

Efficient and robust estimation of tail parameters for Pareto and exponential models

Alain Desgagné^a 

Département de mathématiques, Université du Québec à Montréal, Montréal (Québec), Canada,

^adesgagne.alain@uqam.ca

Abstract. In this paper, a new efficient and robust estimator of the Pareto tail index is proposed. Although the emphasis is on the Pareto distribution, all results are valid for the estimation of the scale/rate parameter of the two-parameter exponential distribution. The approach is to assume that the observations were generated from the FLLP-contaminated Pareto, that is, a mixture of the Pareto and FLLP distributions. The latter is an original distribution designed specifically to represent any outlier distribution. The parameters are estimated using an iterative process adapted from the expectation-maximization (EM) algorithm to optimize the properties of the estimators in a robustness context. A robust confidence interval for the Pareto tail index is also given. It is shown through different asymptotic results that these estimators reach a breakdown point of 50% with full efficiency. Their simultaneous high efficiency and high robustness are also shown for finite samples in a large Monte Carlo simulation study. Finally, an example with a real dataset of daily crude oil returns is given.

1 Introduction

The Pareto distribution $\mathcal{P}(\sigma, \alpha)$ defined on the domain $[\sigma, \infty)$, whose distribution and density functions are defined by

$$F_{\mathcal{P}}(x|\sigma, \alpha) := 1 - \left(\frac{\sigma}{x}\right)^{\alpha} \quad \text{and} \quad f_{\mathcal{P}}(x|\sigma, \alpha) := \frac{\alpha\sigma^{\alpha}}{x^{1+\alpha}} \quad (1)$$

for fixed $\sigma > 0$ and $\alpha > 0$, is widely used in modeling variables from many fields such as economics, finance, actuarial science, computer science or biology, to name a few (see, e.g., Brzezinski, 2016). In particular, the Pareto distribution with its decreasing density having a heavy right tail is suitable for modeling the right tail of many distributions using upper observations. Indeed, an important result (see, e.g., Vandewalle et al., 2007) establishes that if the density of a random variable Z is regularly varying at infinity with index $-(1 + \alpha)$, then the conditional distribution of the relative excesses $X := (\frac{Z}{t}|Z > t)$ over a threshold $t > 0$ converges to a $\mathcal{P}(1, \alpha)$ as $t \rightarrow \infty$. In other words, if the right tail decay of the density of Z is predominantly polynomial ($\sim 1/z^{1+\alpha}$), like the heavy tailed distributions of Pareto, Cauchy, Student, Fréchet, Burr, inverse-gamma, log-logistic or log-gamma, then the distribution of the relative excesses over the threshold t is closely approximated by a $\mathcal{P}(1, \alpha)$ if t is large enough.

The estimation of the Pareto tail index α —or equivalently $\gamma := 1/\alpha$ —is a widely studied problem in statistics as evidenced by the review of more than a hundred estimators by Fedotenkov (2020). The maximum likelihood estimator (MLE), based on a random sample x_1, \dots, x_n assumed to come from a $\mathcal{P}(\sigma, \alpha)$, is given by

$$\frac{1}{\hat{\alpha}^{\text{MLE}}} = \hat{\gamma}^{\text{MLE}} := \frac{1}{n} \sum_{i=1}^n \log(x_i/\sigma), \quad (2)$$

Key words and phrases. Heavy tails, M -estimator, robust weighted maximum likelihood estimator, relative excesses over a large threshold, Monte Carlo simulations, outliers.

Received October 2022; accepted January 2024.

where the scale parameter σ , treated as a nuisance parameter in this article, is estimated by the sample minimum $x_{(1)} \equiv \min\{x_1, \dots, x_n\}$ if it is unknown. The MLE is considered the benchmark given its desirable asymptotic properties such as consistency and minimum variance among unbiased estimates. However, it is well known that the MLE is very sensitive to very large observations which conflict with most of the data originating from the assumed Pareto distribution, which is our definition of outliers in this paper. These outliers have an undesirable impact on the estimation of the Pareto tail index and therefore on the whole inference. For example, for the sample $x = (1.01, 1.02, 1.04, 1.05, 1.07, 1.10, 1.13, 1.17, 1.26) \approx F_{\mathcal{P}}^{-1}(0.1, \dots, 0.9|1, 10)$ “perfectly” generated from a $\mathcal{P}(1, 10)$, we obtain $\hat{\alpha}^{\text{MLE}} = 11.38$. If we add a single large observation from the top 0.1% of this Pareto model, say 2, the estimate of α drops drastically to $\hat{\alpha}^{\text{MLE}} = 6.74$.

Robust estimation methods are generally designed to resolve this conflict by limiting the impact of outliers and focusing the estimation on non-outliers. A few robust methods using different approaches have been proposed in the literature. According to the Monte Carlo simulation comparisons of Brazauskas and Serfling (2001) and Brzezinski (2016), the best estimators of the Pareto tail index in terms of the efficiency-robustness trade-off are as follows. The optimal B-robust estimator (OBRE) defined in Hampel et al. (1986) was adapted by Victoria-Feser and Ronchetti (1994) for the estimation of the Pareto tail index and Dupuis and Victoria-Feser (2006) introduced the weighted maximum likelihood estimator (WMLE). OBRE and WMLE are both M -estimators whose ψ -functions and influence functions have been bounded. Their ψ -functions are based on a weighted score function to reduce the impact of upper observations deviating from the Pareto model. The generalized median estimator (GME), proposed by Brazauskas and Serfling (2000a) and Brazauskas and Serfling (2000b), first consists of evaluating the MLE of α or γ (adjusted to be median unbiased) over all subsets of observations taken k at a time. The GME is the median of these $\binom{n}{k}$ subsets. Finkelstein, Tucker and Veeh (2006) introduced the probability integral transform statistic estimator (PITSE) as the solution in α of the equation $n^{-1} \sum_{i=1}^n (\sigma/x_i)^{\alpha t} = (t+1)^{-1}$, for a fixed tuning constant $t > 0$. The trimmed mean estimator (TME), introduced by Kimber (1983), consists first in discarding a determined proportion of the uppermost observations, then in calculating the MLE (corrected for the bias) on the others. More recently, other robust estimators of the Pareto tail index have been proposed, namely a non-parametric estimator using the extended Pareto distribution (Dierckx, Goegebeur and Guillou, 2014), a trimmed version of the Hill estimator (Bhattacharya, Kallitsis and Stoev, 2019) and estimators using an exponential regression model (Ghosh, 2017; Minkah, de Wet and Ghosh, 2021). Identification of large outliers in a Pareto model using boxplot methods was also studied by Safari, Masseran and Ibrahim (2019).

As can be seen, these approaches are quite different, but they all depend on a tuning constant to adjust the trade-off between efficiency and robustness. Naturally, high breakdown point robust estimators can easily be obtained with these methods, for example, by discarding the 40% upper observations. However, this unrefined way comes at a high cost in terms of efficiency, where the performance of a very robust estimator in the absence of outliers is poor. Conversely, if the tuning constant is adjusted to obtain an efficient estimator, the robustness becomes rather limited. Common usage is therefore to compromise by finding the sweet spot to obtain a moderately efficient and robust estimator.

To fill this gap, we propose in this paper a novel Pareto tail index estimator exhibiting simultaneously high breaking point and high efficiency, without compromise. Like the WMLE, it can be thought of as an M -estimator with a ψ -function based on the weighted score function, where the weights depend on a constant ω representing the proportion of non-outliers. However, we go further by estimating ω , which allows us to achieve both efficiency and robustness. Our approach is inspired by the work of Yohai (1987), Gervini and Yohai (2002)

and mainly [Desgagné \(2021\)](#) which also achieve both efficiency and robustness in a linear regression context. The first step is to assume that the observations were generated from a mixture of the Pareto distribution and another distribution that we design specifically to represent outliers. The second step is to develop an iterative estimation method by modifying the expectation-maximization (EM) algorithm to optimize the estimation in a context of outliers. We obtain highly efficient estimators under the Pareto model for all sample sizes, including full asymptotic efficiency at a fast rate. Moreover, under the Pareto model, our estimators generate exactly the MLE benchmark most of the time (in a proportion ranging from 0.84 to 0.91 depending on the sample size), which then means no cost for use of a robust estimator. We also obtain very robust estimators that reach an asymptotic breakdown point of 50%. Unlike most robust estimators of the Pareto tail index, inference is entirely based on observations, which means that there is no tuning constant to be set by the user. Therefore, no trade-off is made between efficiency and robustness; on the contrary, we get both. We also obtain for each observation an estimated probability that it comes from the Pareto component (versus the outlier component), which makes the detection of outliers very easy and efficient.

Although the focus is on the Pareto distribution in this article, all the results can be easily adapted for the exponential distribution $\mathcal{E}(\mu, \alpha)$ having the density $f_{\mathcal{E}}(y|\mu, \alpha) := \alpha e^{-\alpha(y-\mu)}$, with $y \in [\mu, \infty)$, $\mu \in \mathbb{R}$, $\alpha > 0$. This can be done using a simple log/exponential transformation since $X \stackrel{\mathcal{L}}{\sim} \mathcal{P}(\sigma, \alpha) \Leftrightarrow Y \stackrel{\mathcal{L}}{\sim} \mathcal{E}(\mu, \alpha)$ if $Y = \log X$ and $\mu = \log \sigma$. The exponential distribution is suitable for modeling the right tail—if the threshold is chosen large enough—of a density whose decay is predominantly exponential ($\sim \exp(-\alpha z)$), like the distributions of Laplace, gamma (including χ^2 and exponential), logistic or Gumbel.

In Section 2, we design an original contaminated Pareto distribution called the P-FLLP distribution. Robust inference of the Pareto model using the P-FLLP distribution is studied in Section 3. We construct an iterative estimation method that yields the robust P-FLLP estimators in Section 3.1, the method is validated by asymptotic theoretical results in Section 3.2, a connection between our approach and M -estimators is made in Section 3.3 and robust confidence interval for the Pareto tail index is discussed in Section 3.4. In Section 4, we study the performance of the P-FLLP estimator of the Pareto tail index in a Monte Carlo simulation, in the absence and in the presence of outliers. An example is given in Section 5 and we conclude in Section 6. The detailed programming code using R software is provided in the Supplementary Material ([Desgagné \(2024\)](#)).

2 Modeling observations with a contaminated Pareto distribution

In this section, we design an original distribution adapted to our robustness context according to specific optimality criteria. Without loss of generality, the description of our approach is first made assuming the standardized Pareto distribution $\mathcal{P}(1, 1)$ for ease of understanding, with its density given by $f_{\mathcal{P}}(x|1, 1) = x^{-2}$, $x \geq 1$. Recall that outliers are defined, in this article, as very large observations that conflict with most of the data coming from the assumed Pareto distribution. In other words, the right tail of the Pareto density is not heavy enough to accommodate these too large observations called outliers, resulting in a bad trade-off in the tail index estimate. A distinction is made between large observations that are plausible under the Pareto model and outliers that are too large to be plausible. The idea behind our approach is to use a model that describes our context as closely as possible, that is, the observations come from a Pareto distribution except for possible large outliers.

The first step in building our model is to define a threshold τ that defines the non-outlier region $[1, \tau]$ and the outlier region (τ, ∞) , and to replace the right tail of the Pareto density over the outlier region $[\tau, \infty)$ with a heavier tail to account for potential outliers. It is well

known that heavy tail modeling is an efficient way to deal with outliers, but the heaviness of the right tail is yet to be determined. In the context of scale inference—such as our exponential model $\mathcal{E}(\mu, 1/\gamma)$ —Desgagné (2013) showed that super heavy tails with logarithmic decay (e.g., log-Pareto decay) allow full robustness, in that the influence of outliers disappears completely when their distance with the bulk of data goes to infinity. Under the Pareto model $\mathcal{P}(\sigma, \alpha)$, this result translates into log-log-Pareto decay. Therefore, to achieve this full robustness, we replace the right tail of the Pareto density with a log-log-Pareto decay of the form $(x \log x)^{-1} (\log \log x)^{-(\lambda+1)}$, where $\lambda > 0$ determines the level of tail decay. To maintain a density function which integrates to 1, it is necessary in return to multiply the modified Pareto density by a normalizing constant $\omega > 0$, which depends on the two added parameters τ and λ .

The second step consists in rethinking our model as a contaminated Pareto distribution, that is, a mixture of a Pareto $\mathcal{P}(1, 1)$ distribution defined on $[1, \infty)$ and a second component defined on (τ, ∞) which models the outliers. The normalizing constant ω becomes a parameter which now plays the role of the mixture weight, which is possible if we add the constraint $0 < \omega < 1$. These ω and $1 - \omega$ weights can be interpreted as the proportion of non-outliers and the proportion of outliers, respectively. To compensate for the fact that ω is now a free parameter, the threshold τ is now set automatically as a function of ω . Finally, to ensure that the outlier component is positive and well-defined as a density—or equivalently, that the right tail of the modified Pareto is heavier everywhere than that of the original Pareto—we need to make sure the right tail is heavy enough or equivalently, that λ is small enough. To achieve this, we set λ appropriately as a function of ω . We obtain a contaminated Pareto distribution with only the parameter ω (and eventually the parameters σ and α for the non-standardized version).

We now see in more details how the contaminated Pareto with the three parameter ω, σ, α is constructed and then discuss its properties at the end of this section. Similar to the family of log-Pareto-tailed symmetric distributions introduced by Desgagné (2015), the first step, as described above, is to replace the right tail of the Pareto distribution with a log-log-Pareto decay. We obtain the following continuous density:

$$f(x|\sigma, \alpha, \tau, \lambda) = \begin{cases} \omega \alpha x^{-1} z^{-1} & \text{if } \sigma \leq x \leq \sigma \tau^{1/\alpha} \Leftrightarrow 1 \leq z \leq \tau, \\ \omega \alpha x^{-1} \tau^{-1} \frac{\log \tau}{\log z} \left(\frac{\log \log \tau}{\log \log z} \right)^{\lambda+1} & \text{if } x > \sigma \tau^{1/\alpha} \Leftrightarrow z > \tau, \end{cases} \quad (3)$$

where $z := (x/\sigma)^\alpha$ is the standardized version of x , $\sigma > 0$ is a scale parameter, $\alpha > 0$ is a power parameter, $\tau > e$ is a threshold that determines the outlier region, $\lambda > 0$ determines the level of the log-log-Pareto tail decay and ω acts, at this stage, as a normalizing constant. Note that the term $\alpha x^{-1} z^{-1}$ (if $\sigma \leq x \leq \sigma \tau^{1/\alpha}$) corresponds to the Pareto density $f_{\mathcal{P}}(x|\sigma, \alpha)$.

Proposition 2.1. *The normalizing constant ω is given by*

$$\omega = (1 - \tau^{-1} + \lambda^{-1} \tau^{-1} (\log \tau) \log \log \tau)^{-1}. \quad (4)$$

Proof. Using the changes of variables $z = (x/\sigma)^\alpha$ and $u = (\log \log z)/(\log \log \tau)$, respectively in the second and fourth equalities, we have

$$\begin{aligned} \omega^{-1} &= \omega^{-1} \int_{\sigma}^{\infty} f(x|\sigma, \alpha, \tau, \lambda) dx = \omega^{-1} \int_1^{\infty} f(z|1, 1, \tau, \lambda) dz \\ &= \int_1^{\tau} z^{-2} dz + \tau^{-1} (\log \tau) \int_{\tau}^{\infty} \frac{1}{z \log z} \left(\frac{\log \log \tau}{\log \log z} \right)^{\lambda+1} dz \end{aligned}$$

$$\begin{aligned}
&= 1 - \tau^{-1} + \tau^{-1}(\log \tau)(\log \log \tau) \int_1^\infty \left(\frac{1}{u}\right)^{\lambda+1} du \\
&= 1 - \tau^{-1} + \lambda^{-1} \tau^{-1}(\log \tau) \log \log \tau. \quad \square
\end{aligned}$$

The density $f(x|\sigma, \alpha, \tau, \lambda)$ in (3) with its log-log-Pareto right tail decay ensures full robustness, but it is not guaranteed to be a contaminated Pareto distribution. The next step then consists in rewriting this density as a mixture of a Pareto with an outlier component. We observe that the normalizing constant ω can also play the role of a mixture weight if we add the constraint $0 < \omega < 1$. Since ω is now a free parameter, in return we make the parameter τ act as the normalizing constant, so τ is set automatically using (4). We obtain

$$f(x|\omega, \sigma, \alpha, \lambda) = \omega f_{\mathcal{P}}(x|\sigma, \alpha) + (1 - \omega) f_{\mathcal{C}}(x|\omega, \sigma, \alpha, \lambda),$$

where $0 < \omega < 1$ is the mixture weight for the Pareto component and $f_{\mathcal{C}}(x|\omega, \sigma, \alpha, \lambda)$ is the contaminating component. If $\omega = 1$, we define $f(x|\omega, \sigma, \alpha, \lambda) = f_{\mathcal{P}}(x|\sigma, \alpha)$, that is, the contaminated Pareto is simply the Pareto distribution. Therefore, if $0 < \omega < 1$, we can deduce that the contaminating component is defined on $x \in (\sigma \tau^{1/\alpha}, \infty) \Leftrightarrow z \in (\tau, \infty)$ and given by

$$\begin{aligned}
&f_{\mathcal{C}}(x|\omega, \sigma, \alpha, \lambda) \\
&= (1 - \omega)^{-1} f(x|\omega, \sigma, \alpha, \lambda) - (1 - \omega)^{-1} \omega f_{\mathcal{P}}(x|\sigma, \alpha) \\
&= \begin{cases} 0 & \text{if } \sigma \leq x \leq \sigma \tau^{1/\alpha} \Leftrightarrow 1 \leq z \leq \tau, \\ (1 - \omega)^{-1} \omega \alpha x^{-1} \left[\tau^{-1} \frac{\log \tau}{\log z} \left(\frac{\log \log \tau}{\log \log z} \right)^{\lambda+1} - z^{-1} \right] & \text{if } x > \sigma \tau^{1/\alpha} \Leftrightarrow z > \tau, \end{cases}
\end{aligned}$$

where $z := (x/\sigma)^\alpha$. By construction, $f_{\mathcal{C}}(x|\omega, \sigma, \alpha, \lambda)$ integrates to 1 over its domain. Therefore, the last step is to fix the parameter λ such that $f_{\mathcal{C}}(x|\omega, \sigma, \alpha, \lambda) > 0$ for $x > \sigma \tau^{1/\alpha}$, so that $f_{\mathcal{C}}(x|\omega, \sigma, \alpha, \lambda)$ is well defined as a density.

Proposition 2.2. *The density $f_{\mathcal{C}}(x|\omega, \sigma, \alpha, \lambda)$ is strictly positive for $x > \sigma \tau^{1/\alpha} \Leftrightarrow z > \tau$ if and only if $0 < \lambda \leq (\log \tau - 1)(\log \log \tau) - 1$.*

Proof. We first observe that the condition $\tau > e \Leftrightarrow \log \tau > 1$ ensures that $\log z > 1$ and $\log \log z > 0$ for all $z \geq \tau$. We want to show, for $z > \tau \Leftrightarrow \log z > \log \tau$, that

$$\begin{aligned}
f_{\mathcal{C}}(x|\omega, \sigma, \alpha, \lambda) > 0 &\Leftrightarrow \tau^{-1} \frac{\log \tau}{\log z} \left(\frac{\log \log \tau}{\log \log z} \right)^{\lambda+1} - z^{-1} > 0 \\
&\Leftrightarrow \tau^{-1} (\log \tau) (\log \log \tau)^{\lambda+1} > z^{-1} (\log z) (\log \log z)^{\lambda+1} \\
&\Leftrightarrow \log z - \log \log z - (\lambda + 1) \log \log \log z > \log \tau - \log \log \tau - (\lambda + 1) \log \log \log \tau \\
&\Leftrightarrow h(\log z) > h(\log \tau),
\end{aligned}$$

where

$$h(u) := u - \log u - (\lambda + 1) \log \log u,$$

with

$$h'(u) = 1 - u^{-1} - (\lambda + 1)(u \log u)^{-1} \quad \text{and} \quad h''(u) = u^{-2} + (\lambda + 1)(\log u + 1)(u \log u)^{-2}.$$

We observe that $h(u)$ is convex for all $u \geq \log \tau > 1$ since $h''(u) > 0$ for all $u > 1$. Therefore,

$$\begin{aligned}
h(\log z) > h(\log \tau) \quad \text{for all } \log z > \log \tau &\Leftrightarrow h'(\log \tau) \geq 0 \\
&\Leftrightarrow 1 - (\log \tau)^{-1} - (\lambda + 1)(\log \tau)^{-1} (\log \log \tau)^{-1} \geq 0 \\
&\Leftrightarrow \lambda \leq (\log \tau - 1)(\log \log \tau) - 1. \quad \square
\end{aligned}$$

While Proposition 2.2 says that all values of $0 < \lambda \leq (\log \tau - 1)(\log \log \tau) - 1$ generate a valid density, we set $\lambda := (\log \tau - 1)(\log \log \tau) - 1$ to obtain the lightest log-log-Pareto right tail decay. In doing so, we simplify the estimation of the density by reducing the number of parameters to three, namely ω , σ and α .

We are now ready to define formally our contaminated Pareto model. We name the contaminating component “filtered-log-log-Pareto (FLLP)” distribution (specifically, filtered of the Pareto) and we name the mixture “FLLP-contaminated Pareto distribution” or simply P-FLLP distribution.

Definition 2.1. A random variable X is said to have a filtered-log-log-Pareto (FLLP) distribution (specifically, filtered of the Pareto), if its density is given by

$$f_{\text{FLLP}}(x|\omega, \sigma, \alpha) := (1 - \omega)^{-1} \omega \alpha x^{-1} \left[\tau^{-1} \frac{\log \tau}{\log z} \left(\frac{\log \log \tau}{\log \log z} \right)^{\lambda+1} - z^{-1} \right],$$

for $x > \sigma \tau^{1/\alpha} \Leftrightarrow z > \tau$, where $z := (x/\sigma)^\alpha$ is the standardized version of x , $0 < \omega < 1$, $\sigma > 0$ is a scale parameter, $\alpha > 0$ is a power parameter, the decay of the right tail is determined by $\lambda > 0$ as a function of τ by

$$\lambda \equiv \lambda(\tau) := (\log \tau - 1)(\log \log \tau) - 1,$$

and the domain is determined by $\tau > 9.3931236$ as a function of ω , where $\tau \equiv \tau(\omega)$ is the unique solution of

$$\omega^{-1} = 1 - \tau^{-1} + \lambda^{-1} \tau^{-1} (\log \tau) \log \log \tau. \tag{5}$$

Note that $\omega \rightarrow 0 \Leftrightarrow \lambda \rightarrow 0 \Leftrightarrow \tau \rightarrow 9.39312358 \dots$ and $\omega \rightarrow 1 \Leftrightarrow \lambda \rightarrow \infty \Leftrightarrow \tau \rightarrow \infty$.

Definition 2.2. A random variable X is said to have an FLLP-contaminated Pareto distribution—also called a P-FLLP distribution—if its distribution is this following mixture:

$$\text{P-FLLP}(\omega, \sigma, \alpha) := \omega \mathcal{P}(\sigma, \alpha) + (1 - \omega) \text{FLLP}(\omega, \sigma, \alpha),$$

with $\mathcal{P}(\sigma, \alpha)$ denoting a Pareto distribution as defined in (1) and $\text{FLLP}(\omega, \sigma, \alpha)$ denoting the filtered-log-log-Pareto (FLLP) distribution as given in Definition 2.1. If $\omega = 1$, we define $\text{P-FLLP}(\omega = 1, \sigma, \alpha) := \mathcal{P}(\sigma, \alpha)$. The density of the P-FLLP(ω, σ, α) distribution, if $0 < \omega < 1$, can be written as

$$f_{\text{P-FLLP}}(x|\omega, \sigma, \alpha) := \omega f_{\mathcal{P}}(x|\sigma, \alpha) + (1 - \omega) f_{\text{FLLP}}(x|\omega, \sigma, \alpha) \\ = \begin{cases} \omega \alpha x^{-1} z^{-1} & \text{if } \sigma \leq x \leq \sigma \tau^{1/\alpha} \Leftrightarrow 1 \leq z \leq \tau, \\ \omega \alpha x^{-1} \tau^{-1} \frac{\log \tau}{\log z} \left(\frac{\log \log \tau}{\log \log z} \right)^{\lambda+1} & \text{if } x > \sigma \tau^{1/\alpha} \Leftrightarrow z > \tau, \end{cases}$$

where $z := (x/\sigma)^\alpha$ is the standardized version of x , $0 < \omega < 1$ is the mixture weight, $\sigma > 0$ is a scale parameter and $\alpha > 0$ is a power parameter. The decay of the right tail is determined by $\lambda > 0$ and the outlier region is determined by $\tau > 9.3931236$, where $\lambda \equiv \lambda(\tau)$ is defined as a function of τ and $\tau \equiv \tau(\omega)$ is defined as a function of ω , such as given in Definition 2.1.

We have illustrated in Figure 1 the P-FLLP(ω, σ, α) density for $\sigma = 1$ and different combinations of ω, α . We also showed its $\mathcal{P}(\sigma, \alpha)$ and $\text{FLLP}(\omega, \sigma, \alpha)$ components. We observe that the P-FLLP mixture (solid gray line) and the Pareto density (dashed black line) are visually almost indistinguishable, which is a desirable property for efficient parameter estimation is the absence of outliers. More precisely, as expressed in Definition 2.2 and illustrated in Figure 1, the P-FLLP density is the Pareto density downweighted by the mixture weight ω in

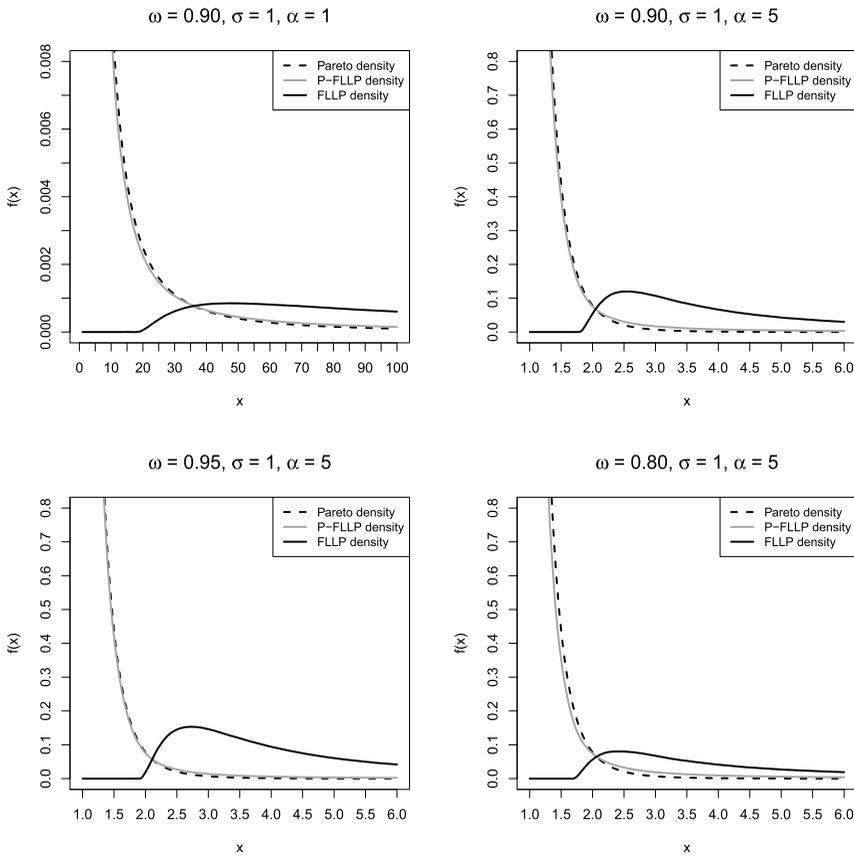


Figure 1 The $P\text{-FLLP}(\omega, \sigma, \alpha)$ density (solid gray line) as a mixture of the $\mathcal{P}(\sigma, \alpha)$ (dashed black line) and the $FLLP(\omega, \sigma, \alpha)$ (solid black line), for $\sigma = 1$ and different combinations of ω, α .

the core part defined as $1 \leq z \leq \tau$ and to compensate, the right tail $z > \tau$ (the outlier region) is raised everywhere to accommodate for possibly more extreme observations than expected under the Pareto model.

We also observe in Figure 1 the outlier component represented by the $FLLP(\omega, \sigma, \alpha)$ density (solid black line). The FLLP density is defined only on the right tail ($z > \tau$) in the spirit of the outlier region defined by [Davies and Gather \(1993\)](#). It is a rather diffuse or vague density with its log-log-Pareto decay, similar to the non-informative prior in Bayesian inference. We do not seek with the outlier component to precisely capture a specific contamination distribution. Rather, it aims to approximate any upper outlier distribution through its super heavy tail. Indeed, as mentioned above, it is shown in [Desgagné \(2013\)](#) that super heavy tails with logarithmic decay allow full robustness in estimating a scale model or, equivalently, with log-log decay in the estimation of a power model such as the Pareto tail index, and this, for any outlier pattern. We focus on modeling non-outliers, not outliers.

The last but one of the most important features of the $P\text{-FLLP}(\omega, \sigma, \alpha)$ mixture in a robust estimation context is its automatic adaptation to the proportion of outliers. Both the outlier region threshold τ and the tail decay λ depend uniquely on the expected proportion of Pareto observations ω , or equivalently on the expected proportion of outliers $1 - \omega$. A larger proportion of outliers (i.e., a smaller value of ω) is accompanied by smaller τ and λ values, meaning a larger outlier region with a heavier right tail, and vice-versa. For example, a value of $\omega = 0.95$ generates $\tau = 25.421$ and $\lambda = 1.625$, while a smaller value of $\omega = 0.90$ generates $\tau = 18.171$ and $\lambda = 1.023$. Naturally, if $\omega = 1$, the $P\text{-FLLP}$ model has no contaminating component and becomes the Pareto distribution.

3 Robust inference of the Pareto model using the P-FLLP distribution

In this section, the efficient and robust estimator of the Pareto tail index using the FLLP-contaminated Pareto distribution is introduced. The P-FLLP estimator is presented in Section 3.1. In Section 3.2, we present asymptotic results for the P-FLLP estimator, where topics such as infinite outliers, convergence in probability and breakdown point are investigated. A connection between our approach and M -estimators is made in Section 3.3. We go a step further in Section 3.4 where robust confidence interval for the Pareto tail index is discussed.

3.1 Robust estimation of the Pareto tail index using the P-FLLP distribution

Consider a random sample $\mathbf{x} \equiv (x_1, \dots, x_n)$ assumed to come from the Pareto model $\mathcal{P}(\sigma, \alpha)$, except for potential upper outliers. Recall that σ is treated as a nuisance parameter. It is therefore assumed that σ is known, otherwise it suffices to replace it by its estimator $x_{(1)}$. Robust estimation of the Pareto tail index α should ideally be based only on Pareto observations, while excluding outliers. Our approach then consists in extending the Pareto model to the FLLP-contaminated Pareto distribution as introduced in Definition 2.2, that is,

$$x_i \stackrel{\mathcal{L}}{\sim} \text{P-FLLP}(\omega, \sigma, \alpha) := \omega \mathcal{P}(\sigma, \alpha) + (1 - \omega) \text{FLLP}(\omega, \sigma, \alpha).$$

The particular case $\omega = 1$, which represents the absence of outliers, gives the Pareto distribution. The outlier component is modeled by the FLLP distribution, which as mentioned earlier is a rather diffuse or vague density with a log-log-Pareto decay that aims to represent all upper outlier distributions through its super heavy tail. The objective is to estimate the ω and α parameters efficiently in a robustness context.

Our approach is to adapt the well-known expectation-minimization (EM) algorithm (Dempster, Laird and Rubin, 1977) to our robustness context. More precisely, the modification of the algorithm is done at step M, while step E is identical. We introduce an unobserved vector of n independent latent Bernoulli variables

$$\mathbf{v} := (v_1, \dots, v_n)^T, \quad (6)$$

defined as $v_i = 1$ with probability ω if the observation x_i comes from the Pareto component or $v_i = 0$ with probability $1 - \omega$ if it comes from the FLLP contaminating distribution. Since the EM algorithm is an iterative process, consider our current estimates to be $\hat{\omega}^{(t)}$ and $\hat{\alpha}^{(t)}$ at step (t) . Using Bayes' theorem, we can verify that $v_i | \mathbf{x}, \hat{\omega}^{(t)}, \hat{\alpha}^{(t)} \equiv v_i | x_i, \hat{\omega}^{(t)}, \hat{\alpha}^{(t)}$ has a Bernoulli distribution with its mean given by

$$\begin{aligned} \hat{\pi}_i^{(t)} &:= E_{v_i | x_i, \hat{\omega}^{(t)}, \hat{\alpha}^{(t)}}(v_i) = \frac{\hat{\omega}^{(t)} f_{\mathcal{P}}(x_i | \sigma, \hat{\alpha}^{(t)})}{\hat{\omega}^{(t)} f_{\mathcal{P}}(x_i | \sigma, \hat{\alpha}^{(t)}) + (1 - \hat{\omega}^{(t)}) f_{\text{FLLP}}(x_i | \hat{\omega}^{(t)}, \sigma, \hat{\alpha}^{(t)})} \\ &= \frac{\hat{\omega}^{(t)} f_{\mathcal{P}}(x_i | \sigma, \hat{\alpha}^{(t)})}{f_{\text{P-FLLP}}(x_i | \hat{\omega}^{(t)}, \sigma, \hat{\alpha}^{(t)})}. \end{aligned}$$

Ideally, if the v_i 's were known, we could accurately distinguish non-outliers ($v_i = 1$) from outliers ($v_i = 0$) and base the MLE on the non-outliers. Since they are unknown, the idea is to estimate each v_i by their posterior mean $\hat{\pi}_i^{(t)}$ and to base the inference on the observations weighted by the $\hat{\pi}_i^{(t)}$.

Step E consists in estimating the log-likelihood of the complete data by its expectation. Since the likelihood function is given by

$$L(\omega, \alpha | \mathbf{x}, \mathbf{v}) = \prod_{i=1}^n [\omega f_{\mathcal{P}}(x_i | \sigma, \alpha)]^{v_i} [(1 - \omega) f_{\text{FLLP}}(x_i | \omega, \sigma, \alpha)]^{1-v_i},$$

we have

$$\begin{aligned}
Q(\omega, \alpha | \hat{\omega}^{(t)}, \hat{\alpha}^{(t)}) &:= E_{\mathbf{v} | \mathbf{x}, \hat{\omega}^{(t)}, \hat{\alpha}^{(t)}}(\log L(\omega, \alpha | \mathbf{x}, \mathbf{v})) = \sum_{i=1}^n E_{v_i | x_i, \hat{\omega}^{(t)}, \hat{\alpha}^{(t)}} \\
&\quad (v_i(\log \omega + \log f_{\mathcal{P}}(x_i | \sigma, \alpha)) \\
&\quad + (1 - v_i)(\log(1 - \omega) + \log f_{\text{FLLP}}(x_i | \omega, \sigma, \alpha))) \\
&= \sum_{i=1}^n \hat{\pi}_i^{(t)} (\log \omega + \log f_{\mathcal{P}}(x_i | \sigma, \alpha)) \\
&\quad + (1 - \hat{\pi}_i^{(t)}) (\log(1 - \omega) + \log f_{\text{FLLP}}(x_i | \omega, \sigma, \alpha)).
\end{aligned}$$

Step M normally consists of estimating a new set of parameters $\hat{\omega}^{(t+1)}$ and $\hat{\alpha}^{(t+1)}$ by maximizing $Q(\omega, \alpha | \hat{\omega}^{(t)}, \hat{\alpha}^{(t)})$ with respect to ω and α . However, this is not optimal in our robustness context since the outlier component $f_{\text{FLLP}}(x_i | \omega, \sigma, \alpha)$ depends on both ω and α . Only Pareto observations should contain information about α and only the estimated proportion of Pareto observations should be used for the estimation of ω . Also, we focus on modeling non-outliers, not outliers. This is where we adapt the algorithm to our robustness context. We instead assume that the outlier component depends on a distinct set of parameters denoted by ω_0 and α_0 , so that the quantity to be maximized becomes

$$\begin{aligned}
&Q(\omega, \alpha, \omega_0, \alpha_0 | \hat{\omega}^{(t)}, \hat{\alpha}^{(t)}) \\
&= \sum_{i=1}^n \hat{\pi}_i^{(t)} (\log \omega + \log f_{\mathcal{P}}(x_i | \sigma, \alpha)) + (1 - \hat{\pi}_i^{(t)}) (\log(1 - \omega) + \log f_{\text{FLLP}}(x_i | \omega_0, \sigma, \alpha_0)).
\end{aligned}$$

Then we proceed normally by estimating $\hat{\omega}^{(t+1)}$ and $\hat{\alpha}^{(t+1)}$ by $\arg_{\omega, \alpha} \max < Q(\omega, \alpha, \omega_0, \alpha_0 | \hat{\omega}^{(t)}, \hat{\alpha}^{(t)})$ and we find

$$\hat{\omega}^{(t+1)} := \frac{1}{n} \sum_{i=1}^n \hat{\pi}_i^{(t)} \quad \text{and} \quad 1/\hat{\alpha}^{(t+1)} := \frac{\sum_{i=1}^n \hat{\pi}_i^{(t)} \log(x_i/\sigma)}{\sum_{i=1}^n \hat{\pi}_i^{(t)}}. \quad (7)$$

As for the estimation of ω_0 and α_0 , we simply set

$$\hat{\omega}_0^{(t+1)} := \hat{\omega}^{(t+1)} \quad \text{and} \quad \hat{\alpha}_0^{(t+1)} := \hat{\alpha}^{(t+1)}.$$

In doing so, our estimators are no longer maximum likelihood estimators. However, our approach has several advantages. We keep a small number of parameters to estimate and we obtain the explicit formulas given in (7), which would not be the case if we used the original EM algorithm. The estimated model remains an P-FLLP(ω, σ, α) distribution, which would not be the case if we kept distinct parameters for the FLLP component. Thanks to the weights $\hat{\pi}_i^{(t)}$, the estimation of the Pareto tail index is essentially done on Pareto observations and the estimation of ω represents, as it should, the mixture weight or the proportion of non-outliers. Moreover, our approach is validated by theoretical results presented in Section 3.2. For example, we show that our estimators achieve an asymptotic breakdown point (BP) of 50%, the highest desired value. In this paper, BP is defined as the largest proportion of infinite upper outliers that an estimator can robustly handle. We also show that our estimator of α , under the Pareto model, behaves asymptotically like the MLE and therefore shares the same asymptotic properties such as full efficiency. We also show in Section 4.1, that under the Pareto model, our estimators generate exactly the MLE benchmark in a proportion ranging from 0.84 to 0.91 (depending on the sample size), which is a rare feature for a robust estimator. We now formally present the P-FLLP estimators of the parameters ω and α .

Definition 3.1. The robust P-FLLP estimators of ω (the expected proportion of Pareto observations) and α (the Pareto tail index) are given by

$$\hat{\omega} := \frac{1}{n} \sum_{i=1}^n \hat{\pi}_i \quad \text{and} \quad \hat{\alpha}^{-1} = \hat{\gamma} := \frac{\sum_{i=1}^n \hat{\pi}_i \log(x_i/\sigma)}{\sum_{i=1}^n \hat{\pi}_i}, \quad \text{with}$$

$$\hat{\pi}_i := \frac{\hat{\omega} f_{\mathcal{P}}(x_i|\sigma, \hat{\alpha})}{f_{\text{P-FLLP}}(x_i|\hat{\omega}, \sigma, \hat{\alpha})}.$$

These estimators are the results of an iterative process. When different solutions are found, only those with $\hat{\omega} > 1/2$ are considered. The P-FLLP estimators are then defined as the solution with the largest value of $\hat{\alpha}$, or equivalently, with the smallest value of $\hat{\gamma}$. If σ is unknown, it is replaced by its estimate $x_{(1)}$.

An efficient way to perform multiple searches with different initial values is to consider values of $\hat{\omega}$ uniformly distributed between $1/2$ and 1 , for example, $\hat{\omega} \in 1/2 + \{\frac{1}{2(k+1)}, \frac{2}{2(k+1)}, \dots, \frac{k}{2(k+1)}\}$ for say k initial values. For each given $\hat{\omega}$, the initial value of $\hat{\alpha}$ is then set at the MLE (as defined in (2)) computed on the $\lfloor \hat{\omega}n \rfloor$ smallest observations, where $\lfloor \cdot \rfloor$ is the floor function. In practice, we have observed that $k = 5$ searches using this strategy are generally sufficient to find all existing solutions. We also noted that the iterative process always converges and that one or two solutions are usually found. In addition, the computation time is fast. For example, it took about 0.25 seconds with the R software to launch 5 different searches on a sample of size $n = 1000$. Note that we can consider that we have reached convergence when the difference between the estimators of two successive iterations is less than a chosen threshold such as 10^{-9} .

We now examine the strategy adopted in Definition 3.1 when the iterative process produces more than one solution. First, it is trivial to see that MLE with $\hat{\omega} = 1, \hat{\pi}_1 = 1, \dots, \hat{\pi}_n = 1$ is either the only solution (in the absence of outliers) or one of the solutions (in the presence of outliers). When solutions other than MLE are also found, only those with $\hat{\omega} > 0.50$ are considered, which has the effect (see Section 3.2) of generating a breakdown point of 50%, the highest desirable value. Recall that $\hat{\omega}$ is the estimate of the proportion of non-outliers and it is common sense that non-outliers represent at least half of the data, which is why we only consider solutions with $\hat{\omega} > 0.50$. Among these solutions, the P-FLLP estimators are defined as the solution with the largest value of $\hat{\alpha}$, which is the strategy generating the highest robustness to outliers while sacrificing very little efficiency (which remains very high) in the absence of outliers.

In addition to $\hat{\omega}$ which estimates the expected proportion of Pareto observations and $\hat{\alpha}$ which estimates the Pareto tail index, we also obtain a powerful tool for identifying outliers using $\hat{\pi}_i$, which estimates the probability that the observation x_i comes from the Pareto component. If an observation is located in the core part, that is $\sigma < x_i \leq \sigma \hat{\tau}^{1/\hat{\alpha}} \Leftrightarrow 1 \leq (x_i/\sigma)^{\hat{\alpha}} \leq \hat{\tau}$, then we have $\hat{\pi}_i = 1$, which means that the probability of being a non-outlier is estimated at 1. Note that $\hat{\tau} \equiv \hat{\tau}(\hat{\omega})$ is the unique solution in τ of equation (5), replacing ω by $\hat{\omega}$. If an observation is located in the outlier region, that is $x_i > \sigma \hat{\tau}^{1/\hat{\alpha}} \Leftrightarrow (x_i/\sigma)^{\hat{\alpha}} > \hat{\tau}$, then $\hat{\pi}_i$ decreases gradually from 1 to 0 as x_i increases to infinity. Equivalently, the estimated probability of being an outlier, given by $1 - \hat{\pi}_i$, increases from 0 to 1 as x_i increases from $\sigma \hat{\tau}^{1/\hat{\alpha}}$ to infinity. If we want to make a binary decision about the outlier nature of an observation, we can simply identify a value as an outlier if $\hat{\pi}_i < 0.5$ and as a non-outlier if $\hat{\pi}_i \geq 0.5$.

The P-FLLP estimator of α or γ in Definition 3.1 can be modified for some bias correction under the Pareto model. It is well known that if σ is known, $\hat{\gamma}^{\text{MLE}} := n^{-1} \sum_{i=1}^n \log(x_i/\sigma)$ is unbiased since $\hat{\gamma}^{\text{MLE}} \stackrel{\mathcal{L}}{\sim} \Gamma(n, \gamma/n)$. However, $\hat{\alpha}^{\text{MLE}} := n / \sum_{i=1}^n \log(x_i/\sigma)$ is biased and it is rather the estimator $\hat{\alpha}^{\text{MLE-U}} := (n-1) / \sum_{i=1}^n \log(x_i/\sigma)$ that is unbiased since it can be

shown that $\hat{\alpha}^{\text{MLE}} \overset{\mathcal{L}}{\sim} \text{IG}(n, n\alpha)$, where IG stands for the inverse gamma distribution. If σ is unknown and estimated by $\hat{\sigma} := x_{(1)}$, then it is rather $\hat{\gamma}^{\text{MLE-U}} := (n - 1)^{-1} \sum_{i=1}^n \log(x_i/\hat{\sigma})$ that is unbiased for γ since $\hat{\gamma}^{\text{MLE}} \overset{\mathcal{L}}{\sim} \Gamma(n - 1, \gamma/n)$ in this case. Finally, $\hat{\alpha}^{\text{MLE-U}} := (n - 2)/\sum_{i=1}^n \log(x_i/\hat{\sigma})$ is unbiased for α since it can be shown that $\hat{\alpha}^{\text{MLE}} \overset{\mathcal{L}}{\sim} \text{IG}(n - 1, n\alpha)$ when σ is estimated by $x_{(1)}$.

One way to reconcile these disagreements—due to the asymmetry in the distributions of $\hat{\alpha}$ and $\hat{\gamma}$ —is to make a compromise as done in [Brazauskas and Serfling \(2000a\)](#). Note that we use σ in the following equations, but recall that we replace σ with its estimate $x_{(1)}$ if it is unknown. It consists in dividing $\sum_{i=1}^n \log(x_i/\sigma)$ by $0.5\chi_{0.5;2(n-1_\sigma)}^2 \approx n - 1_\sigma - 1/3$ instead of n , where $\chi_{0.5;\nu}^2 \approx \nu - 2/3$ is the median of a chi-square distribution with ν degrees of freedom and 1_σ is an indicator function defined as

$$1_\sigma := \begin{cases} 0, & \text{if } \sigma \text{ is known;} \\ 1, & \text{if } \sigma \text{ is unknown and estimated by } x_{(1)}. \end{cases} \tag{8}$$

In doing so, we obtain a median-unbiased (MU) version of the MLE for both α and γ :

$$1/\hat{\alpha}^{\text{MLE-MU}} = \hat{\gamma}^{\text{MLE-MU}} := \frac{\sum_{i=1}^n \log(x_i/\sigma)}{0.5\chi_{0.5;2(n-1_\sigma)}^2} \approx \frac{\sum_{i=1}^n \log(x_i/\sigma)}{n - 1_\sigma - 1/3}. \tag{9}$$

Therefore, to achieve different bias corrections in our robust context, we modify $\sum_{i=1}^n \hat{\pi}_i = \hat{\omega}n$ (which represents the estimated number of Pareto observations) in the denominator of $\hat{\gamma}$ as follows:

$$\begin{aligned} \hat{\gamma}^{\text{P-FLLP-U}} &:= \frac{\sum_{i=1}^n \hat{\pi}_i \log(x_i/\sigma)}{\hat{\omega}n - 1_\sigma} && \text{for a correction of bias for } \gamma, \\ \hat{\alpha}^{\text{P-FLLP-U}} &:= \frac{\hat{\omega}n - 1_\sigma - 1}{\sum_{i=1}^n \hat{\pi}_i \log(x_i/\sigma)} && \text{for a correction of bias for } \alpha, \\ \frac{1}{\hat{\alpha}^{\text{P-FLLP-MU}}} &= \hat{\gamma}^{\text{P-FLLP-MU}} := \frac{\sum_{i=1}^n \hat{\pi}_i \log(x_i/\sigma)}{0.5\chi_{0.5;2(\hat{\omega}n-1_\sigma)}^2} \\ &\approx \frac{\sum_{i=1}^n \hat{\pi}_i \log(x_i/\sigma)}{\hat{\omega}n - 1_\sigma - 1/3} && \text{for a correction of median-bias.} \end{aligned} \tag{10}$$

These estimators are used in combination with $\hat{\omega}$ and $\hat{\pi}_1, \dots, \hat{\pi}_n$ in the iterative process described in [Definition 3.1](#).

3.2 Asymptotic properties of the P-FLLP estimators

We first consider asymptotic properties of P-FLLP estimators in [Section 3.2.1](#) in the sense that the upper outliers are infinite, for any fixed sample size n , which allows us to determine their breakdown point (BP). In [Section 3.2.2](#), we consider the classical nature of the asymptotic, that is, when $n \rightarrow \infty$. We study the convergence of the P-FLLP estimators, which allows us to determine the asymptotic BP for different distributions of the observations. Finally, we investigate the asymptotic behavior of the P-FLLP estimator of $\gamma := \alpha^{-1}$ when the true distribution is a contaminated Pareto distribution with different levels of contamination.

3.2.1 Behavior of the P-FLLP estimators in the presence of infinite upper outliers. We suppose that our sample x_1, \dots, x_n ($n \geq 3$) is divided into $k \geq 2$ fixed observations (say x_1, \dots, x_k) and $n - k$ observations (say x_{k+1}, \dots, x_n) which behave as infinite upper outliers, in the sense that $x_i \rightarrow \infty, i = k + 1, \dots, n$. The k fixed observations consist of non-outliers and possibly finite upper outliers. Note that the term ‘‘P-FLLP iterative process’’ refers below to the iterative process described in [Definition 3.1](#) for finding the P-FLLP estimators.

Proposition 3.1. *If $x_{k+1} \rightarrow \infty, \dots, x_n \rightarrow \infty$, the P-FLLP iterative process produces at least one robust solution such that $\hat{\gamma} = \hat{\alpha}^{-1} < \infty \Leftrightarrow \hat{\alpha} > 0$. Furthermore, $\hat{\gamma} = \hat{\alpha}^{-1} < \infty \Leftrightarrow \hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0$, that is, a robust solution such that $\hat{\gamma} = \hat{\alpha}^{-1} < \infty$ is necessarily a solution which completely rejects the $n - k$ infinite upper outliers, and vice versa.*

Proof. We first show that $\hat{\gamma} = \hat{\alpha}^{-1} < \infty \Rightarrow \hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0$. Suppose that the P-FLLP iterative process converged to a solution where $\hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0$ is not satisfied, that is, at least one infinite outlier is not completely rejected. Without loss of generality, suppose $\hat{\pi}_n > 0$. Then we have $\hat{\alpha}^{-1} := (\sum_{i=1}^n \hat{\pi}_i)^{-1} \sum_{i=1}^n \hat{\pi}_i \log(x_i/\sigma) \geq (\sum_{i=1}^n \hat{\pi}_i)^{-1} \hat{\pi}_n \log(x_n/\sigma) \rightarrow \infty$ since $\hat{\pi}_n > 0, x_n \rightarrow \infty, x_i/\sigma \geq 1$ and $0 \leq \hat{\pi}_i \leq 1$ for $i = 1, \dots, n$.

We now show that $\hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0 \Rightarrow \hat{\alpha}^{-1} < \infty$. Suppose we have $\hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0$, then

$$\hat{\alpha}^{-1} := \left(\sum_{i=1}^n \hat{\pi}_i \right)^{-1} \sum_{i=1}^n \hat{\pi}_i \log(x_i/\sigma) = \left(\sum_{i=1}^k \hat{\pi}_i \right)^{-1} \sum_{i=1}^k \hat{\pi}_i \log(x_i/\sigma) < \infty$$

since it is based on the fixed observations x_1, \dots, x_k . Note that σ is fixed, whether known or estimated by $x_{(1)}$, since the sample minimum is necessarily found among the k fixed observations.

Finally, we show that the P-FLLP iterative process generates at least one robust solution such that $\hat{\alpha}^{-1} < \infty$, or equivalently as we have just shown, such that $\hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0$. Suppose we have $\hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0$ at the initial step or at some other step in the iterative process. At the next iteration, we obtain $\hat{z}_i := (x_i/\sigma)^{\hat{\alpha}} \rightarrow \infty$ for $i = k+1, \dots, n$ since $x_i \rightarrow \infty, \sigma$ is fixed and $\hat{\alpha}^{-1} < \infty \Leftrightarrow \hat{\alpha} > 0$. We therefore always obtain $\hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0$, at this iteration and for all the following ones, since

$$\hat{\pi}_i := \frac{\hat{\omega} f_{\mathcal{P}}(x_i|\sigma, \hat{\alpha})}{f_{\text{P-FLLP}}(x_i|\hat{\omega}, \sigma, \hat{\alpha})} = \frac{\hat{\omega} f_{\mathcal{P}}(\hat{z}_i|1, 1)}{f_{\text{P-FLLP}}(\hat{z}_i|\hat{\omega}, 1, 1)}$$

and the right tail of the P-FLLP density is heavier than that of the Pareto, which implies that $\hat{\pi}_i \rightarrow 0$ as $\hat{z}_i \rightarrow \infty$. \square

Although Proposition 3.1 establishes that the P-FLLP iterative process generates at least one robust solution ($\hat{\gamma} = \hat{\alpha}^{-1} < \infty$) and that a robust solution completely rejects infinite outliers ($\hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0$), it is not guaranteed that one of them is such that $\hat{\omega} > 1/2$, that is, a valid candidate for the P-FLLP estimators. Note that a non-robust solution in this context of infinite outliers is a solution with $\hat{\gamma} \rightarrow \infty \Leftrightarrow \hat{\alpha} \rightarrow 0$.

Proposition 3.2. *If we have*

- $x_{k+1} \rightarrow \infty, \dots, x_n \rightarrow \infty$, and
- *there exists at least one solution satisfying $\hat{\omega} > 1/2$ among the robust solutions (those with $\hat{\gamma} = \hat{\alpha}^{-1} < \infty \Leftrightarrow \hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0$) generated by the P-FLLP iterative process (see Proposition 3.1),*

then the P-FLLP estimators (see Definition 3.1) are found among the robust solutions satisfying $\hat{\omega} > 1/2$, precisely the solution with the largest $\hat{\alpha}$.

Proof. According to Definition 3.1, in the case of multiple solutions, the P-FLLP estimators are defined as the solution with the largest value of $\hat{\alpha}$, provided that $\hat{\omega} > 1/2$. It is therefore sufficient to observe that $\hat{\alpha} \rightarrow 0$ for a non-robust solution and $\hat{\alpha} > 0$ for a robust solution, so the solution with the largest $\hat{\alpha}$ is necessarily among the robust ones. \square

To better interpret these propositions, recall that there are k fixed observations and $n - k$ infinite outliers. The proportion and number of outliers are respectively estimated by $1 - \hat{\omega}$ and $n - \hat{\omega}n$, which include the $n - k$ infinite outliers and hence the potential $k - \hat{\omega}n \geq 0$ finite upper outliers (among the k fixed observations). Therefore, the proportion and number of non-outliers (necessarily found among the k fixed observations) in the sample are respectively estimated by $\hat{\omega}$ and $\hat{\omega}n$.

Proposition 3.2 ensures that the P-FLLP estimators can handle robustly the $n - \hat{\omega}n$ outliers, that is, the $n - k$ infinite outliers and the $k - \hat{\omega}n$ finite outliers (if any), if they represent less than half of the sample, that is, if $n - \hat{\omega}n < n/2 \Leftrightarrow 1 - \hat{\omega} < 1/2$. Or, equivalently, if the estimated number of non-outliers constitutes the majority of the sample, that is, if $\hat{\omega}n > n/2 \Leftrightarrow \hat{\omega} > 1/2$.

Consider the special case where $k = \hat{\omega}n \Leftrightarrow \hat{\omega} = k/n$, which means that the estimated probability of finding finite upper outliers among the k fixed observations is 0, that is, $k - \hat{\omega}n = 0$. Since $\hat{\omega}n = \sum_{i=1}^n \hat{\pi}_i$ and $\hat{\pi}_{k+1} = 0, \dots, \hat{\pi}_n = 0$ for robust solutions, $\hat{\omega} = k/n$ is equivalent to $\hat{\pi}_1 = 1, \dots, \hat{\pi}_k = 1$. The condition $\hat{\omega} > 1/2$ in Proposition 3.2 is then equivalent to $k > n/2 \Leftrightarrow n - k < n/2$. In this case, we can determine the breakdown point (BP), which can be defined in our context as the largest proportion of infinite upper outliers that an estimator can handle robustly. Since the $n - k$ infinite outliers are rejected as long as they represent less than 50% of the sample, this corresponds to a BP of $n^{-1} \max\{m \in \mathbb{N} : m < n/2\}$ (for example, BP = 49/100 if $n = 100$), which tends to a BP of 50% when $n \rightarrow \infty$, the highest desired value.

Proposition 3.2 is also useful for determining the BP of the P-FLLP estimators when $n \rightarrow \infty$ and the distribution of the fixed observations is specified. First, we need to study the convergence of the P-FLLP estimators as $n \rightarrow \infty$.

3.2.2 Behavior of the P-FLLP estimators when $n \rightarrow \infty$. We take the analysis further by studying the asymptotic behavior of the P-FLLP estimators, where the asymptotic nature is now understood as the sample size n approaching infinity. Instead of directly assuming that $x_i \rightarrow \infty$ for $i = k + 1, \dots, n$ as in Section 3.2.1, suppose now that the observations x_1, \dots, x_n come from a density g defined on $[\sigma, \infty)$. The outlying nature of the observations x_i is then integrated into the model by their density g , for example, via a contaminated Pareto distribution. Given the well-known result $x_{(1)} \xrightarrow{\mathcal{P}} \sigma$ as $n \rightarrow \infty$, we assume in this section that σ is known.

Proposition 3.3. *Consider a random sample x_1, \dots, x_n generated from a density $g(x)$ defined on $[\sigma, \infty)$. Consider the iterative process leading to the robust P-FLLP estimators described in Definition 3.1, except that at iteration $(t + 1)$ the values of $\hat{\omega}^{(t)}$ and $\hat{\alpha}^{(t)}$ are replaced by known constants $\hat{\omega}^{*(t)}$ and $\hat{\alpha}^{*(t)}$, that is,*

$$\hat{\omega}^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \hat{\pi}_i^{(t)} \quad \text{and} \quad \hat{\delta}^{(t+1)} := \frac{\hat{\omega}^{(t+1)}}{\hat{\alpha}^{(t+1)}} = \frac{1}{n} \sum_{i=1}^n \hat{\pi}_i^{(t)} \log(x_i/\sigma),$$

with

$$\hat{\pi}_i^{(t)} = \frac{\hat{\omega}^{*(t)} f_{\mathcal{P}}(x_i|\sigma, \hat{\alpha}^{*(t)})}{f_{\text{P-FLLP}}(x_i|\hat{\omega}^{*(t)}, \sigma, \hat{\alpha}^{*(t)})} \quad \text{and} \quad \frac{1}{\hat{\alpha}^{(t+1)}} = \hat{\gamma}^{(t+1)} = \frac{\hat{\delta}^{(t+1)}}{\hat{\omega}^{(t+1)}}.$$

Then we have, as $n \rightarrow \infty$,

$$\hat{\omega}^{(t+1)} \xrightarrow{\mathcal{P}} \hat{\omega}^{*(t)} \int_{\sigma}^{\infty} \frac{f_{\mathcal{P}}(x|\sigma, \hat{\alpha}^{*(t)})}{f_{\text{P-FLLP}}(x|\hat{\omega}^{*(t)}, \sigma, \hat{\alpha}^{*(t)})} g(x) dx,$$

$$\begin{aligned} \hat{\delta}^{(t+1)} &\xrightarrow{\mathcal{P}} \hat{\omega}^{*(t)} \int_{\sigma}^{\infty} \frac{f_{\mathcal{P}}(x|\sigma, \hat{\alpha}^{*(t)})}{f_{\text{P-FLLP}}(x|\hat{\omega}^{*(t)}, \sigma, \hat{\alpha}^{*(t)})} \log(x/\sigma) g(x) dx, \\ \frac{1}{\hat{\alpha}^{(t+1)}} &= \hat{\gamma}^{(t+1)} \xrightarrow{\mathcal{P}} \frac{\int_{\sigma}^{\infty} \frac{f_{\mathcal{P}}(x|\sigma, \hat{\alpha}^{*(t)})}{f_{\text{P-FLLP}}(x|\hat{\omega}^{*(t)}, \sigma, \hat{\alpha}^{*(t)})} \log(x/\sigma) g(x) dx}{\int_{\sigma}^{\infty} \frac{f_{\mathcal{P}}(x|\sigma, \hat{\alpha}^{*(t)})}{f_{\text{P-FLLP}}(x|\hat{\omega}^{*(t)}, \sigma, \hat{\alpha}^{*(t)})} g(x) dx}. \end{aligned}$$

Proof. Since the values of $\hat{\omega}^{(t)}$ and $\hat{\alpha}^{(t)}$ are replaced by known constants $\hat{\omega}^{*(t)}$ and $\hat{\alpha}^{*(t)}$, the only random components of $\hat{\omega}^{(t+1)}$ and $\hat{\delta}^{(t+1)}$ are x_1, \dots, x_n , more precisely x_i in each term $\hat{\pi}_i^{(t)}$. Therefore, using the weak law of large numbers, we have, as $n \rightarrow \infty$,

$$\begin{aligned} \hat{\omega}^{(t+1)} &= \frac{1}{n} \sum_{i=1}^n \hat{\pi}_i^{(t)} \xrightarrow{\mathcal{P}} \mathbb{E}(\hat{\pi}_1^{(t)}) = \hat{\omega}^{*(t)} \int_{\sigma}^{\infty} \frac{f_{\mathcal{P}}(x|\sigma, \hat{\alpha}^{*(t)})}{f_{\text{P-FLLP}}(x|\hat{\omega}^{*(t)}, \sigma, \hat{\alpha}^{*(t)})} g(x) dx, \\ \hat{\delta}^{(t+1)} &= \frac{1}{n} \sum_{i=1}^n \hat{\pi}_i^{(t)} \log(x_i/\sigma) \xrightarrow{\mathcal{P}} \mathbb{E}(\hat{\pi}_1^{(t)} \log(X/\sigma)) \\ &= \hat{\omega}^{*(t)} \int_{\sigma}^{\infty} \frac{f_{\mathcal{P}}(x|\sigma, \hat{\alpha}^{*(t)})}{f_{\text{P-FLLP}}(x|\hat{\omega}^{*(t)}, \sigma, \hat{\alpha}^{*(t)})} \log(x/\sigma) g(x) dx, \end{aligned}$$

and the convergence of $1/\hat{\alpha}^{(t+1)} = \hat{\gamma}^{(t+1)}$ follows directly from Slutsky's theorem since $\hat{\gamma}^{(t+1)} = \hat{\delta}^{(t+1)}/\hat{\omega}^{(t+1)}$. \square

Using the idea of Proposition 3.3, we construct an algorithm for computing the asymptotic estimators of ω and α . This corresponds to the asymptotic version of the P-FLLP iterative process described in Definition 3.1, where the density $g(x)$ replaces the finite sample x_1, \dots, x_n .

Definition 3.2. The asymptotic robust P-FLLP estimators of ω (the expected proportion of Pareto observations) and α (the Pareto tail index), if the observations x_1, \dots, x_n (with $n \rightarrow \infty$) are assumed to come from a density g defined on $[\sigma, \infty)$, are denoted $\hat{\omega}^*$ and $\hat{\alpha}^* = 1/\hat{\gamma}^*$ and are the results of the iterative process given by

$$\begin{aligned} \hat{\omega}^{*(t+1)} &= \hat{\omega}^{*(t)} \int_{\sigma}^{\infty} \frac{f_{\mathcal{P}}(x|\sigma, \hat{\alpha}^{*(t)})}{f_{\text{P-FLLP}}(x|\hat{\omega}^{*(t)}, \sigma, \hat{\alpha}^{*(t)})} g(x) dx, \\ \hat{\delta}^{*(t+1)} &= \hat{\omega}^{*(t)} \int_{\sigma}^{\infty} \frac{f_{\mathcal{P}}(x|\sigma, \hat{\alpha}^{*(t)})}{f_{\text{P-FLLP}}(x|\hat{\omega}^{*(t)}, \sigma, \hat{\alpha}^{*(t)})} \log(x/\sigma) g(x) dx, \\ \frac{1}{\hat{\alpha}^{*(t+1)}} &= \hat{\gamma}^{*(t+1)} = \frac{\hat{\delta}^{*(t+1)}}{\hat{\omega}^{*(t+1)}}. \end{aligned}$$

When different solutions are found, only those with $\hat{\omega}^* > 1/2$ are considered. The asymptotic P-FLLP estimators are then defined as the solution with the largest value of $\hat{\alpha}^*$, or equivalently, with the smallest value of $\hat{\gamma}^*$. If σ is unknown, it is replaced by its estimate $x_{(1)}$.

Initial values $\hat{\omega}^{*(0)}$ and $\hat{\alpha}^{*(0)}$ must be provided. If we set $\hat{\omega}^{*(0)} = 1$, the convergence is immediate and we obtain $\hat{\omega}^* = 1$ and the asymptotic value of the MLE of α expressed in a functional form:

$$1/\hat{\alpha}^{*\text{MLE}} = \hat{\gamma}^{*\text{MLE}} = \int_{\sigma}^{\infty} \log(x/\sigma) g(x) dx = E_g(\log(X/\sigma)). \quad (11)$$

For the remainder of this section, we explore the behavior of the asymptotic P-FLLP estimators $\hat{\omega}^*$ and $\hat{\alpha}^* = 1/\hat{\gamma}^*$ for different densities g . First, consider that g is the P-FLLP(ω, σ ,

α) density defined on $[\sigma, \infty)$. Using the iterative process described in Definition 3.2, we find $\hat{\omega}^* = \omega$ and $\hat{\alpha}^* = \alpha$ for any values of ω, σ, α . Given that the P-FLLP distribution is the underlying model of the P-FLLP estimators, these results are not surprising, but reassuring nonetheless. In the particular case where $\omega = 1$, we have that g is the $\mathcal{P}(\sigma, \alpha)$ density. Then we find $\hat{\omega}^* = 1$ and $\hat{\alpha}^{*\text{MLE}} = [E_{\mathcal{P}(\sigma, \alpha)}(\log(X/\sigma))]^{-1} = \alpha$ using (11). Note that, as expected, if we instead use the P-FLLP estimators described in Definition 3.1 with a very large sample, say of size $n = 10,000$, the estimates $\hat{\omega}$ and $\hat{\alpha} = 1/\hat{\gamma}$ approximate very closely to the asymptotic P-FLLP estimators $\hat{\omega}^*$ and $\hat{\alpha}^* = 1/\hat{\gamma}^*$. As a result, $\hat{\alpha}^{\text{P-FLLP}}$ behaves asymptotically like $\hat{\alpha}^{\text{MLE}}$ under the Pareto model and therefore shares the same asymptotic properties.

Second, we study the behavior of the asymptotic P-FLLP estimators when g is the density of a contaminated Pareto, specifically the mixture $\xi \mathcal{P}(1, 1) + (1 - \xi)\delta_{x_0}$, where δ_{x_0} is the probability measure that places mass 1 on point x_0 . We consider six levels of contamination, namely $1 - \xi \in \{0.02, 0.05, 0.10, 0.20, 0.35, 0.48\}$. For each level of contamination, the asymptotic value of $\hat{\gamma}^* = 1/\hat{\alpha}^*$ is calculated for a set of values of x_0 covering the range $\log x_0 \in [\log 2, 15]$. The asymptotic P-FLLP estimator of $\gamma := \alpha^{-1}$ is compared to the asymptotic non-robust MLE in (11) and the robust probability integral transform statistic estimator (PITSE) introduced by Finkelstein, Tucker and Veeh (2006). For the latter, five tuning constants are considered to obtain asymptotic BP of 0.1, 0.2, 0.3, 0.4 and 0.5 and the asymptotic estimates are approximated with a sample size $n = 10,000$. Note that we also considered other robust estimators such as the optimal B-robust estimator (OBRE) proposed by Victoria-Feser and Ronchetti (1994), the generalized median estimator (GME) proposed by Brazauskas and Serfling (2000a) or the trimmed mean estimator (TME) introduced by Kimber (1983). However, they are not presented here since they exhibit the same behavior as the PITSE. Furthermore, Brzezinski (2016) recommended in their Monte Carlo simulation comparisons the use of the PITSE for its simplicity and good performance in terms of efficiency and robustness.

The results are presented in Figure 2. We observe that the asymptotic P-FLLP estimator of $\gamma := \alpha^{-1}$ shows the same full robustness for all contamination levels. As long as the value of x_0 agrees with a Pareto distribution, no outliers are detected and $\hat{\gamma}^{*\text{P-FLLP}} = \hat{\gamma}^{*\text{MLE}}$. When x_0 exceeds a certain threshold and conflicts with the Pareto model, then the influence of x_0 on $\hat{\gamma}^{*\text{P-FLLP}}$ gradually decreases to almost nothing as x_0 increases. In comparison, the PITSE offers partial robustness, in that the influence of outliers is limited but not completely eliminated, provided that the right tuning constant is chosen beforehand to match the level of contamination. Recall that, for the PITSE and other robust estimators in general, an increase in robustness is associated with a decrease in efficiency (performance in the absence of outliers). As for the P-FLLP estimators, the parameter ω acts as a tuning constant that determines the level of robustness, but its value is estimated by $\hat{\omega}^{*\text{P-FLLP}}$ based on the level of contamination of each sample. Moreover, since the P-FLLP estimator of $\gamma := \alpha^{-1}$ behaves asymptotically like the MLE under the Pareto model, it achieves full asymptotic efficiency.

We observe in Figure 2 that the asymptotic P-FLLP estimators can robustly handle a proportion of $1 - \xi = 48\%$ “infinite” outliers represented by $x_0 = e^{15}$, when the other 52% observations have a $\mathcal{P}(1, 1)$ distribution. If we take the analysis a little further, we find that the asymptotic P-FLLP estimators can robustly handle up to a proportion of 48.54% infinite outliers, which means a BP of 48.54% when the “fixed” observations come from a Pareto distribution. Specifically, we obtain $\hat{\omega}^{*\text{P-FLLP}} = 0.50$, which means 50% non-outliers to compensate for 48.54% infinite outliers and a remaining proportion of 1.46% finite outliers resulting from the upper values generated by the heavy-tailed Pareto distribution. It is possible to achieve a BP of 50% if the Pareto component in the mixture g is replaced by a distribution with a lighter right tail such as the half-normal distribution.

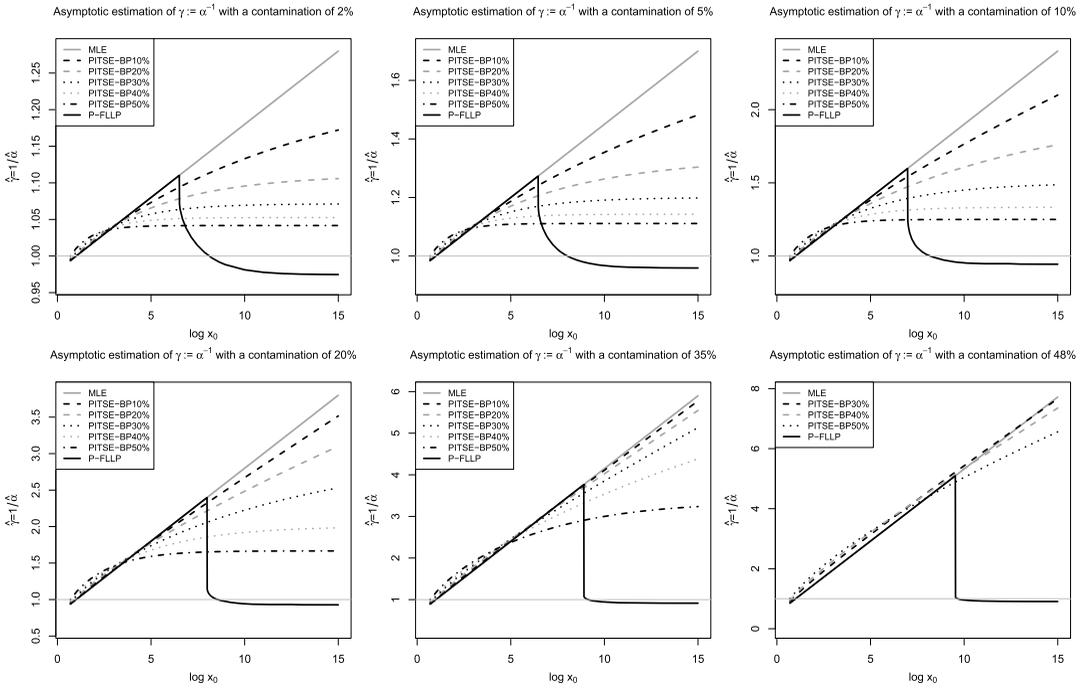


Figure 2 Asymptotic estimation of $\gamma := \alpha^{-1}$, when the true distribution is the mixture $\xi \mathcal{P}(1, 1) + (1 - \xi)\delta_{x_0}$, with six levels of contamination: $1 - \xi \in \{0.02, 0.05, 0.10, 0.20, 0.35, 0.48\}$.

3.3 Comparison of the P-FLLP estimator of the Pareto tail index with M -estimators

The family of M -estimators, in our context, can be defined as the solution in α of the equation $\sum_{i=1}^n \psi_\alpha(x_i) = 0$ for a given ψ -function $\psi_\alpha(x)$. Given that

$$\hat{\alpha}^{-1} = \frac{\sum_{i=1}^n \hat{\pi}_i \log(x_i/\sigma)}{\sum_{i=1}^n \hat{\pi}_i} \Leftrightarrow \sum_{i=1}^n \hat{\pi}_i (\log(x_i/\sigma) - 1/\hat{\alpha}) = 0,$$

we can define the ψ -function for the P-FLLP estimator of α as

$$\psi_\alpha(x) := \frac{\omega f_{\mathcal{P}}(x|\sigma, \alpha)}{f_{\text{P-FLLP}}(x|\omega, \sigma, \alpha)} (\log(x/\sigma) - 1/\alpha),$$

if the true value of ω is assumed to be known ($\hat{\omega}$ depends on the sample x_1, \dots, x_n and $\psi_\alpha(x)$ must be defined as a function of x only). In particular, if $\omega = 1$, the ψ -function for the MLE is given by

$$\psi_\alpha(x) = \log(x/\sigma) - 1/\alpha,$$

which is the negative of the score function $s_\alpha(x) := \frac{\partial}{\partial \alpha} \log f_{\mathcal{P}}(x|\sigma, \alpha)$. Note that $\psi_\alpha(x) = s_\alpha(x)$ would also be a valid ψ -function for the MLE since any multiplicative constant with respect to x can be added to $\psi_\alpha(x)$.

Therefore, like the WMLE, the P-FLLP estimator of α can be thought of as an M -estimator with a ψ -function based on the weighted score function, where the weights depend on a constant ω representing the proportion of non-outliers. However, we go one step further by estimating ω in the iterative process, which largely explains the simultaneous efficiency and robustness of $\hat{\alpha}^{\text{P-FLLP}}$.

The ψ -functions are plotted in Figure 3 for $\hat{\alpha}^{\text{MLE}}$ and $\hat{\alpha}^{\text{P-FLLP}}$ with $\sigma = 1, \alpha = 1$ and ω set to 0.50, 0.95 and 0.99. We observe that the P-FLLP estimators (for any fixed value of $\omega < 1$) are re-descending, meaning that $\psi_\alpha(x) \rightarrow 0$ as $x \rightarrow \infty$.

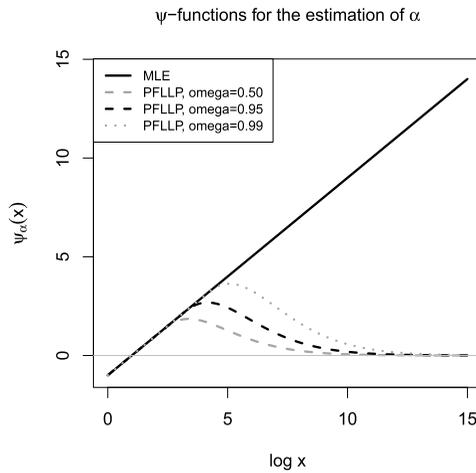


Figure 3 The ψ -functions for $\hat{\alpha}^{MLE}$ and $\hat{\alpha}^{P-FLLP}$, with $\sigma = 1$, $\alpha = 1$ and ω set to 0.50, 0.95 and 0.99.

3.4 Robust confidence interval for the Pareto tail index

We take inference a little further in this section with a robust confidence interval for the Pareto tail index α . We construct the confidence interval with the same approach as that used for the robust estimators. If the v_i 's were known, we could accurately distinguish non-outliers ($v_i = 1$) from outliers ($v_i = 0$) and proceed with the usual technique on the non-outliers. Since they are unknown, the idea is to estimate each v_i by their posterior mean $\hat{\pi}_i$ and to base the confidence interval on the observations weighted by the $\hat{\pi}_i$.

We first find the usual confidence interval on the Pareto observations assuming that the vector \mathbf{v} as defined in (6) is known (recall that $v_i = 1$ if x_i comes from the Pareto component, and $v_i = 0$ otherwise). It is well known that if $x_1, \dots, x_n \stackrel{\mathcal{L}}{\sim} \mathcal{P}(\sigma, \alpha)$, then $2\alpha \sum_{i=1}^n \log(x_i/\sigma) \stackrel{\mathcal{L}}{\sim} \chi_{2n}^2$ and $2\alpha \sum_{i=1}^n \log(x_i/x_{(1)}) \stackrel{\mathcal{L}}{\sim} \chi_{2(n-1)}^2$. Note that we use σ in the following equations, but recall that we replace σ with its estimate $x_{(1)}$ if it is unknown. Therefore, if $x_i | (v_i = 1) \stackrel{\mathcal{L}}{\sim} \mathcal{P}(\sigma, \alpha)$ for $i = 1, \dots, n$, then we have

$$2\alpha \sum_{i=1}^n v_i \log(x_i/\sigma) \Big| \mathbf{v} \stackrel{\mathcal{L}}{\sim} \chi_{2(\sum_{i=1}^n v_i - 1\sigma)}^2,$$

where 1_σ is the indicator function defined in (8). Therefore, given that \mathbf{v} is known, a $100(1 - a)\%$ confidence interval for α is given by

$$\frac{0.5\chi_{a/2; 2(\sum_{i=1}^n v_i - 1\sigma)}^2}{\sum_{i=1}^n v_i \log(x_i/\sigma)} \leq \alpha \leq \frac{0.5\chi_{1-a/2; 2(\sum_{i=1}^n v_i - 1\sigma)}^2}{\sum_{i=1}^n v_i \log(x_i/\sigma)},$$

where $\chi_{p;v}^2$ is the p -th percentile of a χ_v^2 distribution. If \mathbf{v} were known, the true coverage probability of this confidence interval (assuming the Pareto model for non-outliers) would be equal to the nominal coverage probability, for any sample size.

In a second step, the latent variables v_i are estimated by $\hat{\pi}_i$, which represents the estimated probability that the observation x_i comes from the Pareto component. Given that $\sum_{i=1}^n \hat{\pi}_i = \hat{\omega}n$, we obtain this robust P-FLLP $100(1 - a)\%$ confidence interval for α :

$$\frac{0.5\chi_{a/2; 2(\hat{\omega}n - 1\sigma)}^2}{\sum_{i=1}^n \hat{\pi}_i \log(x_i/\sigma)} \leq \alpha \leq \frac{0.5\chi_{1-a/2; 2(\hat{\omega}n - 1\sigma)}^2}{\sum_{i=1}^n \hat{\pi}_i \log(x_i/\sigma)}. \tag{12}$$

It can be useful to express (12) using $\hat{\alpha} \equiv \hat{\alpha}^{\text{P-FLLP}}$ as defined in (3.1) or using $\hat{\alpha}^{\text{P-FLLP-MU}}$ as defined in (10). We obtain

$$\begin{aligned} \frac{\chi_{a/2; 2(\hat{\omega}n-1\sigma)}^2}{2\hat{\omega}n} \hat{\alpha}^{\text{P-FLLP}} &\leq \alpha \leq \frac{\chi_{1-a/2; 2(\hat{\omega}n-1\sigma)}^2}{2\hat{\omega}n} \hat{\alpha}^{\text{P-FLLP}}, \\ \frac{\chi_{a/2; 2(\hat{\omega}n-1\sigma)}^2}{\chi_{0.5; 2(\hat{\omega}n-1\sigma)}^2} \hat{\alpha}^{\text{P-FLLP-MU}} &\leq \alpha \leq \frac{\chi_{1-a/2; 2(\hat{\omega}n-1\sigma)}^2}{\chi_{0.5; 2(\hat{\omega}n-1\sigma)}^2} \hat{\alpha}^{\text{P-FLLP-MU}}. \end{aligned} \quad (13)$$

The P-FLLP confidence interval for $\gamma := 1/\alpha$ is found by inverting the bounds. The true coverage probability of this robust confidence interval under the Pareto model is investigated by Monte Carlo simulations in Section 4 for a nominal coverage probability of 95% and different sample sizes.

4 Comparison of MLE, PITSE and P-FLLP estimators using Monte Carlo simulations

We now investigate the performance of the P-FLLP estimator of the Pareto tail index for finite samples using Monte Carlo simulations. The performance in the absence of outliers is examined in Section 4.1, where efficiency and coverage probability of confidence intervals are established for different sample sizes. Robustness to outliers is studied in Section 4.2, where the P-FLLP estimator of α is compared to the MLE and PITSE. Efficiency versus robustness is analyzed in Section 4.3. The detailed programming code using R software is provided in the Supplementary Material (Desgagné (2024)).

4.1 Performance in the absence of outliers

We first compare the performance of the P-FLLP estimator of the Pareto tail index α to the MLE benchmark when the model to be estimated is $\mathcal{P}(\sigma, \alpha)$ as defined in (1) and the true model from which the observations $\mathbf{x} \equiv (x_1, \dots, x_n)$ were generated is $\mathcal{P}(\sigma_0, \alpha_0)$ —the uncontaminated model. The scale parameter σ , treated as a nuisance parameter, is estimated by the sample minimum $x_{(1)}$ and we compare the median-unbiased estimators $\hat{\alpha}^{\text{MLE-MU}}$ and $\hat{\alpha}^{\text{P-FLLP-MU}}$ as defined in (9) and (10), for the sample sizes $n = 50, 100, 200, 500$ and 1000. Note that the robust competitors compared in the comprehensive Monte Carlo study of Brzezinski (2016), namely optimal B-robust estimator (OBRE), weighted maximum likelihood estimator (WMLE), generalized median estimator (GME), probability integral transform statistic estimator (PITSE) and trimmed mean estimator (TME), depend on a tuning constant. Therefore, any level of efficiency can be achieved in the absence of outliers (always making a trade-off for the level of robustness), so there is no need to include these estimators for this analysis under the uncontaminated model.

The performance of a given estimator $\hat{\alpha} := 1/\hat{\gamma}$ is measured in this paper by

$$\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|) = \mathbb{E}(|\log(\hat{\gamma}/\gamma_0)|), \quad (14)$$

where $\gamma_0 := 1/\alpha_0$ and the expectation is taken with respect to the random sample. We can show that the corrected MLE which minimizes (14) is given precisely by the median-unbiased benchmark $\hat{\alpha}^{\text{MLE-MU}} := 1/\hat{\gamma}^{\text{MLE-MU}}$. Moreover, this measure is invariant whether we study the performance of the Pareto tail index α or the index $\gamma := 1/\alpha$.

In the absence of outliers, there is a price to pay for using robust estimators instead of the MLE benchmark. We measure this cost for a given estimator $\hat{\alpha}$ by the relative efficiency (RE) that we define as

$$\text{RE} := \left(\frac{\mathbb{E}(|\log(\hat{\alpha}^{\text{MLE-MU}}/\alpha_0)|)}{\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|)} \right)^2. \quad (15)$$

Table 1 Relative efficiency of the P-FLLP estimator of α (to the MLE) for the uncontaminated model

Sample size	50	100	200	500	1000	∞
Relative efficiency	0.932	0.948	0.961	0.975	0.984	1

Table 2 Probability in %, under the uncontaminated model, that the P-FLLP estimate of α is identical to that of the MLE, for different sample sizes

Sample size	50	100	200	500	1000
Probability in %	91.0	89.1	87.2	85.1	84.0

Note that the square of the ratio is used to obtain results very similar to the relative efficiency computed with mean squared errors, that is $\mathbb{E}((\hat{\alpha}^{\text{MLE-MU}} - \alpha_0)^2) / \mathbb{E}((\hat{\alpha} - \alpha_0)^2)$, as is often done in the literature.

We can verify that for most known estimators of α (including MLE, P-FLLP and PITSE), the choice of the parameters σ_0 and α_0 in the true model $\mathcal{P}(\sigma_0, \alpha_0)$ has no impact on the value of $\hat{\alpha}/\alpha_0 = (\hat{\gamma}/\gamma_0)^{-1}$, whether σ_0 is known or estimated. As a result, we can fix $\sigma_0 = \alpha_0 = 1$ without loss of generality in the simulations.

To study the performance under the uncontaminated model, 100,000 random samples were simulated using $x_1, \dots, x_n \stackrel{\mathcal{L}}{\sim} \mathcal{P}(\sigma_0, \alpha_0)$, for each of the sample sizes $n = 50, 100, 200, 500$ and 1000. For each sample generated, $|\log(\hat{\alpha}^{\text{MLE-MU}}/\alpha_0)|$ and $|\log(\hat{\alpha}^{\text{P-FLLP-MU}}/\alpha_0)|$ were computed and their expectation were estimated by the average of the corresponding 100,000 simulated distances, for a given sample size. The estimated RE of $\hat{\alpha}^{\text{P-FLLP-MU}}$ are given in Table 1. We observe that the RE of 0.932 is already high at $n = 50$ and that it increases with the size of the sample to approach 1 as $n \rightarrow \infty$ for full asymptotic efficiency.

In Table 2, we have estimated using 100,000 Monte Carlo simulations the probability that $\hat{\alpha}^{\text{P-FLLP-MU}}$ is equal to $\hat{\alpha}^{\text{MLE-MU}}$, or equivalently that $\hat{\omega}^{\text{P-FLLP-MU}} = 1$, under the uncontaminated model and for sample sizes n of 50, 100, 200, 500 and 1000. Again, note that the choice of the parameters σ_0 and α_0 in the true model $\mathcal{P}(\sigma_0, \alpha_0)$ has no impact on these probabilities. We observe high probabilities ranging from 0.84 to 0.91, which means that most of the time, the P-FLLP estimators generate the MLE benchmark under the Pareto model. This largely explains the high efficiency of the P-FLLP estimator. Moreover, this rare feature for a robust method represents an asset for practitioners familiar with MLE.

We also investigated the true coverage probability and the average length (divided by α_0) of the confidence intervals (CI) for the Pareto tail index α , for a nominal coverage probability of $1 - a = 95\%$ and under the uncontaminated $\mathcal{P}(\sigma_0, \alpha_0)$ model. To this end, 100,000 random samples were simulated under the $\mathcal{P}(\sigma_0, \alpha_0)$ model for each of the sample sizes $n = 50, 100, 200, 500$ and 1000. For each sample generated, 95% CI using the P-FLLP and MLE estimates were computed using (13) (it suffices to fix $\hat{\pi}_1 = 1, \dots, \hat{\pi}_n = 1, \hat{\omega} = 1$ to obtain MLE CI). The true coverage probability was estimated by the proportion of the 100,000 CI that included α_0 , for a given sample size. The results are given in Table 3. The true coverage probabilities are very close to 95% for the P-FLLP estimator, starting at 94.3% for $n = 50$ and increasing toward 0.95 as $n \rightarrow \infty$. We also found that the average length (divided by α_0) of the P-FLLP CIs are slightly larger than the MLE CIs, and the difference decreases toward 0 as n increases. This demonstrates on the one hand that the robust P-FLLP confidence intervals defined in (12) and (13) are well constructed and on the other hand that the price to pay compare to the MLE CI is low.

Table 3 True coverage probabilities (in %) and average length (divided by α_0) of CI for α , under the uncontaminated model $\mathcal{P}(\sigma_0, \alpha_0)$, for a nominal coverage probability of 95%

Sample size	50	100	200	500	1000
MLE true coverage	95.0	95.0	95.0	95.0	95.0
P-FLLP true coverage	94.3	94.3	94.4	94.7	94.8
Average length of the MLE CI	0.571	0.398	0.279	0.176	0.124
Average length of the P-FLLP CI	0.578	0.401	0.280	0.176	0.124

4.2 Performance in the presence of outliers

We now study the performance of the robust P-FLLP estimator when the model to be estimated is $\mathcal{P}(\sigma, \alpha)$ and the true model from which the observations $\mathbf{x} \equiv (x_1, \dots, x_n)$ were generated is contaminated. We included the non-robust MLE in our study as a benchmark in terms of efficiency. Regarding the robust competitors, Brzezinski (2016) compared in a comprehensive Monte Carlo analysis the best robust estimators, namely optimal B-robust estimator (OBRE), weighted maximum likelihood estimator (WMLE), generalized median estimator (GME), probability integral transform statistic estimator (PITSE) and trimmed mean estimator (TME). They concluded that the robustness performance of the best estimators, for a given level of efficiency, was quite similar. They recommended PITSE for its simplicity and good performance in terms of robustness. Since these estimators exhibit the same behavior against outliers, which is largely explained by the fact that they all depend on a tuning constant, we included in our study PITSE with five tuning constants chosen to obtain asymptotic breakdown points (BP) of 10%, 20%, 30%, 40%, 50%, which corresponds respectively to asymptotic relative efficiencies (ARE) of 0.99, 0.96, 0.91, 0.84, 0.75. At one extreme we find a PITSE estimator that is robust (BP = 50%) but not very efficient (ARE = 0.75) and at the other extreme we find a PITSE estimator that is barely robust (BP = 10%) but very efficient (ARE = 0.99). The comparison of PITSE and the P-FLLP estimator is delicate because the first depends on a tuning constant chosen by the user which fixes the efficiency/robustness compromise and the second depends only on the observations (its implicit tuning constant ω is estimated for each dataset) so its ARE = 1 and its asymptotic BP $\approx 50\%$ are fixed. For traditional estimators with a tuning constant, the robustness is generally compared for the same level of efficiency, which is however not possible when the comparison is made with an estimator whose tuning constant is fixed automatically by the data, such as the P-FLLP estimator. It will therefore be important in our comparison to consider the performance of robustness in relation to the level of efficiency. This efficiency-robustness trade-off is analyzed using the premium-protection approach in Section 4.3

We now compare the performance of the P-FLLP estimator with the MLE benchmark and the robust competitor PITSE when the model to be estimated is $\mathcal{P}(\sigma, \alpha)$ and the true model from which the observations $\mathbf{x} \equiv (x_1, \dots, x_n)$ were generated is a $\mathcal{P}(\sigma_0, \alpha_0)$ contaminated with a $\mathcal{P}(\sigma_1^*, \alpha_0)$, $\sigma_1^* > \sigma_0$. Specifically,

- a fixed proportion ξ of the observations is generated from a $\mathcal{P}(\sigma_0, \alpha_0)$, and
- a fixed proportion $1 - \xi$ is generated from the contaminating component modeled by a $\mathcal{P}(\sigma_1^*, \alpha_0)$, with $\sigma_1^* = \sigma_0 \sigma_1^{1/\alpha_0}$ and $\sigma_1^* > \sigma_0 \Leftrightarrow \sigma_1 > 1$.

Note that if we set $\mathbf{y} := (\mathbf{x}/\sigma_0)^{\alpha_0}$ to obtain the standardized case, this is equivalent to simulating a proportion ξ of the observations \mathbf{y} from a $\mathcal{P}(1, 1)$ and a proportion $1 - \xi$ from a $\mathcal{P}(\sigma_1, 1)$, $\sigma_1 > 1$. Here again, the choice of the parameters σ_0 and α_0 in the true contaminated model has no impact on the value of $\hat{\alpha}/\alpha_0 = (\hat{\gamma}/\gamma_0)^{-1}$, whether σ_0 is known or estimated. As a result, we can fix $\sigma_0 = \alpha_0 = 1$ without loss of generality in the simulations.

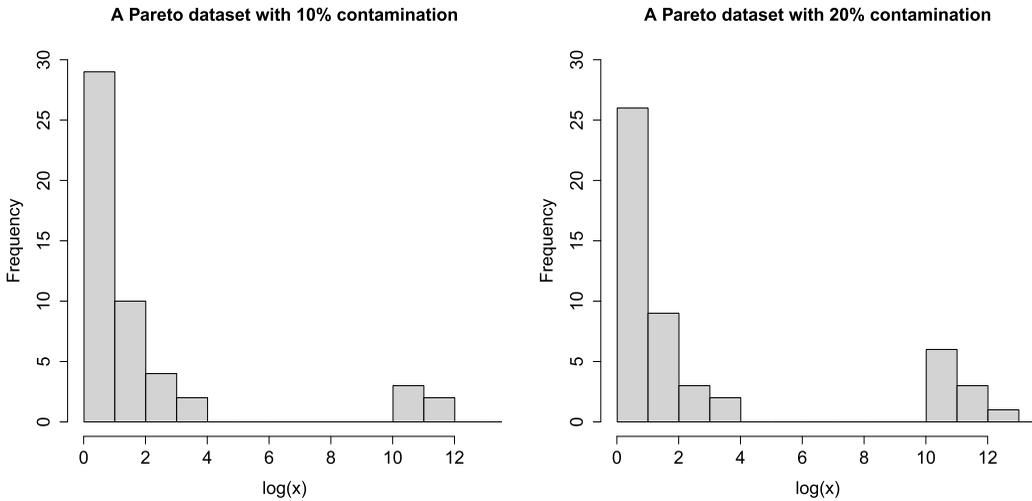


Figure 4 Example of two datasets of size $n = 50$, where 5 outliers (left graph) and 10 outliers (right graph) were generated from a $\mathcal{P}(e^{10}, 1)$. Note that the abscissa is on a logarithmic scale.

The sample size is set to $n = 50$. Other sample sizes such as $n = 100$ were considered, but results were similar. Note that infinite sample size is studied in Section 3.2.2, especially in Figure 2. We consider six levels of contamination, namely $1 - \xi \in \{0.04, 0.06, 0.10, 0.20, 0.30, 0.40\}$, which corresponds to 2, 3, 5, 10, 15 and 20 observations for a sample of size $n = 50$. The first four levels, from 4% to 20%, represent in practice a spectrum of light, moderate and severe contamination. The extreme levels of 30% and 40%, of more theoretical interest, allows us to evaluate the BP of an estimator, which we wish to be as high as possible and ideally close to 50%. The scale parameter σ is estimated by the sample minimum $x_{(1)}$ and we compare the median-unbiased estimators $\hat{\alpha}^{\text{P-FLLP-MU}}$ and $\hat{\alpha}^{\text{MLE-MU}}$ with the competitor PITSE. The performance of a given estimator $\hat{\alpha} := 1/\hat{\gamma}$ is again measured by $\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|) = \mathbb{E}(|\log(\hat{\gamma}/\gamma_0)|)$.

Figure 4 shows an example of two contaminated datasets of size $n = 50$. In the left graph, 45 observations ($\xi = 0.90$) were generated from a $\mathcal{P}(1, 1)$ and 5 outliers ($1 - \xi = 0.10$) were generated from a $\mathcal{P}(\log \sigma_1 = 10, 1)$. In the right graph, the ratio is instead $\xi = 80\%$ non-outliers (40 observations) and $1 - \xi = 20\%$ outliers (10 observations).

We must set the parameter $\sigma_1 > 1$, which represents the left bound of the domain of the standardized contamination component $\mathcal{P}(\sigma_1, 1)$, which amounts to selecting the parameter σ_1^* in the contamination component $\mathcal{P}(\sigma_1^*, \alpha_0)$. Each value of σ_1 represents a different scenario. Instead of choosing a few rather arbitrary values, as is generally the case in the literature, we go further by choosing, for all practical purposes, all the values of σ_1 . To achieve this, an interval $\log \sigma_1 \in [\log 2, B]$ is selected for each of the 6 contamination levels, so that for any $\log \sigma_1$ beyond B , the influence of outliers on the robust estimators becomes stable. We selected $B = 15$ for the four lowest contamination levels and $B = 18$ for the other two. The interval $[\log 2, B]$ is then divided equally into 30 values of $\log \sigma_1$ to obtain a smooth performance curve. This represents 30 different specific scenarios for each of the 6 contamination levels, for a total of 180 scenarios. For each of these 180 scenarios, the measure $\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|)$ is estimated for each estimator in the study using 2000 Monte Carlo simulations.

The results are shown in Figure 5, in which a separate graph is presented for each level of contamination. In each graph, we find 7 curves representing the estimators P-FLLP, MLE and PITSE with five tuning constants $t \in (1/9, 1/4, 3/7, 2/3, 1)$ chosen to obtain asymptotic BP of 10%, 20%, 30%, 40%, 50% respectively, which corresponds to ARE of 0.99, 0.96,

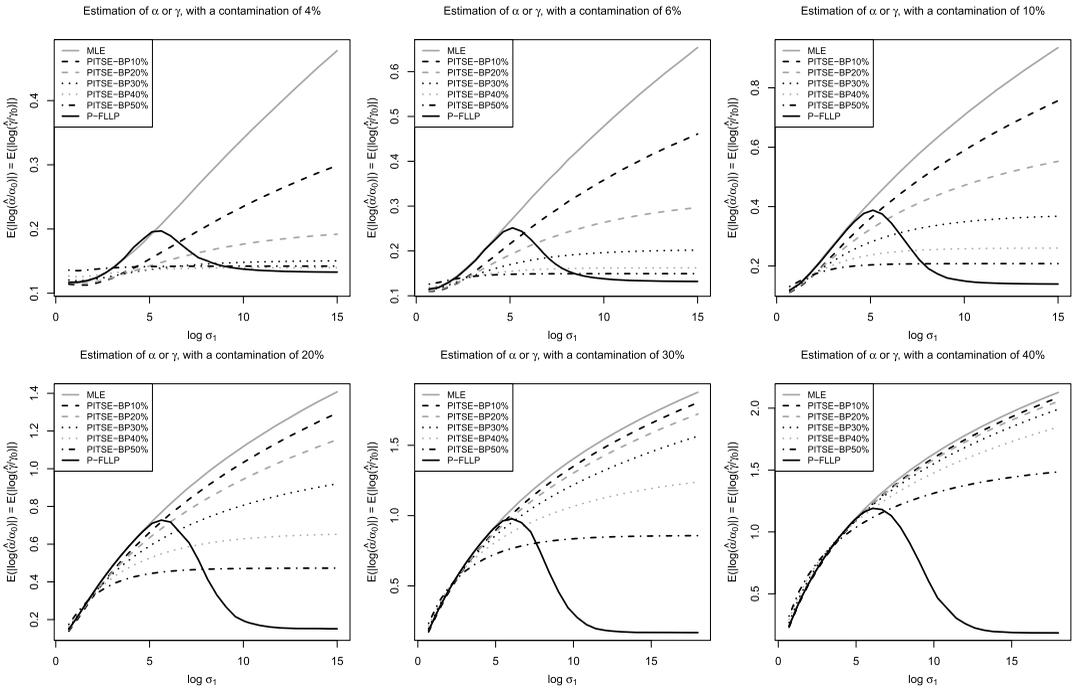


Figure 5 Performance of different estimators of α (or γ) for six levels of contamination.

Table 4 Average distances for the estimation of α or γ . The best value for each row appears in bold

Contamination	P-FLLP	PITSE					MLE
		BP50%	BP40%	BP30%	BP20%	BP10%	
4%	0.15	0.14	0.14	0.14	0.16	0.20	0.28
6%	0.16	0.15	0.15	0.18	0.22	0.29	0.38
10%	0.21	0.20	0.24	0.30	0.39	0.47	0.56
20%	0.36	0.43	0.54	0.67	0.77	0.83	0.90
30%	0.44	0.75	0.95	1.09	1.17	1.20	1.25
40%	0.57	1.18	1.33	1.39	1.42	1.43	1.46
Avg. 4% to 10%	0.17	0.16	0.18	0.21	0.26	0.32	0.41
Rank	2	1	3	4	5	6	7
Avg. 4% to 20%	0.22	0.23	0.27	0.32	0.38	0.45	0.53
Rank	1	2	3	4	5	6	7
Avg. 4% to 40%	0.31	0.47	0.56	0.63	0.69	0.74	0.81
Rank	1	2	3	4	5	6	7

0.91, 0.84, 0.75. The values of $\log \sigma_1$ lie in the x-axis, and the measure $\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|) = \mathbb{E}(|\log(\hat{\gamma}/\gamma_0)|)$ lies in the y-axis. Note that the upper bound of the y-axis was chosen such that the MLE line finishes at the top-right corner to consistently visualize the performance of each estimator relative to the MLE.

The results are also summarized in Table 4 by calculating, for each contamination level and each estimator, the average of the distances $\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|)$ computed at the 30 different values of $\log \sigma_1$. This average distance can be interpreted as a measure of the distance between a curve in a graph and the horizontal line at the origin. Equivalently, this can be interpreted as the measure $\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|)$ when a proportion ξ of the observations is gener-

ated from a $\mathcal{P}(\sigma_0, \alpha_0)$ and a proportion $1 - \xi$ is generated from a $\mathcal{P}(\sigma_0\sigma_1^{1/\alpha_0}, \alpha_0)$, with $\log \sigma_1$ randomly selected from the interval $[\log 2, B]$. The smaller this average distance, the better. The six rows in Table 4 labeled 4% to 40% correspond to the six graphs, respectively. The best performance for each row appears in bold. Overall scores are assigned to each estimator by calculating the average of the distance averages for (i) the 4%, 6% and 10% contamination scenarios, (ii) for the 4%, 6%, 10% and 20% contamination scenarios and (iii) for all contamination scenarios ranging from 4% to 40%. The seven estimators are finally ranked according to these overall scores.

Let us briefly analyze the robustness results only, without considering the efficiency of the estimators (which is done in Section 4.3). As expected, for all contamination scenarios, MLE has the worst robustness performance and that of PITSE increases with its BP. Looking at the graphs in Figure 5, it can be seen that the P-FLLP estimator is the only one to show the same complete robustness for all contamination levels. As long as the value of σ_1 is below a certain threshold, the observations of the contaminating component are in agreement with the uncontaminated component $\mathcal{P}(\sigma_0, \alpha_0)$ and no outliers are detected by the P-FLLP estimator. Then, an increasing conflict arises between the observations of the two different components as the value of σ_1 increases. The P-FLLP estimator manages this conflict by gradually decreasing the influence of outliers to a point where they are virtually excluded. Note that the graphs in Figure 5 for the case $n = 50$ are very similar to those in Figure 2 for the case $n \rightarrow \infty$, which confirms that our results are relevant for all sample sizes. If we look at Table 4, we observe that for the low contamination scenario of 4%, P-FLLP and PITSE BP20% to BP50% have the best performance, with their average distances ranging from 0.14 to 0.16. For the 6% contamination scenario, P-FLLP and PITSE BP30% to BP50% have the best performance, with their average distances ranging from 0.15 to 0.18. For the 10% contamination scenario, P-FLLP, PITSE BP40% and PITSE BP50% have the best performance, with their average distances ranging from 0.20 to 0.24. For the 20%, 30,% and 40% contamination scenarios, the P-FLLP estimator clearly has the best performance, with respective average distances of 0.36, 0.44 and 0.57.

4.3 Efficiency-robustness trade-off using the premium-protection approach

In this section, we use the premium-protection approach of Anscombe (1960) to investigate the efficiency-robustness trade-off. As in the previous section, we set the sample size to $n = 50$. In the absence of outliers, the price of using a robust estimator instead of the MLE benchmark is called the premium. For a given estimator $\hat{\alpha}$, we define the premium as

$$\text{Premium}(\hat{\alpha}) := \frac{\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|) - \mathbb{E}(|\log(\hat{\alpha}^{\text{MLE-MU}}/\alpha_0)|)}{\mathbb{E}(|\log(\hat{\alpha}^{\text{MLE-MU}}/\alpha_0)|)},$$

where the observations come from the uncontaminated model $\mathcal{P}(\sigma_0, \alpha_0)$. The premiums, estimated using 100,000 Monte Carlo simulations for each of the seven estimators in the study, are given in Table 5. The distances $\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|)$ and RE are also given, where RE is defined in (15). The premium and the RE are two equivalent measures since $\text{Premium}(\hat{\alpha}) =$

Table 5 Distances $\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|)$, RE and premiums of different estimators of α , for a sample size of $n = 50$

	MLE	P-FLLP	PITSE				
			BP10%	BP20%	BP30%	BP40%	BP50%
$\mathbb{E}(\log(\hat{\alpha}/\alpha_0))$	0.1141	0.1182	0.1168	0.1187	0.1223	0.1279	0.1363
RE	1	0.932	0.954	0.924	0.871	0.796	0.700
Premium	0	3.6%	2.4%	4.1%	7.1%	12.1%	19.5%

$\text{RE}^{-1/2} - 1$. For example, using the robust P-FLLP estimator instead of MLE increases the distance $\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|)$ from 0.1141 to 0.1182, under the pure $\mathcal{P}(\sigma_0, \alpha_0)$ model. This increase of 0.0041 represents 3.6% of the MLE distance of 0.1141, or a 3.6% premium to pay for using this robust estimator.

In the presence of outliers, the gain provided by the use of a robust estimator instead of the MLE is called the protection. For a given contamination level, we define the protection of a given estimator $\hat{\alpha}$ as

$$\text{Protection}(\hat{\alpha}) := \frac{\mathbb{E}(|\log(\hat{\alpha}^{\text{MLE-MU}}/\alpha_0)|) - \mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|)}{\mathbb{E}(|\log(\hat{\alpha}^{\text{MLE-MU}}/\alpha_0)|)}, \quad (16)$$

where the observations come from a specified contaminated model. Instead of considering each contaminated model separately, that is, each value of $\log \sigma_1$ in the interval $[\log 2, B]$, we aggregate them as we did in Section 4.2. In other words, for a given contamination level $\xi \in \{4\%, 6\%, 10\%, 20\%, 30\%, 40\%\}$, $\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|)$ is the distance measure when a proportion ξ of the observations is generated from a $\mathcal{P}(\sigma_0, \alpha_0)$ and a proportion $1 - \xi$ is generated from a $\mathcal{P}(\sigma_0 \sigma_1^{1/\alpha_0}, \alpha_0)$, with $\log \sigma_1$ randomly selected from the interval $[\log 2, B]$. The distance measures $E(|\log(\hat{\alpha}/\alpha_0)|)$ and $\mathbb{E}(|\log(\hat{\alpha}^{\text{MLE-MU}}/\alpha_0)|)$ in (16) then correspond to the average distances found in Table 4. For example, using the robust P-FLLP estimator instead of MLE decreases the average distance $\mathbb{E}(|\log(\hat{\alpha}/\alpha_0)|)$ from 0.90 to 0.36, under the 20% contamination model. This decrease of 0.54 represents 60% of the MLE distance of 0.90, or 60% of protection obtained by using this robust estimator.

For each level of contamination, the protections are plotted against the premiums in Figure 6. Estimators with low premium and high protection are desired, which corresponds to the upper left corner of the graphs where the P-FLLP estimator is located. This clearly shows the simultaneous efficiency and robustness of the P-FLLP estimator. The dominance of the P-FLLP estimator over PITSE is clear for the five contamination scenarios 6% to 40%, in

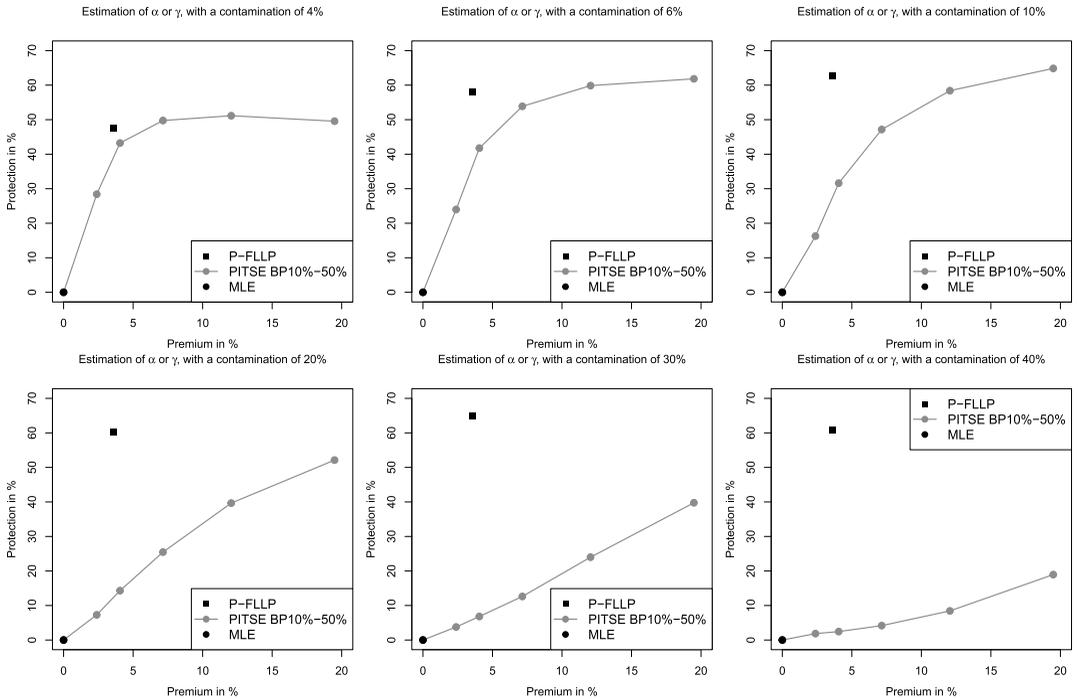


Figure 6 Premium versus protection, for a sample size of 50.

terms of efficiency/robustness. Only PITSE-BP10% has a slightly lower premium than P-FLLP, but it comes with a huge loss of protection. Only PITSE-BP40% and PITSE-BP50% provide slightly more protection than P-FLLP for certain contamination scenarios, but this comes with a huge premium increase. As for the PITSE-BP20% and PISTE-BP30%, they are dominated by the P-FLLP both in premium and protection. For the 4% low contamination scenario, P-FLLP, PITSE-BP20% and PITSE-BP30% show the best performance in terms of premium/protection, making them excellent choices of robust estimators.

5 Fitting the Pareto and exponential models to daily crude oil returns

We collected daily closing prices (denoted $p_t, t = 0, 1, \dots, 998$) of crude oil futures for the 4-year period from October 9, 2017 to October 8, 2021 from the Yahoo Finance website (product code CL = F). We then calculated their $n = 998$ corresponding log returns $r_t := \log(p_t/p_{t-1})$ and discount factors $v_t := \exp(-r_t), t = 1, \dots, 998$. We checked that the r_t 's and the v_t 's are not serially correlated. Here we are interested in modeling losses, which are characterized by negative log returns and discount factors greater than 1.

Specifically, we want to fit the Pareto model $\mathcal{P}(\sigma, \alpha)$ to the right tail of the distribution of discount factors v_t , which is equivalent to fitting the exponential model $\mathcal{E}(\mu = \log \sigma, \alpha)$ to the right tail of the distribution of negative log returns $-r_t = \log v_t$. The first step is to set the parameter σ which defines the right tail. Since the cumulative distribution function (CDF) of the Pareto model defined in (1) can be rewritten as

$$\log v_t = \log \sigma - \alpha^{-1} \log(1 - F_{\mathcal{P}}(v_t|\sigma, \alpha)),$$

we expect a positive linear relationship between $\log v_t$ and $-\log(1 - F_{\mathcal{P}}(v_t|\sigma, \alpha))$, with the intercept and slope given by $\log \sigma$ and α^{-1} . We show in Figure 7 the plot of $\log v_t$ against $-\log(1 - F_n(v_t))$, where F_n is the empirical CDF (adjusted by a factor of $n/(n+1)$). This plot is often used (see, e.g., Dupuis and Victoria-Feser, 2006) to graphically detect the quantile v_t above which the Pareto model is valid, which is found by the value of $\log v_t$ above which the plot yields a straight line. We observe that the points form a straight line for all $\log v_t > 0 \Leftrightarrow v_t > 1$, except for the 4 largest observations which clearly appear as outliers. The Pareto model therefore seems quite adequate for most of the losses and we set the intercept to $\log \sigma = 0$, thereby assuming that $\sigma = 1$ is known.

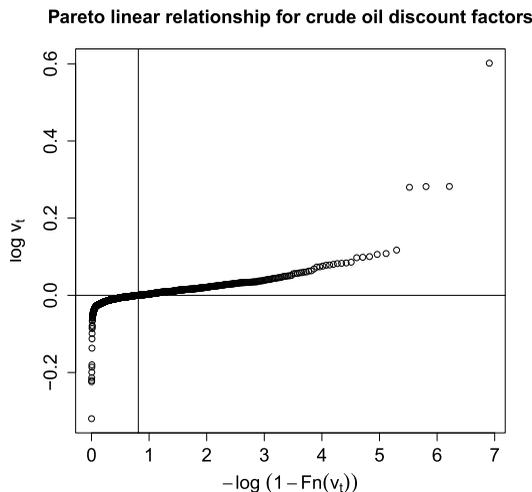


Figure 7 Pareto linear relationship for crude oil discount factors.

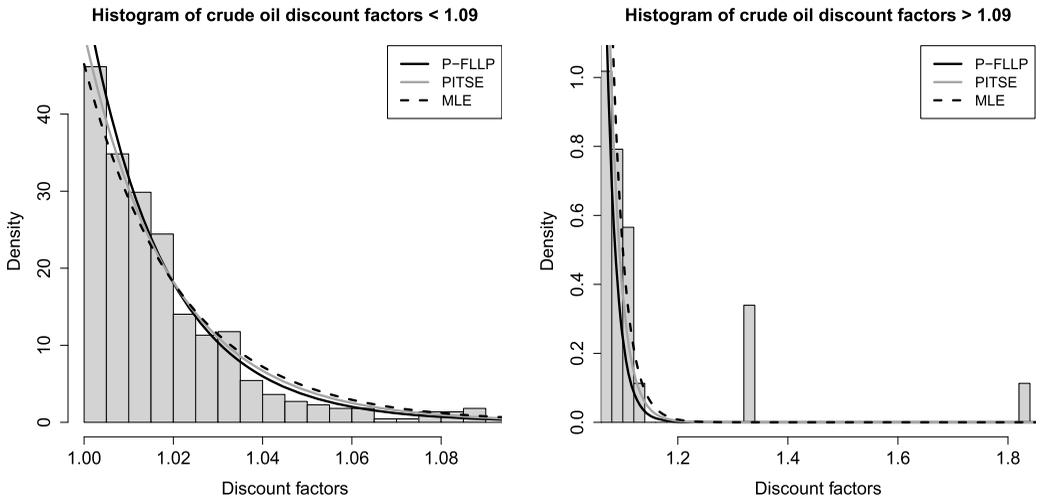


Figure 8 On the left, histogram for crude oil discount factors below 1.09. On the right, histogram for crude oil discount factors above 1.09, that is, the 10 highest values. The histograms are modeled by the Pareto density with the Pareto tail index estimated by the MLE (dashed black line), PITSE (solid grey line) and P-FLLP (solid black line) estimators.

The Pareto model is then fitted to discount factors greater than 1, that is, the 442 largest discount factors denoted by $v_{(557)} \leq \dots \leq v_{(998)}$. The histogram of these discount rates is shown in two parts in Figure 8. On the left, we show the histogram of discount rates below 1.09, which excludes the 10 highest values. On the right, we show the histogram of the 10 highest values, with discount rates greater than 1.09, to better visualize the data set. We now consider the estimation of the tail parameter α using the non-robust MLE and the robust estimators P-FLLP and PITSE. The MLE estimate is $\hat{\alpha}^{\text{MLE}} = 46.501$ with a 95% confidence interval of $\alpha \in [42.266, 50.935]$ and the P-FLLP estimates are $\hat{\omega} = 0.984$ and $\hat{\alpha}^{\text{P-FLLP}} = 56.102$ with a 95% confidence interval of $\alpha \in [50.953, 61.495]$. For the PITSE, it is necessary to fix a tuning constant t by making a compromise between efficiency and robustness. In view of the apparent small proportion of outliers, we choose $t = 1/9$ for an asymptotic relative efficiency (ARE) of 0.99 and a breakdown point (BP) of 10%. Recall that the P-FLLP estimator has the optimal values of $\text{ARE} = 1$ and $\text{BP} = 50\%$. We obtain $\hat{\alpha}^{\text{PITSE}} = 50.570$, halfway between $\hat{\alpha}^{\text{MLE}}$ and $\hat{\alpha}^{\text{P-FLLP}}$. Pareto densities using these MLE, P-FLLP and PITSE estimates of α were added to the histograms in Figure 8. We observe that the Pareto densities with MLE (dashed black line) and P-FLLP (solid black line) estimates have the heaviest and lightest right tail, respectively. Even though the difference in the thickness of the tails might not look that big visually, it still has a major impact on statistical inference. Even though outliers represent a small proportion of the dataset, we see that their impact on the estimation is significant when we compare $\hat{\alpha}^{\text{P-FLLP}} = 56.102$ to $\hat{\alpha}^{\text{MLE}} = 46.501$. As for the PITSE, a partial robustness is obtained with $\hat{\alpha}^{\text{PITSE}} = 50.570$, in the sense that the influence of the outliers is reduced but not completely removed. The impact of outliers is even greater when the estimation is concentrated in the right tail. For example, the probability of observing a discount rate greater than $v_{(994)} = 1.1244$ (the largest observation excluding the 4 clear outliers), given by $S(1.1244 | \sigma = 1, \alpha) := 1.1244^{-\alpha}$, is estimated at 0.0014, 0.0027 and 0.0043, respectively for the P-FLLP, PITSE and MLE, that is, a variation from single to double and even triple.

We now take the analysis of the P-FLLP estimates further. In addition to the non-robust MLE solution with $\hat{\omega} = 1$, the iterative process described in Definition 3.1 generated the unique robust solution $\hat{\omega} = 0.984$ and $\hat{\alpha}^{\text{P-FLLP}} = 56.102$ as given above. The proportion of outliers is therefore estimated at $1 - \hat{\omega} = 0.016$, that is, around 7 observations out of 442.

The estimated probabilities of not being outliers, represented by the $\hat{\pi}_i$'s, are given for the 10 largest observations by 0.703, 0.684, 0.658, 0.572, 0.539, 0.419, 0.0005, 0.0004, 0.0004 and 0. We see that the 4 largest observations $v_{(995)}$, $v_{(996)}$, $v_{(997)}$ and $v_{(998)}$ are clearly identified as outliers and practically completely rejected in the calculation of $\hat{\alpha}^{\text{P-FLLP}}$ with their weights $\hat{\pi}_i$ of 0.0005 or less. As for the other 6 largest observations $v_{(989)}$, \dots , $v_{(994)}$, their estimated probabilities of not being outliers are in a gray area between 0.419 and 0.703, which means that about 3 of them should come from the Pareto component and 3 of them from the outlier component, for our total of about 7 outliers.

Finally, we confirm our results by testing that the observations were drawn from the Pareto distribution. Desgagné, Lafaye de Micheaux and Ouimet (2022) performed an empirical power comparison of 40 goodness-of-fit tests for the univariate Laplace distribution and identified a test named DLO_Z as the most powerful against symmetric alternatives. This test is derived from Rao's score test applied to the symmetric exponential power distribution. Given that the Laplace distribution is a double-exponential, this test can be easily adapted to our situation, where we want to test the composite null hypothesis " $H_0 : X_1, \dots, X_n$ is a random sample from a Pareto model $\mathcal{P}(\sigma, \alpha)$, where σ is known and α is unknown".

The test statistic DLO_Z adapted for our composite null hypothesis is given by

$$K_1(x_1, \dots, x_n, \sigma) := \frac{1}{n} \sum_{i=1}^n z_i \log z_i, \quad \text{with } z_i := y_i / \bar{y} \text{ and } y_i := \log(x_i / \sigma),$$

and K_1 is called the first-power kurtosis. A large (small) value of K_1 indicates a heavier (lighter) right tail than that of a Pareto model, so the test is rejected for sufficiently small or large values of K_1 . The distribution of K_1 and the p-value of the test can be estimated by Monte Carlo simulations.

When the test is applied to the entire sample of $n = 442$ discount rates, the null hypothesis is rejected with a p-value close to 0 because the right tail is too heavy for a Pareto distribution due mainly to the outliers. When the test is applied to the dataset excluding the four largest observations (which were clearly identified as outliers, that is, in conflict with the Pareto model), we obtain a p-value of 0.394, which means that there is no evidence to reject the Pareto model. The p-value reaches a maximum of 0.981 when the 7 largest observations are excluded, which is in agreement with the P-FLLP estimates.

6 Conclusion

In this paper, we introduced an original robust estimator of the Pareto tail index. Although the emphasis is on the Pareto distribution, all results are valid for the estimation of the scale/rate parameter of the two-parameter exponential distribution. This estimator can be thought of as an M -estimator with a ψ -function based on the weighted score function, except that the weights that depend on the proportion of outliers are also estimated. Our approach was to assume that the observations were generated from the FLLP-contaminated Pareto, that is, a mixture of the Pareto and FLLP distributions. The FLLP distribution is an original distribution that we designed specifically to represent any outlier distribution. The parameters are estimated using an iterative process adapted from the expectation-maximization (EM) algorithm to optimize the properties of the estimators in our robustness context. A robust confidence interval for the Pareto tail index is also given.

We showed through different asymptotic results that our estimators reach a breakdown point of 50% with full efficiency. We also showed their simultaneous high efficiency and high robustness for finite samples in a large Monte Carlo simulation study. In particular, the relative efficiency (compared to MLE) increases rapidly from 0.932 to 1 as the sample

size increases from 50 to infinity. This efficiency is largely explained by the fact that the P-FLLP estimators generate MLE-identical results for most uncontaminated samples generated from the Pareto model, an attractive feature for practitioners using MLE. We also compared the performance of MLE, PITSE and P-FLLP estimators for different contamination level scenarios. The P-FLLP estimator had the best overall performance and showed complete robustness for all levels of contamination, where the influence of the outliers is gradually diminished to a point where they are virtually excluded. Finally, we showed how to use our estimators on a real dataset of daily crude oil returns. Unlike most robust estimators of the Pareto tail index, inference is entirely based on observations, which means that there is no tuning constant to be set by the user. We also obtain for each observation an estimated probability that it comes from the Pareto component (versus the outlier component), which makes the detection of outliers very easy and efficient.

Acknowledgments

The financial support of the Canadian Institute of Actuaries is gratefully acknowledged. The author would also like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

The author was supported by the Canadian Institute of Actuaries.

Supplementary Material

Supplement to “Efficient and robust estimation of tail parameters for Pareto and exponential models” (DOI: [10.1214/24-BJPS597SUPP](https://doi.org/10.1214/24-BJPS597SUPP); .zip). R functions and programming code.

References

- Anscombe, F. J. (1960). Rejection of outliers. *Technometrics* **2**, 123–146.
- Bhattacharya, S., Kallitisis, M. and Stoev, S. (2019). Data-adaptive trimming of the Hill estimator and detection of outliers in the extremes of heavy-tailed data. *Electronic Journal of Statistics* **13**, 1872–1925.
- Brazauskas, V. and Serfling, R. (2000a). Robust and efficient estimation of the tail index of a single-parameter Pareto distribution. *North American Actuarial Journal* **4**, 12–27.
- Brazauskas, V. and Serfling, R. (2000b). Robust estimation of tail parameters for two-parameter Pareto and exponential models via generalized quantile statistics. *Extremes* **3**, 231–249.
- Brazauskas, V. and Serfling, R. (2001). Small sample performance of robust estimators of tail parameters for Pareto and exponential models. *Journal of Statistical Computation and Simulation* **70**, 1–19.
- Brzezinski, M. (2016). Robust estimation of the Pareto tail index: A Monte Carlo analysis. *Empirical Economics* **51**, 1–30.
- Davies, L. and Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association* **88**, 782–792.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **39**, 1–22.
- Desgagné, A. (2013). Full robustness in Bayesian modelling of a scale parameter. *Bayesian Analysis* **8**, 187–220.
- Desgagné, A. (2015). Robustness to outliers in location–scale parameter model using log-regularly varying distributions. *The Annals of Statistics* **43**, 1568–1595.
- Desgagné, A. (2021). Efficient and robust estimation of regression and scale parameters, with outlier detection. *Computational Statistics & Data Analysis* **155**, 107114.

- Desgagné, A. (2024). Supplement to “Efficient and robust estimation of tail parameters for Pareto and exponential models.” <https://doi.org/10.1214/24-BJPS597SUPP>
- Desgagné, A., Lafaye de Micheaux, P. and Ouimet, F. (2022). A comprehensive empirical power comparison of univariate goodness-of-fit tests for the Laplace distribution. *Journal of Statistical Computation and Simulation*, 1–46.
- Dierckx, G., Goegebeur, Y. and Guillou, A. (2014). Local robust and asymptotically unbiased estimation of conditional Pareto-type tails. *Test* **23**, 330–355.
- Dupuis, D. J. and Victoria-Feser, M.-P. (2006). A robust prediction error criterion for Pareto modelling of upper tails. *Canadian Journal of Statistics* **34**, 639–658.
- Fedotenkov, I. (2020). A review of more than one hundred Pareto-tail index estimators. *Statistica* **80**, 245–299.
- Finkelstein, M., Tucker, H. G. and Veeh, J. A. (2006). Pareto tail index estimation revisited. *North American Actuarial Journal* **10**, 1–10.
- Gervini, D. and Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *The Annals of Statistics* **30**, 583–616.
- Ghosh, A. (2017). Divergence based robust estimation of the tail index through an exponential regression model. *Statistical Methods & Applications* **26**, 181–213.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*: John Wiley and Sons.
- Kimber, A. C. (1983). Trimming in gamma samples. *Journal of the Royal Statistical Society Series C Applied Statistics* **32**, 7–14.
- Minkah, R., de Wet, T. and Ghosh, A. (2021). Robust estimation of Pareto-type tail index through an exponential regression model. *Communications in Statistics—Theory and Methods*, 1–19.
- Safari, M. A. M., Masseran, N. and Ibrahim, K. (2019). On the identification of extreme outliers and dragon-kings mechanisms in the upper tail of income distribution. *Journal of Applied Statistics* **46**, 1886–1902.
- Vandewalle, B., Beirlant, J., Christmann, A. and Hubert, M. (2007). A robust estimator for the tail index of Pareto-type distributions. *Computational Statistics & Data Analysis* **51**, 6252–6268.
- Victoria-Feser, M.-P. and Ronchetti, E. (1994). Robust methods for personal-income distribution models. *Canadian Journal of Statistics* **22**, 247–258.
- Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics* **15**, 642–656.