further extreme points, with the rationale of eventually excluding all regions of the datasets where missing observations were present. If the m missing top data are all located above $X_{n-k_0,N}$ with $k_0 < k$, then we still have that the spacings $V_{j,n}$ with $j = k_0+1,\dots,k$ are approximately exponentially distributed with mean γ if scaled with the inverse rank numbers $j+m,\ j = k_0+1,\dots,k$. Then, the amount m can be estimated through maximization of

$$L(\gamma, m; k_0) = \prod_{j=k_0+1}^{k} \frac{j+m}{\gamma} \exp\left(-\frac{j+m}{\gamma} V_{j,n}\right),$$

leading to the trimmed log-likelihood function

$$\begin{aligned} l(\gamma, m; k_0) = -(k-k_0)\log\gamma + \sum_{j=k_0+1}^{k} \log(j+m) - \gamma^{-1} \sum_{j=k_0+1}^{k} jV_{j,n} \\ - \frac{m}{\gamma} \log \frac{X_{n-k_0,N}}{X_{n-k,N}}. \end{aligned}$$

As before, approximating $\sum_{j=k_0+1}^{k}(j+m)^{-1}$ by $\log(\frac{m+k}{m+k_0})$, we obtain the following likelihood equations for the maximum likelihood estimators $(\hat{\gamma}_{k_0,k}, \hat{m}_{k_0,k})$:

$$\begin{cases} \hat{\gamma}_{k_0,k} = \frac{1}{(k-k_0)} \sum_{j=k_0+1}^{k} jV_{j,n} + \frac{\hat{m}_{k_0,k}}{k-k_0} \log \frac{X_{n-k_0,N}}{X_{n-k,N}} \\ \log(\frac{\hat{m}_{k_0,k}+k}{\hat{m}_{k_0,k}+k_0}) = \frac{1}{\hat{\gamma}_{k_0,k}} \log \frac{X_{n-k_0,N}}{X_{n-k,N}}. \end{cases} \tag{2.7}$$

Direct optimization of the above likelihood is possible and effective through standard nonlinear numerical procedures. However, we refer the reader to the next subsection for a fast fixed-point iterative solver. Note that $(\hat{\gamma}_{0,k}, \hat{m}_{0,k}) = (\hat{\gamma}_k, \hat{m}_k)$.

In practice, one first selects k large enough so that one believes that all missing datapoints are within the top k observations. Subsequently, one plots $(\hat{\gamma}_{k_0,k}, \hat{m}_{k_0,k})$ as a function of k_0 and with k fixed. One then expects to find a stable region in the plot for a certain value of k_0 and above. The point from which stability happens then provides an estimate k_0 of additional top order statistics required to be deleted from the sample in order to reduce the dataset to the canonical case, i.e. where data is missing above the largest sample observation.

As an example, in Figure 2, such a plot for one sample is given for a sample of size $N = 500$ from a Pareto distribution with $\gamma = 0.5$ where 25 missing observations are all randomly spread in the top 50 of the original data set. We also present the MSE and bias of $(\hat{\gamma}_{k_0,k}, \hat{m}_{k_0,k})$ based on 1000 repetitions. Since all missing observations are situated above $X_{n-25,N}$, we expect the plots $(\hat{\gamma}_{k_0,k}, \hat{m}_{k_0,k})$ to be stable in the region $k_0 \geq 25$ with $\hat{m}_{k_0,k}$ indicating the total number of missing observations, while for $k_0 < 25$ these plots will be decreasing with decreasing k_0 as the number of missing observations above such $X_{n-k_0,N}$ is decreasing with smaller k_0. The estimates $\hat{\gamma}_{k_0,k}$ appear to be reliable for $k_0 \geq 25$. Compare this with a case where the 25 missing data are all situated above $X_{n,N}$ (bottom line of Figure 2). Here the plot is stable over the whole plotted area.

The present approach using likelihood trimming only provides a graphical method assisting in detecting missing observations below the largest observation $X_{n,N}$, and asymptotics are still absent. It appears that methods for detection of change points in the $\hat{\gamma}_{k_0,n-1}$ and $\hat{m}_{k_0,n-1}$ plots could be used to provide an adaptive method to detect a minimal k_0 value above which the missing observations are situated. This will be pursued in subsequent work. Note that formula (2.3) can still be used in this case in order to estimate extreme quantiles.

2.4. Regularized fixed-point optimization

This subsection is devoted to further robustifying our estimator in two ways. First, we impose a regularization term on the number of missing observations (we refrain from also penalizing the tail index since the two quantities are heavily correlated anyway). Subsequently, we prove that we may compute the (possibly) penalized MLE estimator recursively by defining a suitable contraction operator and then invoking Banach's fixed-point theorem. The latter is key to iteratively computing our estimators without having to have good initial guesses.

To this end, consider the penalized log-likelihood given by

$$l_\lambda(\gamma, m; k_0) = \\ -(k-k_0)\log\gamma + \sum_{j=k_0+1}^{k} \log(j+m) - \gamma^{-1} \sum_{j=k_0+1}^{k} jV_{j,n} - \frac{m}{\gamma}\log\frac{X_{n-k_0,N}}{X_{n-k,N}} - \lambda m. \tag{2.8}$$

It is not hard to see that the approximate solution to the score equations

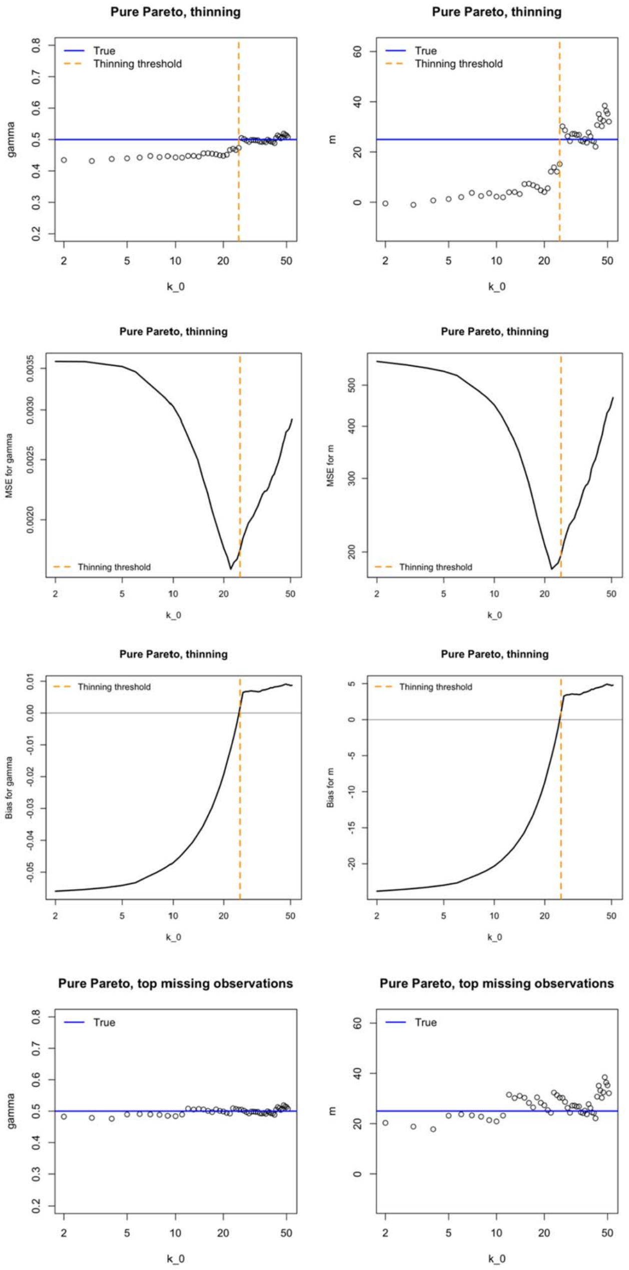

FIG 2. *Simulations results for Pareto data* ($\gamma = 1/2, N = 500, m = 25$). *Top: plots of* $\hat{\gamma}_{k_0,n-1}$ *and* $\hat{m}_{k_0,n-1}$ *using* $k = n - 1$ *as a function of* $\log k_0$ *for one sample when 25 data are randomly deleted from the top 50 original observations. Middle two lines: plot of MSE and bias of* $\hat{\gamma}_{k_0,n-1}$ *and* $\hat{m}_{k_0,n-1}$ *as a function of* $\log k_0$ *with randomly spread missing observations based on* 1000 *repetitions. Bottom: plot of* $\hat{\gamma}_{k_0,n-1}$ *and* $\hat{m}_{k_0,n-1}$ *as a function of* $\log k_0$ *when the original top 25 data are deleted.*

satisfies the following equations

$$\hat{m}_{k_0,k} = \frac{k X_{n-k,N}^{\frac{1}{\hat{\gamma}_{k_0,k}}} - e^{\lambda} k_0 X_{n-k_0,N}^{\frac{1}{\hat{\gamma}_{k_0,k}}}}{e^{\lambda} X_{n-k_0,N}^{\frac{1}{\hat{\gamma}_{k_0,k}}} - X_{n-k,N}^{\frac{1}{\hat{\gamma}_{k_0,k}}}}$$

$$\hat{\gamma}_{k_0,k} = \frac{1}{(k-k_0)} \sum_{j=k_0+1}^{k} j V_{j,n} + \frac{\hat{m}_{k_0,k}}{k-k_0} \log \frac{X_{n-k_0,N}}{X_{n-k,N}}.$$

This gives rise to a recursive perturbed estimator for the tail index, given by

$$\hat{\gamma}^{(\lambda)}_{k_0,k}(r+1) = \frac{1}{(k-k_0)} \sum_{j=k_0+1}^{k} j V_{j,n} + \frac{\log \frac{X_{n-k_0,N}}{X_{n-k,N}}}{k-k_0} \frac{k X_{n-k,N}^{\frac{1}{\hat{\gamma}^{(\lambda)}_{k_0,k}(r)}} - e^{\lambda} k_0 X_{n-k_0,N}^{\frac{1}{\hat{\gamma}^{(\lambda)}_{k_0,k}(r)}}}{e^{\lambda} X_{n-k_0,N}^{\frac{1}{\hat{\gamma}^{(\lambda)}_{k_0,k}(r)}} - X_{n-k,N}^{\frac{1}{\hat{\gamma}^{(\lambda)}_{k_0,k}(r)}}}. \tag{2.9}$$

The corresponding estimator $\hat{m}^{(\lambda)}_{k_0,k}(r)$ for the number of missing observations is defined analogously.

It is clear that whenever (2.8) has a unique maximum then proving that (2.9) is a contraction operator will suffice, by Banach's fixed-point theorem, to obtain convergence of $\hat{\gamma}^{(\lambda)}_{k_0,k}(r)$ as $r \to \infty$ to the maximizer of (2.8).

Proposition 3. *Let $\lambda \geq 0$. Then the recursive map $\hat{\gamma}^{(\lambda)}_{k_0,k}(r)$ is c-Lipschitz continuous on any compact set bounded away from zero, with $c < 1$.*

Proposition 4. *Let $\lambda \geq 0$ and let $(\hat{\gamma}^{(\lambda)}_{k_0,k}, \hat{m}^{(\lambda)}_{k_0,k})$ be the penalized maximum likelihood estimator of (2.8), with $\hat{\gamma}^{(\lambda)}_{k_0,k} > 0$. Then for any positive starting value $\hat{\gamma}^{(\lambda)}_{k_0,k}(0) > 0$,*

$$\lim_{r \to \infty} \left(\hat{\gamma}^{(\lambda)}_{k_0,k}(r), \hat{m}^{(\lambda)}_{k_0,k}(r)\right) = \left(\hat{\gamma}^{(\lambda)}_{k_0,k}, \hat{m}^{(\lambda)}_{k_0,k}\right)$$

Proof. Any c-Lipschitz continuous function with $c \in (0,1)$ is a contraction. Then apply Banach's fixed point theorem. □

2.5. Sample fraction selection

In order to select an appropriate value $\tilde{k}$ of the number k of top data several methods can be used based on goodness-of-fit techniques. For instance:

i) For every k the correlation coefficient r_k can be computed based on the top k points of the adjusted QQ-plot (2.1) with $\hat{m}_k$ substituting $\hat{m}$, and $\tilde{k}_r$ then corresponds to the largest correlation r_k;

ii) For every k the Anderson-Darling (A-D) W-statistic, given by (cf. [1])

$$W_k^2 = -k - \frac{1}{k} \sum_{j=1}^{k} (2j-1) \left[\log u_{j,k} + \log(1 - u_{k-j+1,k})\right]$$

is computed on

$$\left\{u_j = 1 - \exp\left(-(j + \hat{m}_k)V_{j,n}/\hat{\gamma}_k\right);\ j = 1, \ldots, k\right\}$$

and $\tilde{k}_W$ then corresponds to the smallest W-statistic W_k.

From the simulations and practical experiments the use of $\tilde{k}_W$ appears to yield much better results in cases different from the strict Pareto distribution. While such adaptive method does not guarantee a consistent estimator for the asymptotic MSE optimal k value, it does address the finite-sample case appropriately, as it appears from simulation studies. However, it is advised to further validate the choice of k using graphical support from the Pareto QQ-plot adjusted with $\hat{m}_{k_W}$, for instance, by validating linearity in this plot above $X_{n-\tilde{k}_W,N}$.

3. Simulation results and diamond case study

3.1. Simulation results

We conduct a systematic simulation study to investigate the finite-sample behaviour of our estimators for varying k, as well as the A-D approach for the automatic k selection when all missing observations are situated beyond the largest observation. For varying sample fraction we consider several estimators, as follows:

1. The estimator $(\hat{\gamma}_k^{(0.01)}, \hat{m}_k^{(0.01)})$.
2. The estimator $(\hat{\gamma}_k^{(0)}, \hat{m}_k^{(0)})$.
3. The Hill estimator $H_{k,n}$ and an implied number of missing observations given by plugging it into the second equation of (2.2), yielding a 'naive' estimator for the number of missing observations: $k\{(\frac{X_{n,N}}{X_{n-k,N}})^{\frac{1}{H_{k,n}}} - 1\}^{-1}$.
4. The moment estimator for the EVI, and its implied estimator for the number of missing observations through the same construct as the previous case.
5. The Generalized Pareto Distribution (GPD) maximum-likelihood estimator for the EVI, and its implied estimator for the number of missing observations through the same construct as the previous case.

We have also compared with the estimator from [10] as written in their paper, but their likelihood is very sensitive, for instance with respect to the starting value. A stepwise procedure for their estimators could improve the performance, though we refrain from implementing this. Note that the last two estimators often provide negative tail indices, so that automatic selection formulae derived through the Anderson-Darling approach from Subsection 2.5 becomes unstable. Thus, for automatic selection of k we only consider the following three cases:

1a. The estimator $(\hat{\gamma}_k^{(0.01)}, \hat{m}_k^{(0.01)})$ with k selected through the goodness-of-fit criterion with Anderson-Darling statistic outlined in Subsection 2.5.

2a. The estimator $(\hat{\gamma}_k^{(0)}, \hat{m}_k^{(0)})$ with k selected through the goodness-of-fit criterion with Anderson-Darling statistic outlined in Subsection 2.5.
3a. The Hill estimator $H_{k,n}$ with k chosen from the selection of the case 1a.

In order to assess how deviations from pure the Pareto distribution affect the estimation procedure, we consider the following distributions with regularly-varying tails:

i) The Pareto distribution with RTF given by $\bar{F}(x) = x^{-2}$, $x \geq 1$.
ii) The Burr$(2, -2, 2)$ distribution with RTF given by $\bar{F}(x) = ((2+x^4)/2)^{-1/2}$, $x > 0$.
iii) The Fréchet(2) distribution with RTF given by $\bar{F}(x) = 1 - \exp(-(x^{-2}))$, $x > 0$.
iv) The GPD$(1/2, 1)$ distribution with RTF given by $\bar{F}(x) = (1 + x/2)^{-2}$, $x > 0$.

Notice that we have specified $\gamma = 1/2$ in all cases, which allows for deviations in behaviour to be solely attributed to the slowly varying component ℓ of the distributions. For all distributions we consider a sample size of 500, $m = 5, 25, 50$, and compare mean (median when applicable) squared error and bias terms for the estimation of the tail index, for the number of missing observations, and for a high quantile ($p = 1/500$). All results are provided in Appendix B.

Remark 5. We use the iterative procedure from Subsection 2.4, which by Proposition 4 converges for any positive starting value. For convenience, we choose 1/2 as starting value, agreeing with the tail index of the simulations, though any other starting value provides indistinguishable results. Convergence usually happens within a few iterations, but all results were obtained using 100 iterations.

The main conclusions from the figures are the following:

- Our estimators 1, 1a, 2 and 2a behave better at most k than the benchmarks when the true distribution is close to being strictly Pareto, and the automatic selection procedure is effective regardless of the distribution type. The regularization terms arising from using $\lambda = 0.01$ is useful for fixed k but seems to play a less important role when automatically selecting k.
- When the distribution has a significant deviation from strict Pareto tails, the automatic estimators 1a and 2a also outperform the 'naive' estimator 3a, except in the $m = 5$ case when considering the tail index, where there is a slight underestimation by the estimator 3a, and a slight overestimation by 1a and 2a. For the number of missing observations and high quantiles, this effect seems to be less pronounced. Note in case $m = 5$ for all distributions that estimator 3 has a smaller MSE than estimators 1 and 2.
- Estimators allowing negative tail indices, 4 and 5, can only be considered adequate when the number of missing observations is small ($m = 5$ in our study, or 1%), and then in that case, using a very large k. This is

particularly the case for the GPD distribution, iv). Notice, however, that these estimators do not provide a natural estimate of the number of missing observations, so that their effectiveness relies solely on $m \ll n$.[1] This behaviour is extended to the Hill estimator, 3, which in its automatic form 3a can be competitive for tail index estimation.

- Not accounting for missing values *decreases* the value of estimator 3, but not removing bias from the regularly varying component (by choosing k large enough) *increases* estimator 3. These two very different sources of bias often cancel each other out, providing "by chance" a good estimate. This is observed as a sharp decrease in MSE for medium or large k values. However, estimating such a high k where the two biases cancel out is neither straightforward nor in the scope of this article. As expected, the effect vanishes when considering estimators with automatically selected k.

We end this section checking the validity of the Gamma$(m, 1)$ distribution which follows from Theorem 1 as a sampling distribution for $\hat{m}_k$ in case of a small and a moderate m. The goodness-of-fit is illustrated in Figure 17 in case of the Pareto distribution, based on $10,000$ repetitions. When using $\hat{\gamma}_k^{(0)}$ a positive bias with respect to the Gamma model is present with small m. When inserting the Hill estimator $H_{k,n}$ in place of $\hat{\gamma}_k$ in the second equation of (2.2), the distribution of the estimates of m for very small values of m follows the $\Gamma(m, 1)$ distribution quite close and shows a negative bias in case of moderate m.

3.2. Diamond case study

The problem of estimating the amount of missing diamonds in ore mining was considered before in [9]. They used a Bayesian approach to fit a truncated generalized Pareto distribution to part of the data. Based on the estimated tail probability the expected number of diamonds larger than a specific weight was estimated. Ore recovered from alluvial deposits are less subjected to the possibility of breakage. If the stones are not recovered during the metallurgical recovery process, they are discarded onto the tailing dumps from where they can be recovered during a re-mining program. Because of the potentially large monetary value of even a small amount of missing large diamonds, it is of interest to analyze if re-mining of a diamond dump is profitable thanks to the presence of large diamonds. The Pareto-type model is generally accepted to describe and analyze carat data.

Here we use the same data set as in [9]. The Pareto QQ-plot of these observations is given in the bottom left plot of Figure 3. A concave deflection at the upper part of the QQ-plot appears comparable with the graphs in Figure 1 obtained from simulations. This leads to a systematic decline of the Hill estimates with decreasing k below 100 (see Figure 3 bottom right, before correction).

[1] In some figures, the curves corresponding to estimators 4 and 5 are completely out of range due to very negative bias (and thus very large MSE) and related numerical instability.

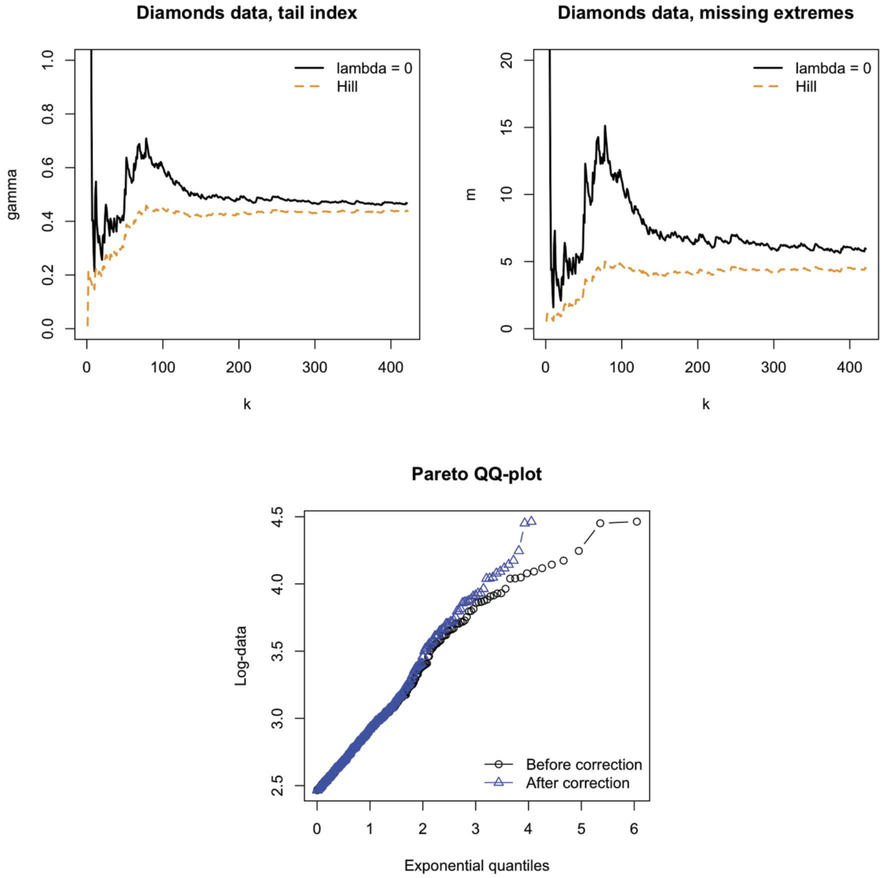

FIG 3. *Diamonds data. Top left: adapted Hill and Hill estimates, $\hat{\gamma}_k$ and $H_{k,n}$ respectively, as a function of k. Top right: $\hat{m}_k$ and the plug-in missing observation estimator derived from inserting the Hill estimator into the second equation of* (2.2)*, as a function of k. Bottom: original Pareto QQ-plot and adapted QQ-plot using $\hat{m}_{273}$.*

We provide plots of $\hat{m}_k$ and $\hat{\gamma}_k$ as a function of k in the top panels of Figure 3 and of the W_k statistic in Figure 4. We obtain

$$\tilde{k}_W = 273, \quad \hat{\gamma}_{\tilde{k}_W} = 0.4798, \quad \hat{m}_{\tilde{k}_W} = 6.4739$$

The corresponding parameter estimates for $\lambda = 0.01$ are

$$\tilde{k}_W^{\lambda} = 273, \quad \hat{\gamma}_{\tilde{k}_W}^{\lambda} = 0.4792, \quad \hat{m}_{\tilde{k}_W}^{\lambda} = 6.3740.$$

The plots are virtually indistinguishable between $\lambda = 0$ and $\lambda = 0.01$, and hence we present only the former. The plug-in missing observation estimator derived from inserting the Hill estimator into the second equation of (2.2) yields a very stable plot as a function of k leading to an estimate of 5 missing observations.

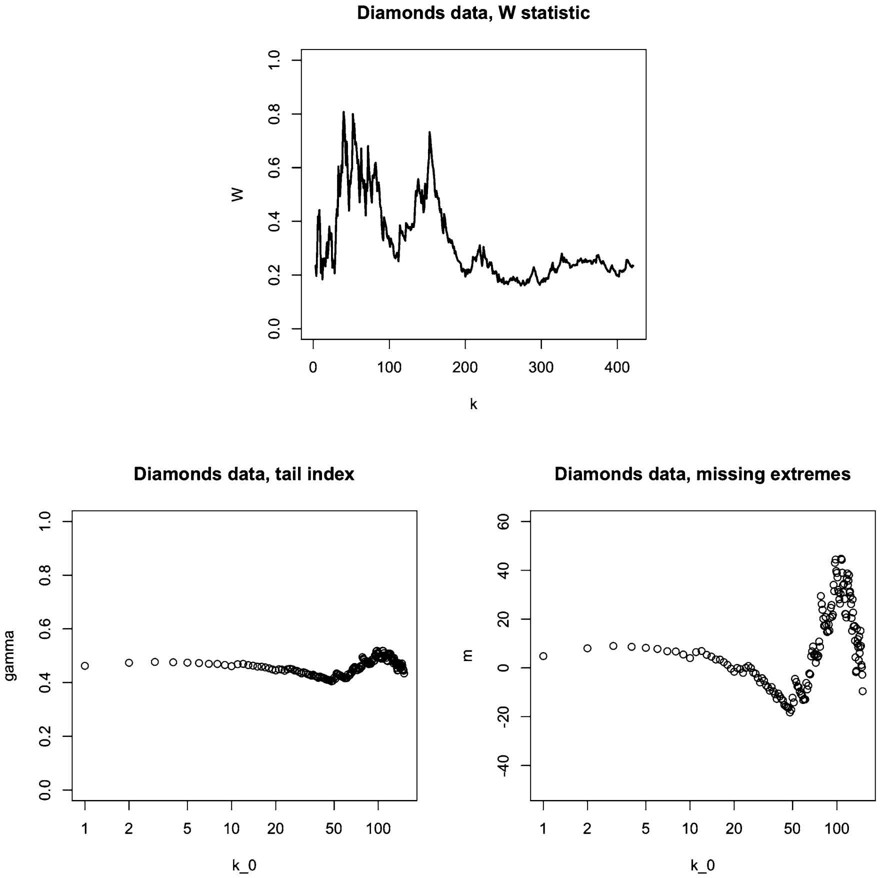

FIG 4. *Diamonds data. Top: W_k statistics as a function of k. Bottom left: estimates $\hat{\gamma}_{k_0,273}$ as a function of k_0. Bottom right: estimates $\hat{m}_{k_0,273}$ as a function of k_0.*

The bottom panels of Figure 3 shows the difference of Pareto QQ plots before and after correcting with 6 missing extremes. In Figure 4 the solutions $(\hat{m}_{k_0,273}, \hat{\gamma}_{k_0,273})$ of the trimmed likelihood with $0 \leq k_0 < 273$ are plotted as a function of k_0. From this it appears that no extra missing observations can be reported apart from 6 missing observations at the top. This is to be compared with the 8 missing observations reported by the method from [9]. Based on Theorem 1, a 95% confidence interval for m is obtained using the 0.025 and 0.975 quantiles of the $\Gamma(6.45, 1)$ distribution: $(2.5, 12.3)$. Based on the Hill imputed missings estimator we obtain the interval $(1.6, 10.2)$.

4. Conclusion

In this paper, we addressed the problem of missing observations in the highest quantiles of a dataset, assuming that the data followed a Pareto-type distribution. We presented solutions for estimating the extreme value index, the number of missing data and extreme quantiles, assuming that all missing data were beyond the observed data. We also proposed an adaptive method for selecting the number of top data used in the estimation. Additionally, we introduced a graphical method in order to infer on the number of missing extremes spread over the largest observed data. We derived asymptotic results and considered robustifying our estimator through regularization. We demonstrated the effectiveness of our approach through simulation experiments and an application in the diamond mining industry.

Appendix A: Proof of Theorems

Proof of Theorem 1. With $U_{j,v}$ $(j = 1, \ldots, v)$ denoting the order statistics of an i.i.d. sample of size v from the uniform (0,1) distribution, we have using (2.6) that

$$\begin{aligned}\frac{X_{n,N}}{X_{n-k,N}} &=_d \left(\frac{U_{k+m,N}}{U_{m,N}}\right)^{\gamma}\left(1+b\big(U_{k+m,N}^{-1}\big)h_{-\beta}\left(\frac{U_{k+m,N}}{U_{m,N}}\right)\big(1+o_p(1)\big)\right)\\ &=_d (U_{m,k+m})^{-\gamma}\left(1+b\left(\frac{N}{k+m}\right)h_{-\beta}\left(\frac{m+k}{m}\right)\big(1+o_p(1)\big)\right)\\ &= O_p\left(\log\frac{k}{m}\right),\end{aligned}$$

so that

$$\left(\frac{X_{n,N}}{X_{n-k,N}}\right)^{-\frac{1}{\gamma}} = U_{m,m+k}\left(1-\frac{1}{\gamma}b\left(\frac{N}{k+m}\right)h_{-\beta}\left(\frac{m+k}{m}\right)\big(1+o_p(1)\big)\right).$$

Using $\frac{1}{\hat{\gamma}_k}-\frac{1}{\gamma} = -(\hat{\gamma}_k-\gamma)/(\gamma\hat{\gamma}_k) = -[\zeta_{k,n}+k^{-1/2}Z_k(1+o_p(1))]\gamma^{-2}(1+o_p(1))$ and $\max(\zeta_{k,n},k^{-1/2})\log(k/m)\to 0$ we obtain

$$\left(\frac{X_{n,N}}{X_{n-k,N}}\right)^{\frac{1}{\hat{\gamma}_k}-\frac{1}{\gamma}} = \exp(-\epsilon_{k,n}) = 1-\epsilon_{k,n}\big(1+o_p(1)\big)$$

with

$$\epsilon_{k,n} = -\big[\zeta_{k,n}+k^{-1/2}Z_k\big(1+o_p(1)\big)\big]\gamma^{-2}\log(k/m)\big(1+o_p(1)\big).$$

Now, as $\lambda\to 0$,

$$\begin{aligned}\frac{k}{e^{\lambda}(\frac{X_{n,N}}{X_{n-k,N}})^{\frac{1}{\hat{\gamma}_k}}-1} &= \frac{k}{(\frac{X_{n,N}}{X_{n-k,N}})^{\frac{1}{\gamma}}(1+\lambda(1+o(1))+\epsilon_{k,n})-1}\\ &= \frac{k}{(\frac{X_{n,N}}{X_{n-k,N}})^{\frac{1}{\gamma}}-1}\end{aligned}$$

$$\times \left\{ 1 + \left[-\lambda\big(1+o(1)\big) + \epsilon_{k,n} \right] \left(1 - \left(\frac{X_{n,N}}{X_{n-k,N}} \right)^{-\frac{1}{\gamma}} \right)^{-1} \right\}.$$

Concerning the distribution of $k\{(\frac{X_{n,N}}{X_{n-k,N}})^{\frac{1}{\gamma}} - 1\}^{-1}$ we have

$$\begin{aligned} &k\left\{ \left(\frac{X_{n,N}}{X_{n-k,N}} \right)^{\frac{1}{\gamma}} - 1 \right\}^{-1} \\ &\quad =_d \frac{kU_{m,m+k}}{1-U_{m,m+k}} \\ &\quad \times \left(1 - \frac{1}{\gamma} b\left(\frac{N}{k+m} \right) h_{-\beta}\left(\frac{m+k}{m} \right) \left(1 + \frac{m}{k} \right) \big(1+o_p(1)\big) \right), \end{aligned}$$

while $(1 - (\frac{X_{n,N}}{X_{n-k,N}})^{-\frac{1}{\gamma}})^{-1} = 1 + o_p(1)$. Furthermore, the density of $R_{m,k} := \frac{kU_{m,m+k}}{1-U_{m,m+k}}$ is given by

$$f_{k,m}(v) = \frac{\Gamma(k+m+1)}{\Gamma(m)\Gamma(k+1)k^m} v^{m-1} \left(1 + \frac{v}{k} \right)^{-(k+2m)} = \frac{1}{\Gamma(m)} v^{m-1} e^{-v} \big(1+o(1)\big)$$

as $k \to \infty$, $m^2/k \to 0$ thanks to Stirling's formula. □

Proof of Theorem 2. First we define

$$\begin{aligned} S^{(1)}_{k,m,N} &= k^{-1} \sum_{j=1}^{k} (j+m) V_{j,n} = k^{-1} \sum_{j=1}^{k} (j+m) \log \frac{X_{N-(j+m)+1,N}}{X_{N-(j+m),N}} \\ S^{(2)}_{k,m,N} &= \sum_{j=1}^{k} V_{j,n} = \sum_{j=1}^{k} \log \frac{X_{N-(j+m)+1,N}}{X_{N-(j+m),N}}. \end{aligned}$$

Using Theorem 4.1 in [3] specifying the exponential representations of $Z_{i,N}$, it follows that

$$\begin{aligned} &\sqrt{k} \begin{pmatrix} S^{(1)}_{k,m,N} - \gamma - b_{N,k}\zeta_{\delta,1+\beta} \\ S^{(2)}_{k,m,N} - \gamma \log(1+\delta^{-1}) - b_{N,k}\zeta_{\delta,\beta} \end{pmatrix} \\ &\quad \to_d \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \gamma^2 \begin{pmatrix} 1 & \log(1+\delta^{-1}) \\ \log(1+\delta^{-1}) & [\delta(1+\delta)]^{-1} \end{pmatrix} \right). \end{aligned} \tag{A.1}$$

Replacing $(j+m)\log\frac{X_{N-(j+m)+1,N}}{X_{N-(j+m),N}}$ by $M_{j,k,N}$ and using classical limit theorems leads to the stated limit distributions. The term based on $R_{j,N}$ in case of $S^{(1)}_{k,m,N}$ can be handled as in the proof of Theorem 4.2 in [3]. Concerning $S^{(2)}_{k,m,N}$ this term equals $\sum_{j=1}^{k} R_{j+m,N}/(j+m) = \sum_{i=m+1}^{m+k} R_{i,N}/i$ which is $o_p(b_{N,k+m})\log(k+m)/m$.

The equation defining $\hat{\delta}_k$ is given by

$$\frac{1 - \hat{\delta}_k \log(1+(\hat{\delta}_k)^{-1})}{\log(1+(\hat{\delta}_k)^{-1})} = \frac{k^{-1}\sum_{j=1}^{k} jV_{j,n}}{S^{(2)}_{k,m,N}}$$

where the right hand side converges to $\frac{1-\delta\log(1+\delta^{-1})}{\log(1+\delta^{-1})}$, so that $\hat{\delta}_k$ is consistent and then also $\hat{\gamma}_k$.

Using that $\sum_{j=1}^{k} jV_{j,n} = S^{(1)}_{k,m,N} - \delta S^{(2)}_{k,m,N} + O(k^{-1})$ as $k \to \infty$, the likelihood equations are given by

$$\begin{cases} \hat{\gamma}_k = S^{(1)}_{k,m,N} + (\hat{\delta}_k - \delta)S^{(2)}_{k,m,N} + O(k^{-1}) \\ \hat{\gamma}_k \log(1 + (\hat{\delta}_k)^{-1}) = S^{(2)}_{k,m,N}, \end{cases}$$

or

$$\begin{cases} (\hat{\gamma}_k - \gamma) - (\hat{\delta}_k - \delta)S^{(2)}_{k,m,N} = S^{(1)}_{k,m,N} - \gamma + O(k^{-1}) \\ (\hat{\gamma}_k - \gamma)\log(1 + (\hat{\delta}_k)^{-1}) + \gamma(\log(1 + (\hat{\delta}_k)^{-1}) - \log(1 + \delta^{-1})) \\ \quad = S^{(2)}_{k,m,N} - \gamma\log(1 + \delta^{-1}). \end{cases}$$

Using the consistency of $\hat{\delta}_k$ and $\hat{\gamma}_k$ we then obtain

$$\begin{cases} (\hat{\gamma}_k - \gamma) - (\hat{\delta}_k - \delta)\gamma\log(1 + \delta^{-1}) = S^{(1)}_{k,m,N} - \gamma + O(k^{-1}) \\ (\hat{\gamma}_k - \gamma)\log(1 + \delta^{-1})(1 + o_p(1)) - (\hat{\delta}_k - \delta)\frac{\gamma}{\delta(1+\delta)}(1 + o_p(1)) \\ \quad = S^{(2)}_{k,m,N} - \gamma\log(1 + \delta^{-1}), \end{cases}$$

and so

$$\begin{pmatrix} \hat{\gamma}_k - \gamma \\ \hat{\delta}_k - \delta \end{pmatrix} \\ = \begin{pmatrix} 1 & -\gamma\log(1 + \delta^{-1}) \\ \log(1+\delta^{-1})(1+o_p(1)) & -\frac{\gamma}{\delta(1+\delta)}(1 + o_p(1)) \end{pmatrix}^{-1} \begin{pmatrix} S^{(1)}_{k,m,N} - \gamma + O(k^{-1}) \\ S^{(1)}_{k,m,N} - \gamma\log(1+\delta^{-1}) \end{pmatrix}.$$

Using the asymptotic result in (A.1) now leads to the asserted result after some algebra. □

Proof of Proposition 3. We ease the notation by defining the constants (with respect to x) $a = \frac{X_{n-k,N}}{X_{n-k_0,N}} \in (0,1)$ and $A = \frac{1}{(k-k_0)}\sum_{j=k_0+1}^{k} jV_{j,n}$. Consequently, we wish to show that the following function is c-Lipschitz continuous:

$$f(x) = A + \frac{1}{k-k_0}\log(1/a)\frac{ka^{1/x} - e^{\lambda}k_0}{e^{\lambda} - a^{1/x}}.$$

It will be enough to establish that its derivative is, uniformly and in absolute value, less than unity. For this, note first that

$$|f'(x)| = \frac{a^{1/x}e^{\lambda}\log^2(a)}{x^2(e^{\lambda} - a^{1/x})^2}.$$

Then $|f'(x)| \leq c < 1$ with $c > 0$ is equivalent to

$$\log^2(a)/c \leq x^2\big(e^{\lambda/2}a^{-1/(2x)} - e^{-\lambda/2}a^{1/(2x)}\big)^2.$$

Taking square roots and replacing the right-hand side with two Taylor expansions, we obtain, after cancelling terms, the equivalent inequality

$$-\log(a)/(x\sqrt{c}) \leq -\sum_{n=0}^{\infty} 2\frac{(\log(a)-\lambda x)^{2n+1}}{(2x)^{2n+1}(2n+1)!},$$

or

$$\begin{aligned} 1 &\leq \sqrt{c}\frac{\lambda x+\log(1/a)}{\log(1/a)}\sum_{n=0}^{\infty}\frac{(\log(a)-\lambda x)^{2n}}{(2x)^{2n}(2n+1)!} \\ &= \sqrt{c}\frac{2x}{\log(1/a)}\sinh\bigl(\bigl(\log(1/a)+\lambda x\bigr)/(2x)\bigr). \end{aligned}$$

But the above assertion is satisfied for any $\lambda \geq 0$, since in that case, using the bounded-away property of the compact K, we get

$$I := \inf_{x\in K}\frac{2x}{\log(1/a)}\sinh\bigl(\bigl(\log(1/a)+\lambda x\bigr)/(2x)\bigr) > 1,$$

and we may simply take $c = (1/I)^2 \in (0,1)$. □

Appendix B: Simulation plots

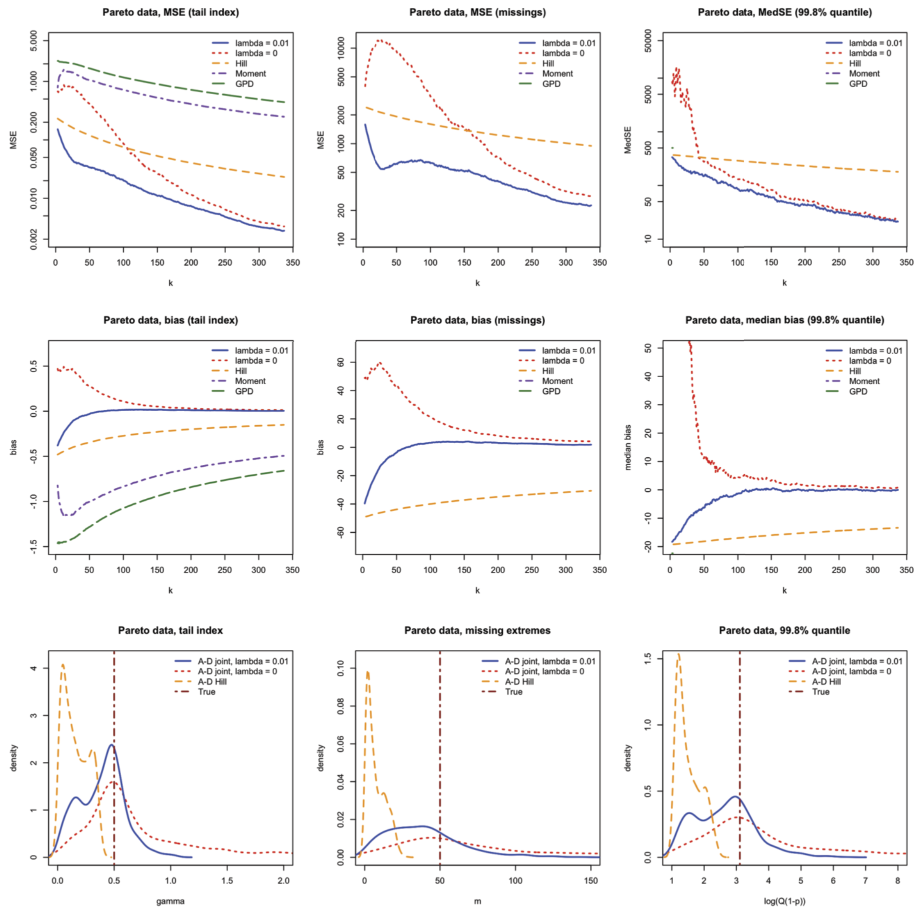

Fig 5. *Simulations results for Pareto data* ($\gamma = 1/2, m = 50$)*. Top: mean square error (MSE) for the tail index, number of missing observations, and* 99.8% *quantile, as a function of top k order statistics used. Center: Corresponding bias plots as a function of top k order statistics. Bottom: density of the estimators with automatically-selected k.*

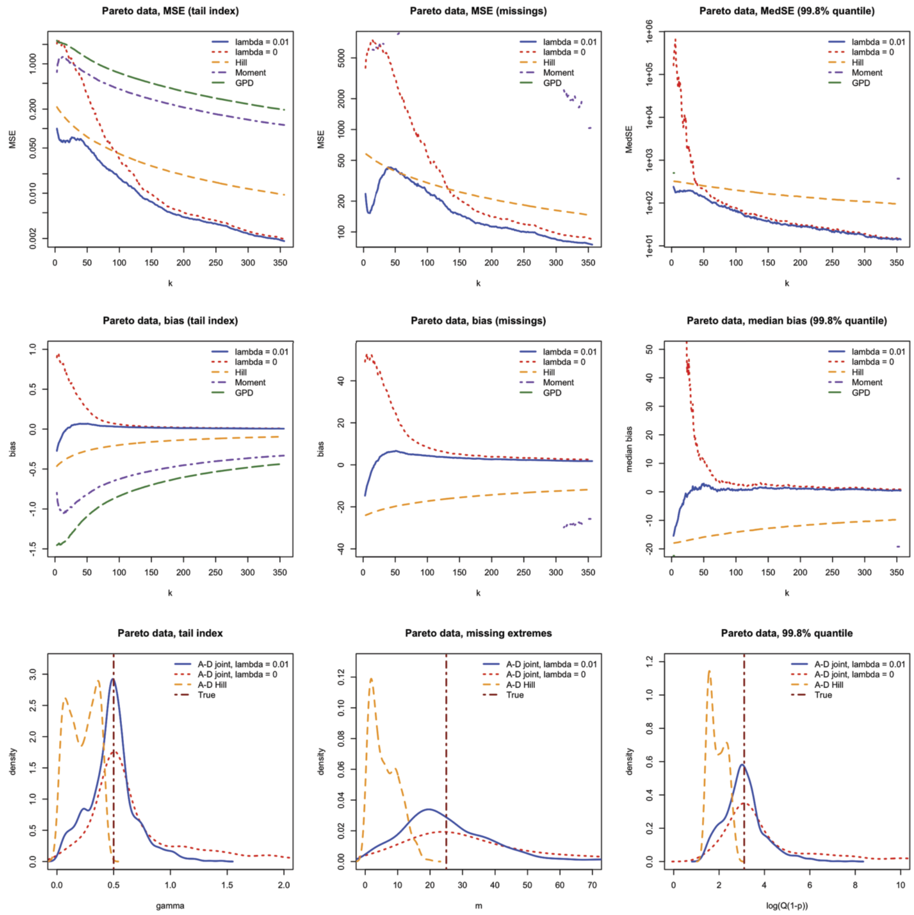

FIG 6. *Simulations results for Pareto data* ($\gamma = 1/2, m = 25$). *Top: mean square error (MSE) for the tail index, number of missing observations, and* 99.8% *quantile, as a function of top* k *order statistics used. Center: Corresponding bias plots as a function of top* k *order statistics. Bottom: density of the estimators with automatically-selected* k.

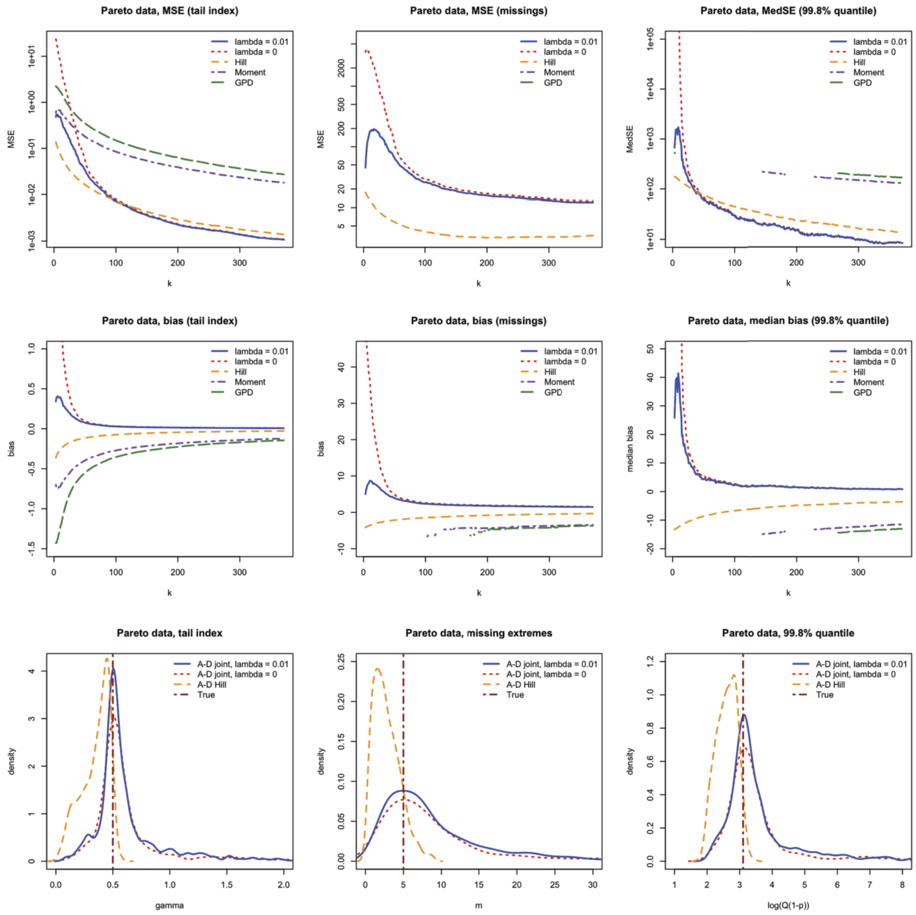

FIG 7. *Simulations results for Pareto data* ($\gamma = 1/2, m = 5$). *Top: mean square error (MSE) for the tail index, number of missing observations, and* 99.8% *quantile, as a function of top* k *order statistics used. Center: Corresponding bias plots as a function of top* k *order statistics. Bottom: density of the estimators with automatically-selected* k.

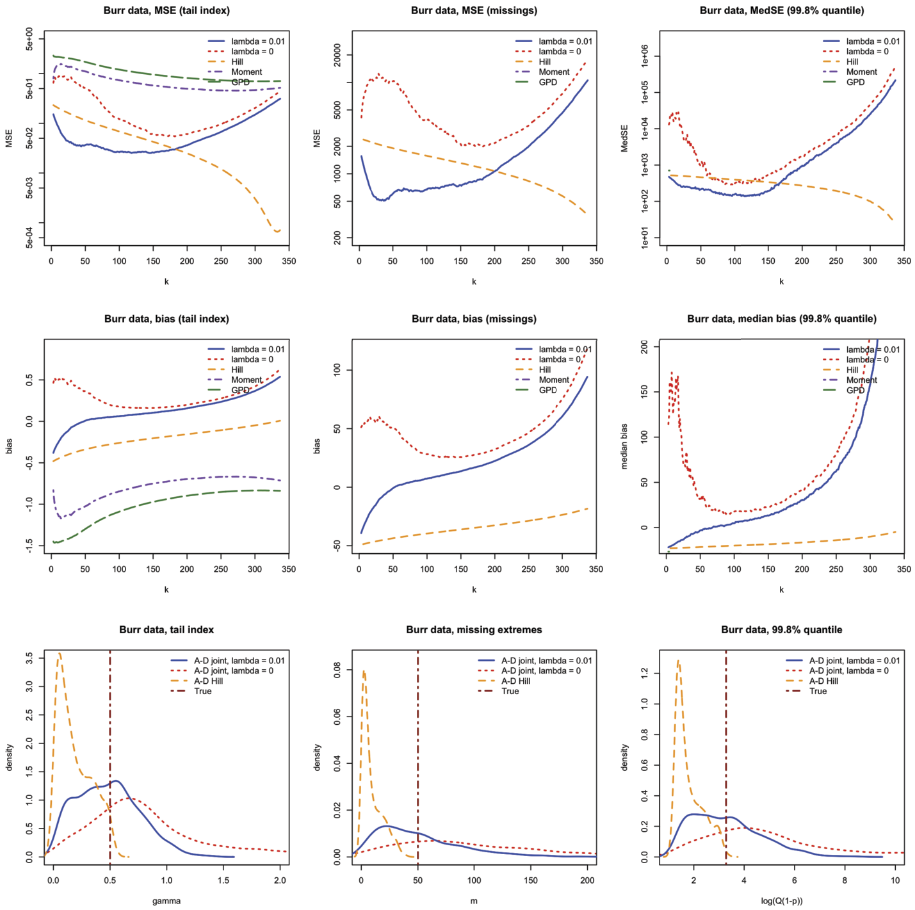

FIG 8. *Simulations results for Burr*$(2, -2, 2)$ *data* $(\gamma = 1/2, m = 50)$. *Top: mean square error (MSE) for the tail index, number of missing observations, and* 99.8% *quantile, as a function of top* k *order statistics used. Center: Corresponding bias plots as a function of top* k *order statistics. Bottom: density of the estimators with automatically-selected* k.

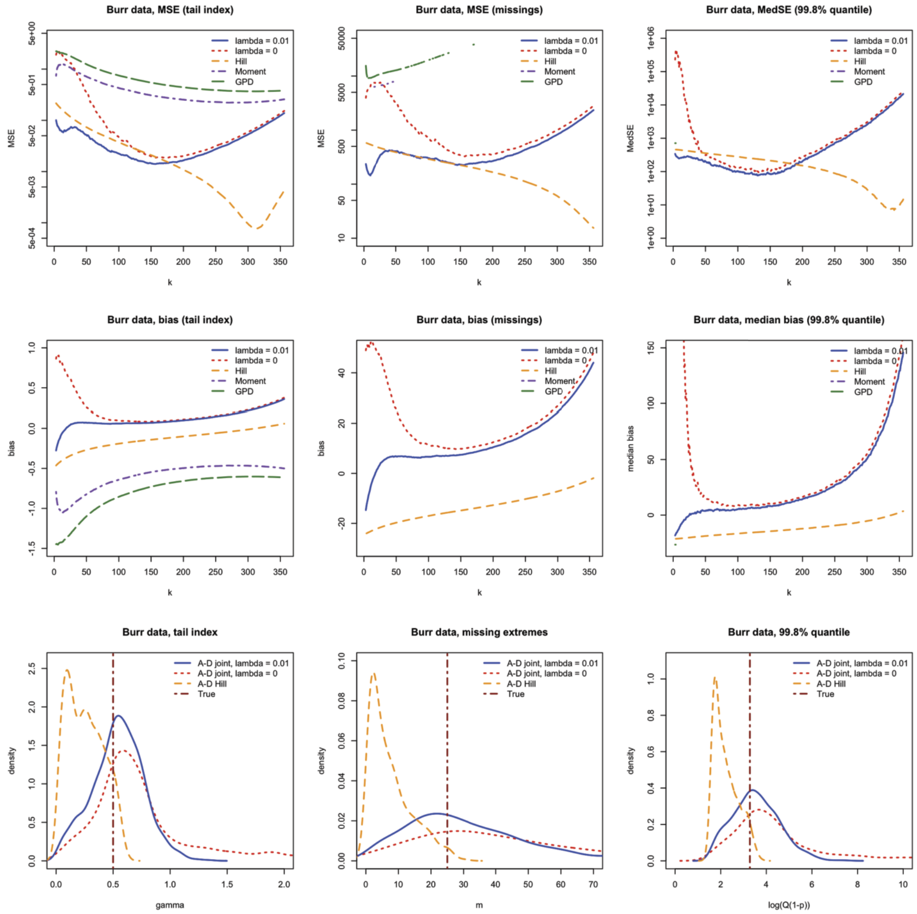

Fig 9. *Simulations results for Burr*$(2, -2, 2)$ *data* $(\gamma = 1/2, m = 25)$. *Top: mean square error (MSE) for the tail index, number of missing observations, and* 99.8% *quantile, as a function of top* k *order statistics used. Center: Corresponding bias plots as a function of top* k *order statistics. Bottom: density of the estimators with automatically-selected* k.

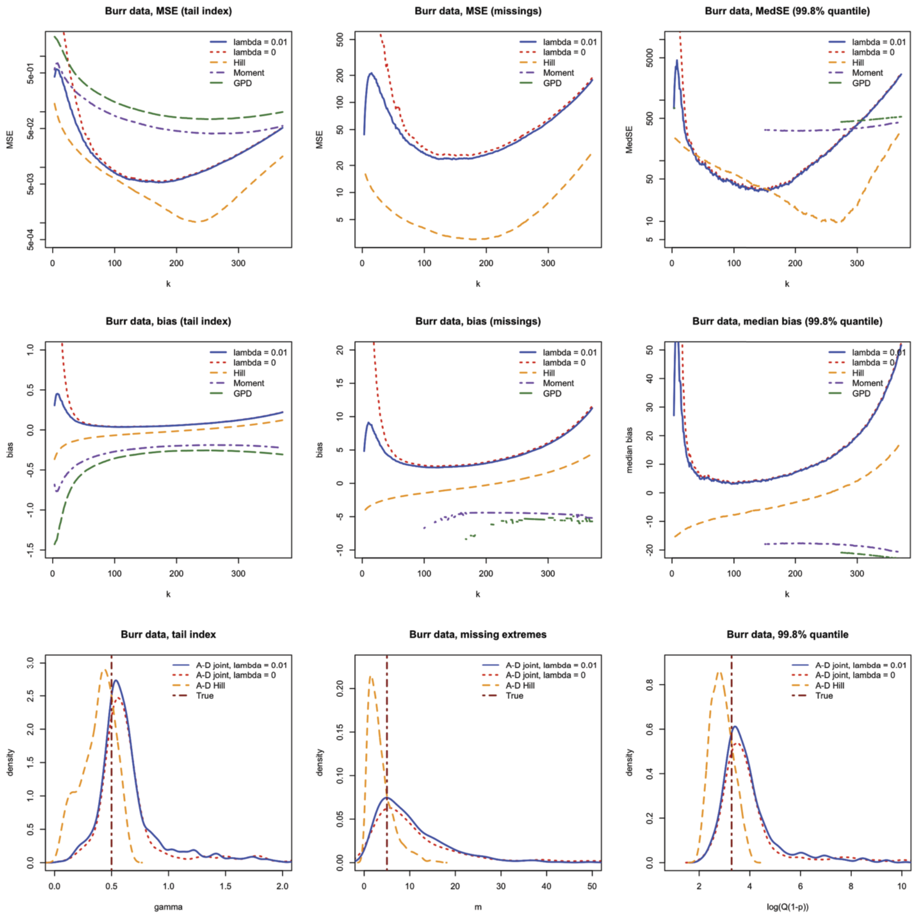

FIG 10. *Simulations results for Burr*$(2, -2, 2)$ *data* $(\gamma = 1/2, m = 5)$. *Top: mean square error (MSE) for the tail index, number of missing observations, and* 99.8% *quantile, as a function of top* k *order statistics used. Center: Corresponding bias plots as a function of top* k *order statistics. Bottom: density of the estimators with automatically-selected* k.

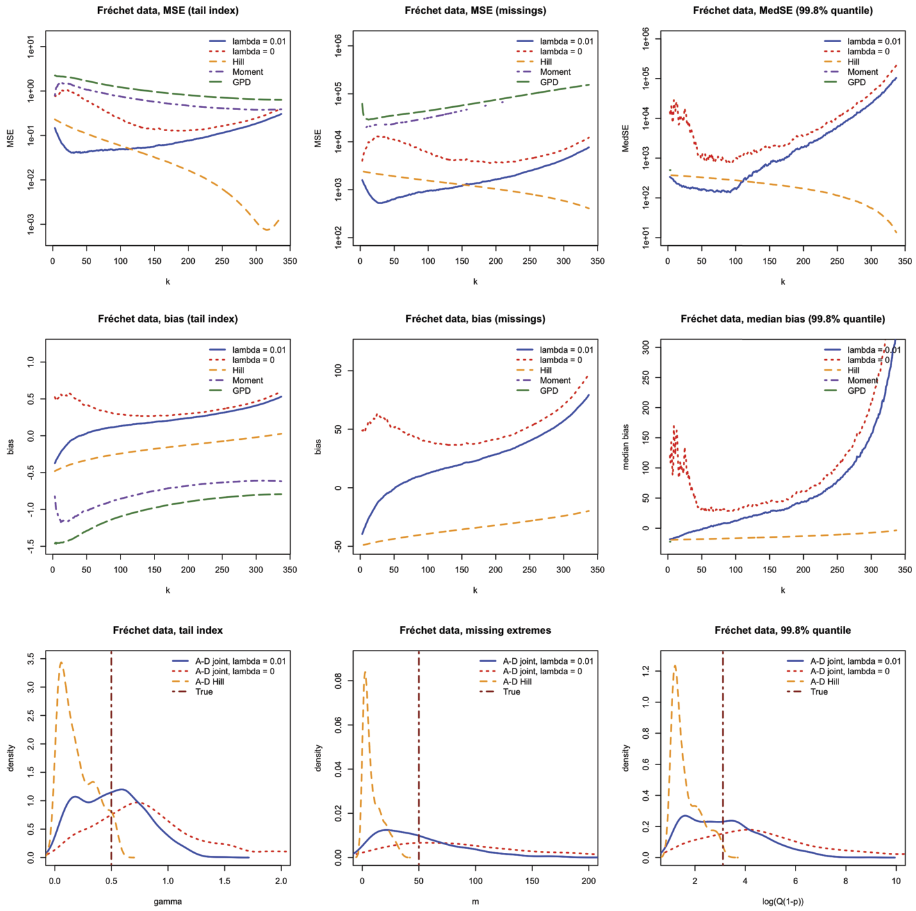

FIG 11. *Simulations results for Fréchet*(2) *data* ($\gamma = 1/2, m = 50$). *Top: mean square error (MSE) for the tail index, number of missing observations, and* 99.8% *quantile, as a function of top k order statistics used. Center: Corresponding bias plots as a function of top k order statistics. Bottom: density of the estimators with automatically-selected k.*

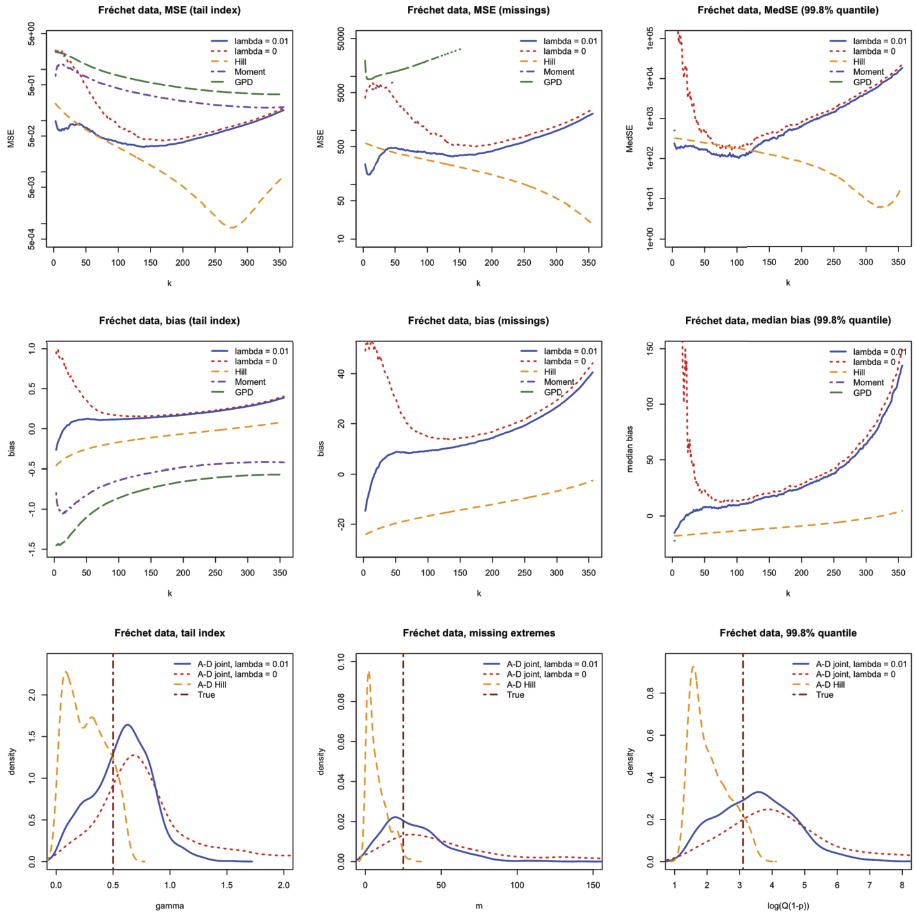

FIG 12. *Simulations results for Fréchet*(2) *data* ($\gamma = 1/2, m = 25$)*. Top: mean square error (MSE) for the tail index, number of missing observations, and* 99.8% *quantile, as a function of top k order statistics used. Center: Corresponding bias plots as a function of top k order statistics. Bottom: density of the estimators with automatically-selected k.*

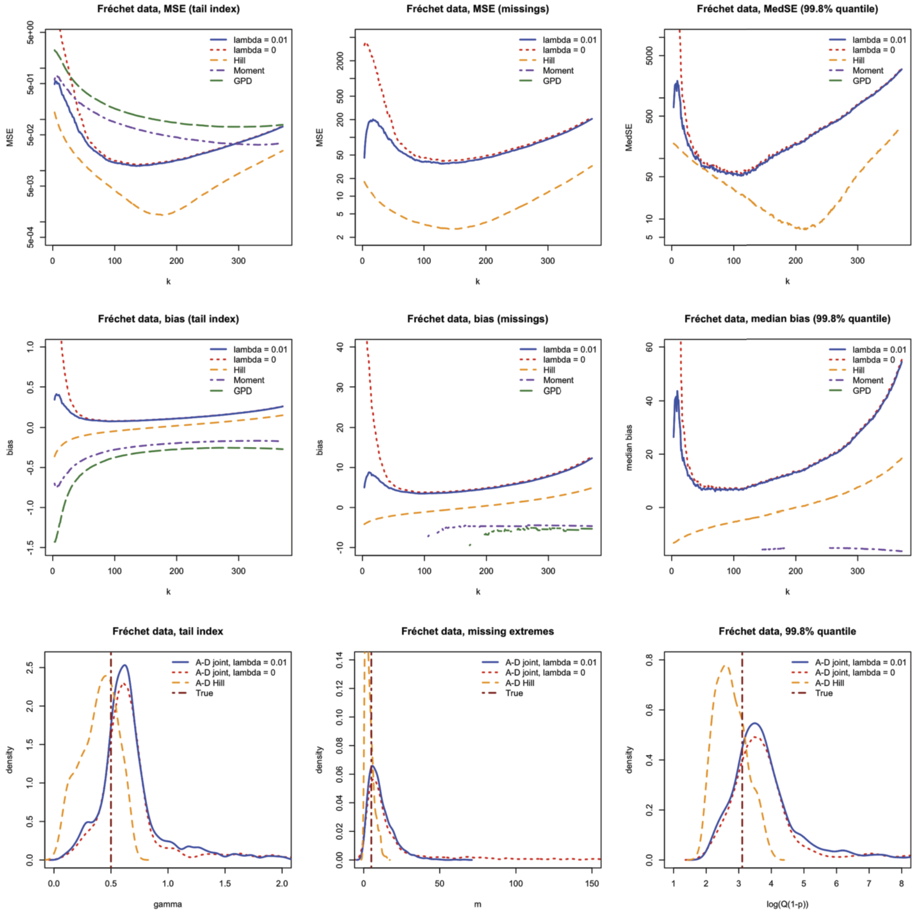

Fig 13. *Simulations results for Fréchet(2) data ($\gamma = 1/2, m = 5$). Top: mean square error (MSE) for the tail index, number of missing observations, and 99.8% quantile, as a function of top k order statistics used. Center: Corresponding bias plots as a function of top k order statistics. Bottom: density of the estimators with automatically-selected k.*

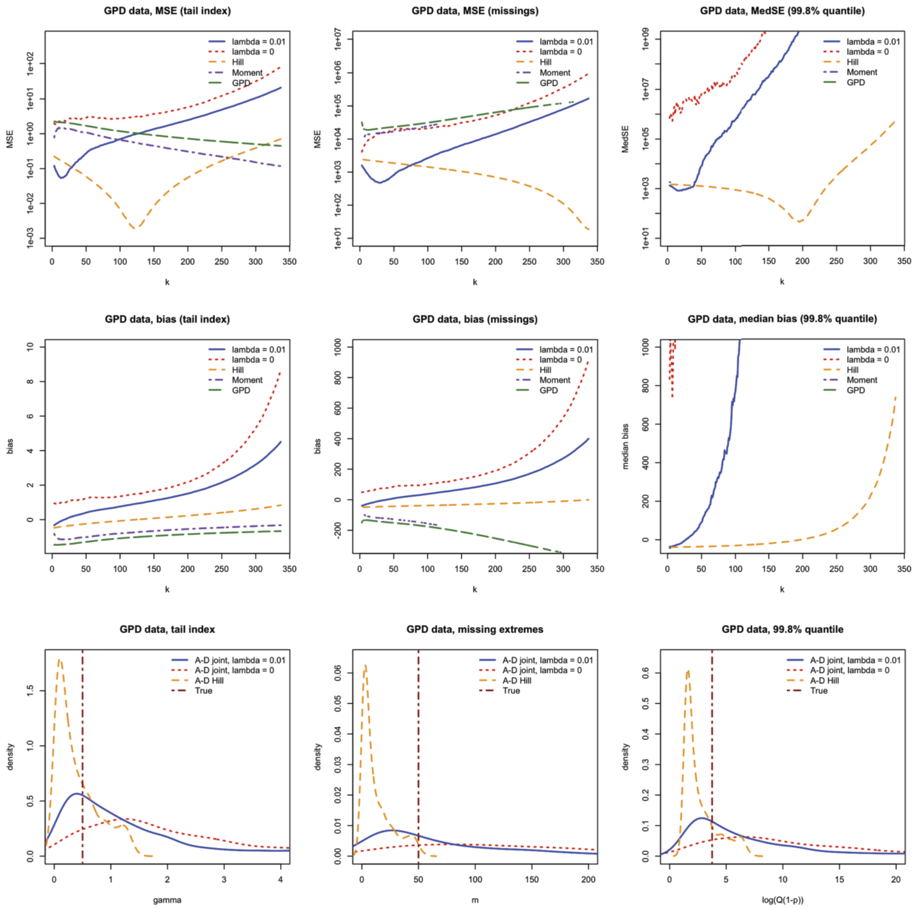

Fig 14. *Simulations results for GPD$(1/2, 1)$ data ($\gamma = 1/2, m = 50$). Top: mean square error (MSE) for the tail index, number of missing observations, and 99.8% quantile, as a function of top k order statistics used. Center: Corresponding bias plots as a function of top k order statistics. Bottom: density of the estimators with automatically-selected k.*

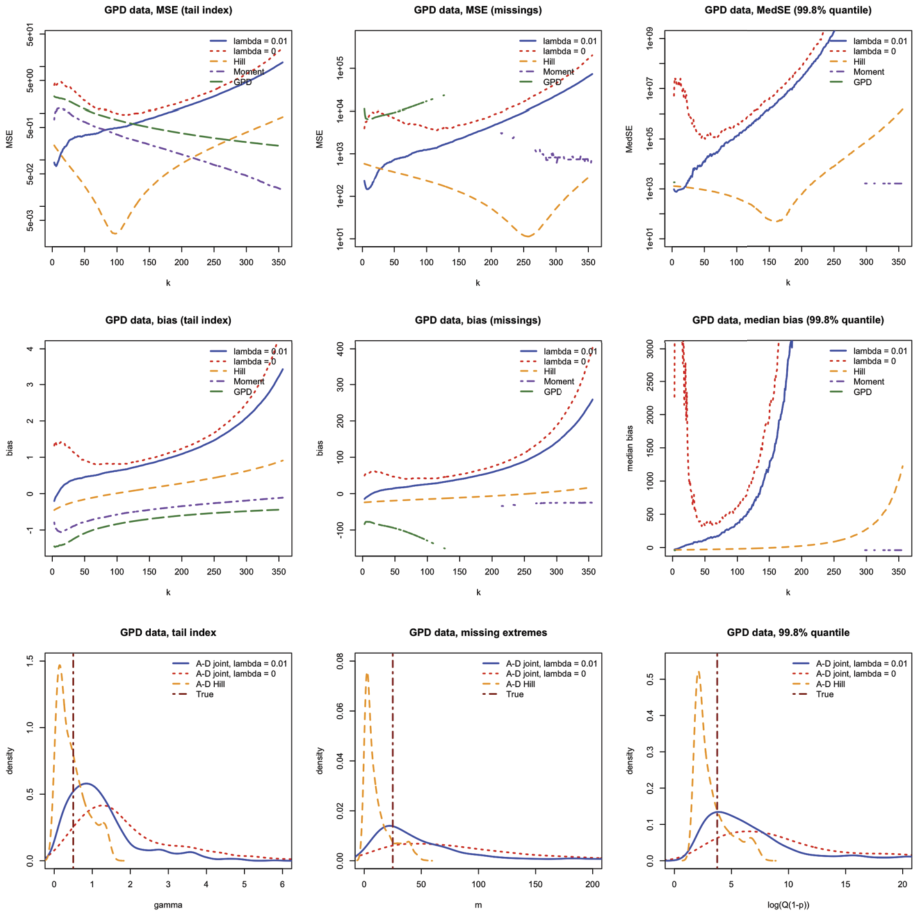

Fig 15. *Simulations results for GPD*$(1/2, 1)$ *data* $(\gamma = 1/2, m = 25)$. *Top: mean square error (MSE) for the tail index, number of missing observations, and* 99.8% *quantile, as a function of top* k *order statistics used. Center: Corresponding bias plots as a function of top* k *order statistics. Bottom: density of the estimators with automatically-selected* k.

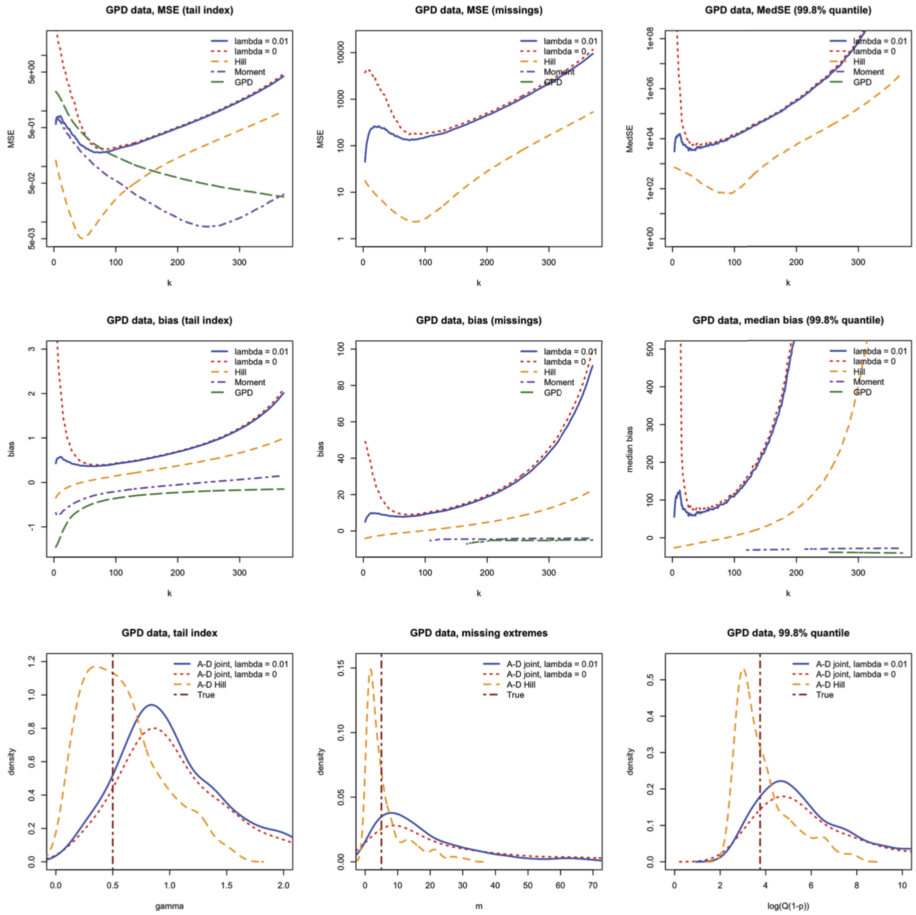

FIG 16. *Simulations results for GPD*$(1/2, 1)$ *data* $(\gamma = 1/2, m = 5)$. *Top: mean square error (MSE) for the tail index, number of missing observations, and* 99.8% *quantile, as a function of top k order statistics used. Center: Corresponding bias plots as a function of top k order statistics. Bottom: density of the estimators with automatically-selected k.*

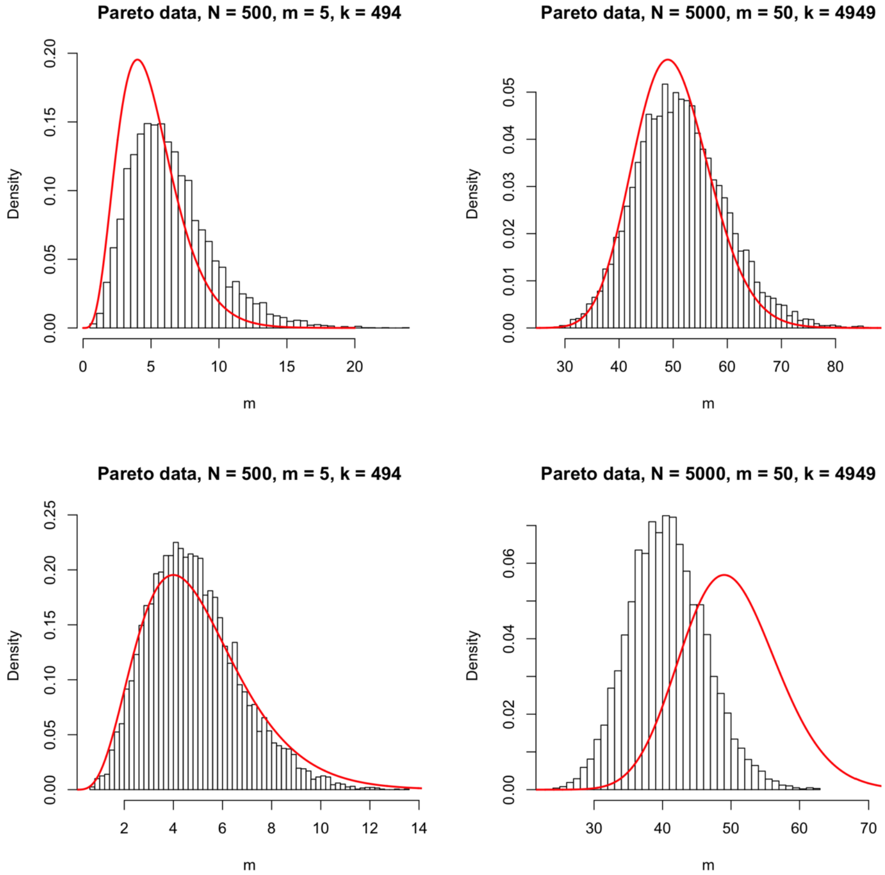

Fig 17. *Histograms and fitted* $\Gamma(m,1)$ *densities for simulated Pareto data. Top left:* $N = 500$, $m = 5$, $k = 494$*; top right:* $N = 5000$, $m = 50$, $k = 4940$*. In the bottom panels are the corresponding histograms using the plug-in missings estimator derived from inserting the Hill estimator into the second equation of* (2.2).

References

[1] Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769. MR0069459

[2] Beirlant, J., Dierckx, G., Goegebeur, Y., and Matthys, G. (1999). Tail index estimation and an exponential regression model. *Extremes*, 2(2):177–200. MR1771132

[3] Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of Extremes: Theory and Applications*, volume 558. John Wiley & Sons. MR2108013

[4] Bhattacharya, S., Kallitsis, M., and Stoev, S. (2019). Data-adaptive trim-

ming of the hill estimator and detection of outliers in the extremes of heavy-tailed data. *Electronic Journal of Statistics*, 13(1):1872–1925. MR3964266

[5] Bladt, M., Albrecher, H., and Beirlant, J. (2020). Threshold selection and trimming in extremes. *Extremes*, 23(4):629–665. MR4165035

[6] Feuerverger, A. and Hall, P. (1999). Estimating a tail exponent by modelling departure from a pareto distribution. *The Annals of Statistics*, 27(2):760–781. MR1714709

[7] Hall, P. and Welsh, A. H. (1985). Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, pages 331–341. MR0773171

[8] Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, pages 1163–1174. MR0378204

[9] Verster, A., de Waal, D., Schall, R., and Prins, C. (2012). A truncated Pareto model to estimate the under recovery of large diamonds. *Math. Geosci.*, 44:91–100.

[10] Xu, H., Davis, R., and Samorodnitsky, G. (2022). Handling missing extremes in tail estimation. *Extremes*, 25(2):199–227. MR4417405

[11] Zou, J., Davis, R. A., and Samorodnitsky, G. (2020). Extreme value analysis without the largest values: what can be done? *Probability in the Engineering and Informational Sciences*, 34(2):200–220. MR4079139