

# Detecting influential data points for the Hill estimator in Pareto-type distributions

M. Hubert<sup>(1)</sup>, G. Dierckx<sup>(1)(2)</sup> and D. Vanpaemel<sup>(1)</sup>

- (1) Katholieke Universiteit Leuven, Department of Mathematics and Leuven Statistics Research Center (LStat), Celestijnenlaan 200B, BE-3001 Heverlee, Belgium.
- (2) Hogeschool-Universiteit Brussel, Stormstraat 2, BE-1000 Brussel, Belgium.

*Key words and phrases:* Pareto-type distribution; extreme value index; tail index estimation; influential data points; robustness.

## Abstract

Pareto-type distributions are extreme value distributions for which the extreme value index  $\gamma > 0$ . Classical estimators for  $\gamma > 0$ , like the Hill estimator, tend to overestimate this parameter in the presence of outliers. In this paper we introduce the empirical influence function plot, which displays the influence each data point has on the Hill estimator. To avoid a masking effect, the empirical influence function is based on an asymptotically normal robust estimator for  $\gamma$ . We also derive a cutoff value for the empirical influence function, allowing to flag data points as highly influential if they exceed this cutoff value.

## 1 Introduction

Consider independent and identically distributed random variables  $X_1, \dots, X_n$  with common cumulative distribution function  $F$  and quantile function  $Q = F^{-1}$ . In extreme value statistics, one is interested in the distribution of the maximum  $X_{n,n}$  where  $X_{1,n} \leq \dots \leq X_{n,n}$  denote the corresponding order statistics. Suppose there exist sequences of constants ( $a_n > 0$ )

and  $(b_n \in \mathbb{R})$  such that the properly centered and scaled sample maxima  $\frac{X_{n,n} - b_n}{a_n}$  converge in distribution to a non-degenerate limit  $H$ . It was shown by Fisher and Tippett [1928] and Gnedenko [1943] that, for certain choices of  $a_n$  and  $b_n$ , the limit distribution  $H$  is of the extreme value type:

$$H_\gamma = \begin{cases} \exp(-(1 + \gamma x)^{-1/\gamma}) & \text{for } 1 + \gamma x > 0, \gamma \neq 0 \\ \exp(-\exp(-x)) & \text{for } x \in \mathbb{R}, \gamma = 0. \end{cases} \quad (1)$$

More details can be found in Beirlant *et al.* [2004] amongst others. For most common distributions, this non-degenerate limit for the distribution of the centered and scaled maxima  $X_{n,n}$  exists. Note that  $H_\gamma$  has only one parameter,  $\gamma$ , which is referred to as the extreme value index (EVI).

Distributions for which  $\gamma > 0$  belong to the Fréchet-Pareto class. They are often referred to as Pareto-type distributions or heavy-tailed distributions, as the tail distribution function  $1 - F(x)$  typically decays polynomially. Examples of this class are Pareto, Fréchet and Burr distributions. In case  $\gamma = 0$  the distribution belongs to the Gumbel class, mainly with exponentially decaying tails. Some well known distributions like the Normal, the Gamma, the Logistic and the Log-normal distribution belong to this class. Distributions for which  $\gamma < 0$ , like the Uniform distribution, belong to the Weibull class and have a finite right endpoint.

Our focus will be on Pareto-type distributions. The most popular estimator of  $\gamma$  is then the Hill estimator [Hill, 1975] which will be described in detail in Section 2. This estimator is based on the  $k + 1$  largest data points, and consequently all these observations have some influence on the estimator. Although the Hill estimator is known to have certain flaws (non-robust, biased), it is still the most popular estimator to estimate the tail index. Several ways of determining an optimal  $k$ -value are known. The goal of this paper is to provide the user of the Hill estimator with a diagnostic plot that is helpful to assess the influence of the different data points on the estimator. Points flagged as influential for the Hill estimator

deserve more attention, as they might result in a biased estimate of the tail index.

Our procedure is based on the concept of influence function (IF), which is a standard tool in robust statistics [Hampel *et al.*, 1986]. From the IF we derive the empirical influence function (EIF) (see Jaeckel [1972] and Mallows [1975] amongst others), which still depends on the unknown extreme value index  $\gamma$ . To avoid the masking effect, we then robustly estimate  $\gamma$ . Our robust estimator will be derived from the robust GLM estimator of Cantoni and Ronchetti [2001]. Next we study whether one or more observations have an abnormal large influence on the Hill estimator. This could point to some unusual observations, eventually coming from another distribution. To flag these observations, we compute a cutoff value from the asymptotic distribution of the EIF.

Note that our ideas of combining robust statistics (which typically downweights extreme values) and extreme value statistics (which models the extremes) are also advocated in Dell'Aquila and Embrechts [2006]. They declare robustness to be an important issue in extreme value theory.

Other robust estimators of the tail index have e.g. been proposed in Brazauskas and Serfling [2000], Victoria-Feser and Ronchetti [1994] and Dupuis and Victoria-Feser [2006] for strict Pareto distributions, Dupuis and Field [1998] and Juárez and Schucany [2004] for generalized Pareto distributions and Vandewalle *et al.* [2004] and Vandewalle *et al.* [2007] for Pareto-type distributions. In these methods, the focus is on fitting the bulk of the data and on the detection of outliers. An outlier can be defined as an observation which does not fit the model imposed by the majority of the data points. In the case of heavy-tailed distributions, it is always discussable whether one or a set of 'very large' observations are outliers or not. For such data sets, it can on one hand be argued that the model distribution is heavy-tailed and contains some outliers, but on the other hand also that the model distribution is 'very' heavy-tailed and that these large observations fit that model well. If we solely use the Hill estimator, the conclusion will usually tend to the second interpretation. Using a robust estimator for the tail index will give preference to the first conclusion. With our work,

we propose a diagnostic tool which leaves the interpretation open, but at least correctly flags the observations that have an unusual large influence on the Hill estimator. Whether these influential points concern some unusual data points (outliers) eventually coming from another distribution or not, often will then be decided on by specialists in the field of the data.

Our paper is organized as follows. In Section 2 we recall the Hill estimator and we derive its influence function and empirical influence function. In Section 3 we construct the robust estimator for  $\gamma$ , derive its asymptotic normality, study some robustness properties and illustrate its importance in the EIF. In Section 4 we study the asymptotic distribution of the EIF, from which we derive an appropriate cutoff value to flag highly influential data points. Throughout, and especially in Section 5, we illustrate our method on a real data set. Section 6 concludes and gives directions for further research. All proofs are collected in the Appendix in Section 7.

## 2 The Hill estimator

### 2.1 Definition

As described in the literature, and summarized in Beirlant *et al.* [2004] amongst others, the tail distribution function  $\bar{F}(x) = 1 - F(x)$  of a Pareto-type distribution can be written as  $\bar{F}(x) = x^{-1/\gamma}\ell_F(x)$  for a certain  $\gamma > 0$  and  $\ell_F$  a slowly varying function satisfying  $\lim_{x \rightarrow \infty} \ell(tx)/\ell(x) = 1$  for each  $t > 0$ . This property can also be described in terms of the tail quantile function  $U(x) = F^{-1}(1 - 1/x)$  with  $F^{-1}$  the quantile function of  $F$ , i.e.  $U(x) = x^\gamma \ell_U(x)$  with  $\ell_U$  again a slowly varying function.

For strict Pareto distributions with  $\bar{F}(x) = Cx^{-1/\gamma}$  for a constant  $C$  and  $\gamma > 0$ ,  $\log(X)$  follows an exponential distribution with mean  $\gamma$ . This provides us with a quick and easy

tool to detect if a data set comes from a Pareto distribution. The Pareto QQ-plot

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

puts on the  $x$ -axis the quantiles of the standard exponential distribution and on the  $y$ -axis the log-transformed ordered data. Using the fact that for Pareto-type distributions  $U(x) = x^\gamma \ell_U(x)$ , we can see that  $\log U(x) = \gamma \log x + \log \ell_U(x)$  and since  $\log \ell_U(x)/\log x \rightarrow 0$  as  $x \rightarrow \infty$ , it is easy to see that  $\log U(x) \sim \gamma \log x$  as  $x \rightarrow \infty$ . This implies that for  $\log \left( \frac{n+1}{j} \right)$  and  $\log x_{n-j+1,n}$  large, or for  $j$  small, the Pareto QQ-plot will show a linear behavior with slope  $\gamma$  for Pareto-type distributions. This is illustrated in Figure 1(a) for a data set of size  $n = 500$  coming from a Fréchet(2) distribution. Note that the tail distribution function for a Fréchet( $\alpha$ ) distribution is given by  $\bar{F}(x) = 1 - \exp(-x^{-\alpha})$  for  $x > 0, \alpha > 0$ , and  $\gamma = \frac{1}{\alpha}$ .

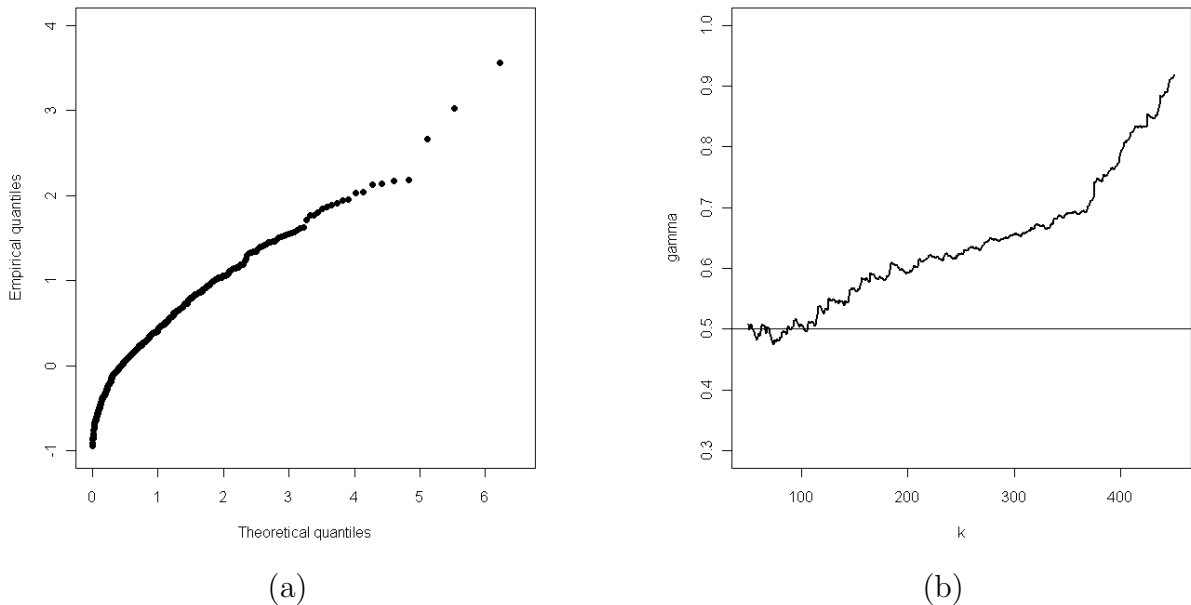


Figure 1: (a) Pareto QQ-plot for a data set coming from a Fréchet(2) distribution with  $n = 500$  and (b) Hill plot of the Hill estimator for the same data, and the true value of  $\gamma = 0.5$  (horizontal line).

Since the slope of the Pareto QQ-plot approximates the unknown parameter  $\gamma$ , an estimate of the slope can be used to estimate  $\gamma$ . This idea was used to construct the Hill

estimator [Hill, 1975]

$$\hat{\gamma}_{k,n}^H = \frac{1}{k} \sum_{j=1}^k \log \frac{X_{n-j+1,n}}{X_{n-k,n}}$$

with  $k$  a well-chosen tuning parameter ( $1 \leq k \leq n-1$ ). This estimator is very popular and easy to calculate although it has the disadvantage that it shows a large bias as  $k$  grows. This can be seen in the so-called Hill plot in Figure 1(b). For the same data set as in Figure 1(a),  $(k, \hat{\gamma}_{k,n}^H)$  is plotted for  $k = 1, \dots, n-1$  whereas the true value of  $\gamma = 0.5$  is indicated by the horizontal line. The number  $k$  thus denotes the number of (largest) observations taken into account to estimate  $\gamma$ . We see that the bias increases enormously as  $k$  grows. The estimates are much better for smaller values of  $k$ .

## 2.2 The influence function of the Hill estimator

As we want to measure the influence each data point has on the Hill estimator, we will make use of its influence function [Hampel *et al.*, 1986]. The influence function is computed at the functional (population) level of the estimator. In general, the influence function of an estimator  $T$  on a point  $y$  at the distribution  $F$  is given by

$$IF(y; T, F) = \lim_{\varepsilon \rightarrow 0} \frac{T(F_{\varepsilon, \Delta_y}) - T(F)}{\varepsilon}$$

(if the limit exists) with  $F_{\varepsilon, \Delta_y} = (1-\varepsilon)F + \varepsilon\Delta_y$  the contaminated distribution where  $\varepsilon$  represents the percentage of contamination and  $\Delta_y$  the Dirac distribution putting all probability mass in  $y$ . The influence function describes the effect of an infinitesimal contamination at the point  $y$  on the functional  $T(F)$ , standardized by the amount of the contamination  $\varepsilon$ .

To derive the IF of the Hill estimator, we first consider its functional form

$$\gamma_{\alpha}^H = T_{\alpha}(F) = \frac{1}{\alpha} \int_{F^{-1}(1-\alpha)}^{+\infty} \log x dF(x) - \log F^{-1}(1-\alpha) \quad (2)$$

where  $0 < \alpha < 1$  denotes the proportion taken into account to estimate  $\gamma$ . From this

functional form (2) the following expression for the influence function of the Hill estimator can be derived (see Appendix Section 7.1):

$$IF(y, \gamma_\alpha^H, F) = \left( -T_\alpha(F) - \frac{1}{\alpha} \log F^{-1}(1 - \alpha) + \frac{1}{\alpha} \log y - \frac{1 - \alpha}{F^{-1}(1 - \alpha)F'(F^{-1}(1 - \alpha))} \right) I(y > F^{-1}(1 - \alpha)). \quad (3)$$

In case  $F$  follows a Pareto-type distribution, the tail distribution function  $\bar{F}(x)$  can be written as  $x^{-1/\gamma}\ell(x)$ , as explained in Section 2.1. This property allows us to simplify expression (3) so that the influence function of the Hill estimator for *Pareto-type distributions* with extreme value index  $\gamma$  can be approximated by

$$IF(y, \gamma_\alpha^H, F) \sim \frac{1}{\alpha} \left( \log \frac{y}{F^{-1}(1 - \alpha)} - \gamma \right) I(y > F^{-1}(1 - \alpha)), \alpha \rightarrow 0. \quad (4)$$

Note that this is in accordance with the influence function for the Hill estimator calculated in Dupuis and Victoria-Feser [2006] for strict Pareto distributions. Figure 2 plots the exact and approximated influence function of the Hill estimator at the Fréchet(2) distribution. We

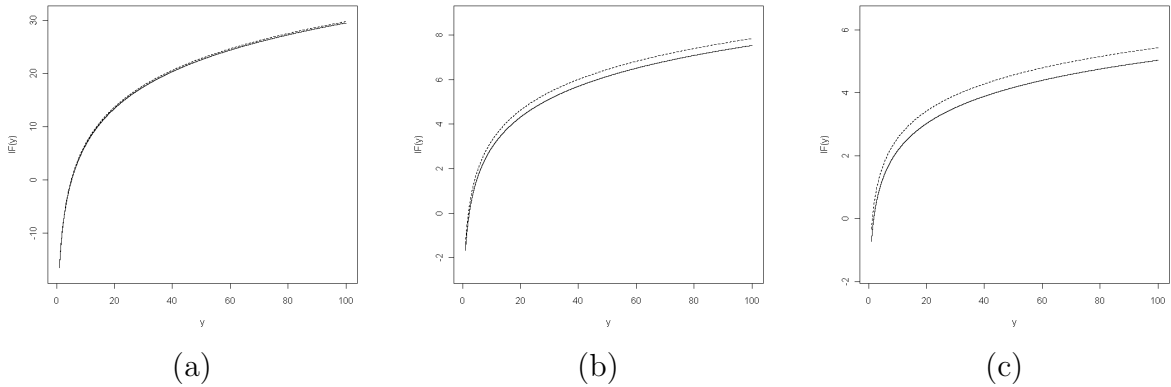


Figure 2: Exact (full line) and approximated (dotted line) influence function for a Fréchet(2) distribution with (a)  $\alpha = 0.1$ , (b)  $\alpha = 0.5$  and (c)  $\alpha = 0.8$ .

see that the approximated influence function lies close to the exact influence function. The approximation is best at the smallest values of  $\alpha$ . Moreover we see that the influence function is monotone increasing. This corresponds with our intuition that the largest observations

have the largest influence on the estimation of the tail behavior of the distribution.

### 2.3 The empirical influence function

Based on expression (4), we define the *empirical influence function* of the Hill estimator  $\hat{\gamma}_{k,n}^H$  on a point  $y$  at a sample of size  $n$  as

$$EIF(y, \hat{\gamma}_{k,n}^H, \hat{F}_n) = \frac{n}{k} \left( \log \frac{y}{x_{n-k,n}} - \gamma \right) I(y > x_{n-k,n}). \quad (5)$$

Of course, this EIF can not be computed as  $\gamma$  is still unknown, and hence, should be replaced with an estimated value  $\hat{\gamma}_{k,n}$ . The most natural choice for  $\hat{\gamma}_{k,n}$  would be the Hill estimator  $\hat{\gamma}_{k,n}^H$  itself. This estimator however will be very biased by highly influential data points, which involves that those highly influential data points will not be detected as such. In robust statistics this is called a masking effect. We illustrate the bias of the Hill estimator in the presence of contamination. For the first example, we have generated 50 data sets containing 500 data points from a Fréchet(2) distribution. We did the same for the second example, but now with data coming from a Burr(1,1,2) distribution. The tail distribution function of a Burr( $\eta, \tau, \lambda$ ) distribution is given by  $\bar{F}(x) = \left( \frac{\eta}{\eta + x^\tau} \right)^\lambda$  for  $x > 0; \eta, \tau, \lambda > 0$  and the extreme value index  $\gamma$  can be calculated as  $\frac{1}{\lambda\tau}$ . For each example and each data set we have contaminated the 2% largest observations by multiplying them by 1000, which is the contamination setting used in Dupuis and Victoria-Feser [2006]. These 2% observations are thus very unlikely to be sampled from a Fréchet(2) distribution (in the first example) or a Burr(1,1,2) distribution (in the second example). This is shown in the Pareto QQ-plot of one contaminated data set coming from a Fréchet(2) distribution in Figure 3(a), and one contaminated data set coming from a Burr(1,1,2) distribution in Figure 3(b). In both plots, the 10 contaminating data points are plotted with a cross sign, and are easy to distinguish since they clearly do not follow the straight line of the regular data points.

In Figure 4 the medians over 50 Hill estimates for  $\gamma$  are plotted for  $k$  between 1 and

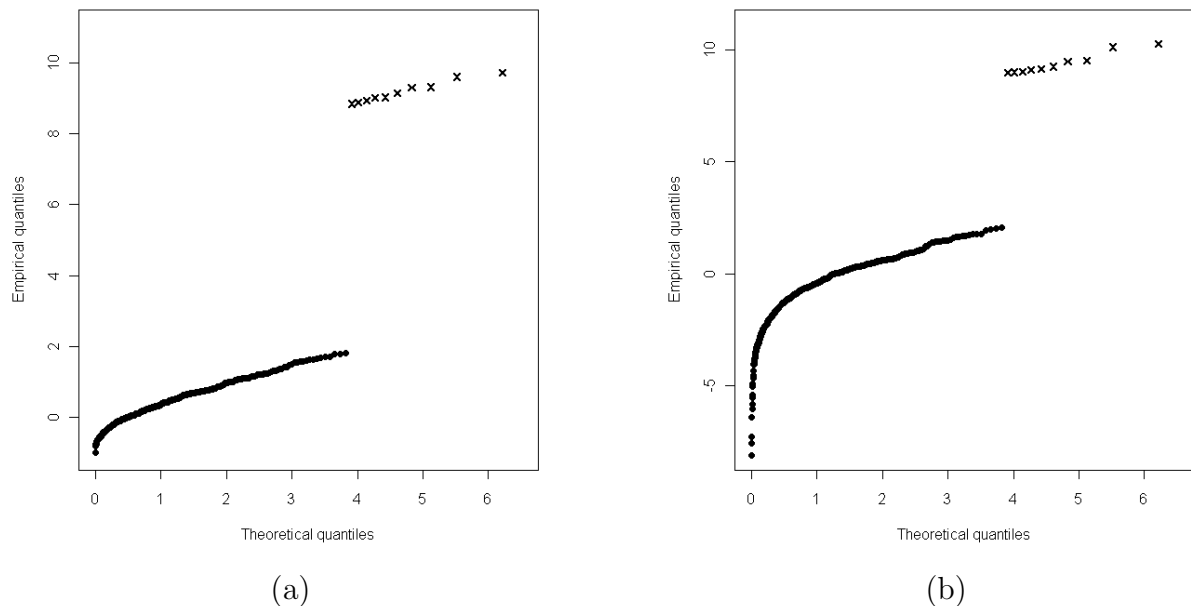


Figure 3: Pareto QQ-plot for a data set coming from (a) a Fréchet(2) distribution and (b) a Burr(1,1,2) distribution with  $n = 500$  and 10 contaminating data points.

499. Figure 4(a) shows the results for data coming from a Fréchet(2) distribution, and Figure 4(b) for data coming from a Burr(1,1,2) distribution. The full line shows the estimates for the uncontaminated data sets, whereas the dashed line exposes the estimates for the contaminated data. We can see that, in case of contamination, for all values of  $k$  the medians of the estimated values for  $\gamma$  are considerably larger than the true value ( $\gamma = 0.5$  in both examples). This will have an impact on the EIF, which will underestimate the real influences of the observations. Hence, it is important to plug in an estimator for  $\gamma$  in the EIF which is not influenced by those unusual data points. Note that the idea of plugging in a robust estimator in the EIF of a non-robust estimator was first proposed by Pison *et al.* [2003] to detect influential observations in factor analysis, and further explored in Pison and Van Aelst [2004] and Debryne *et al.* [2009] in the context of canonical correlation analysis and principal component analysis.

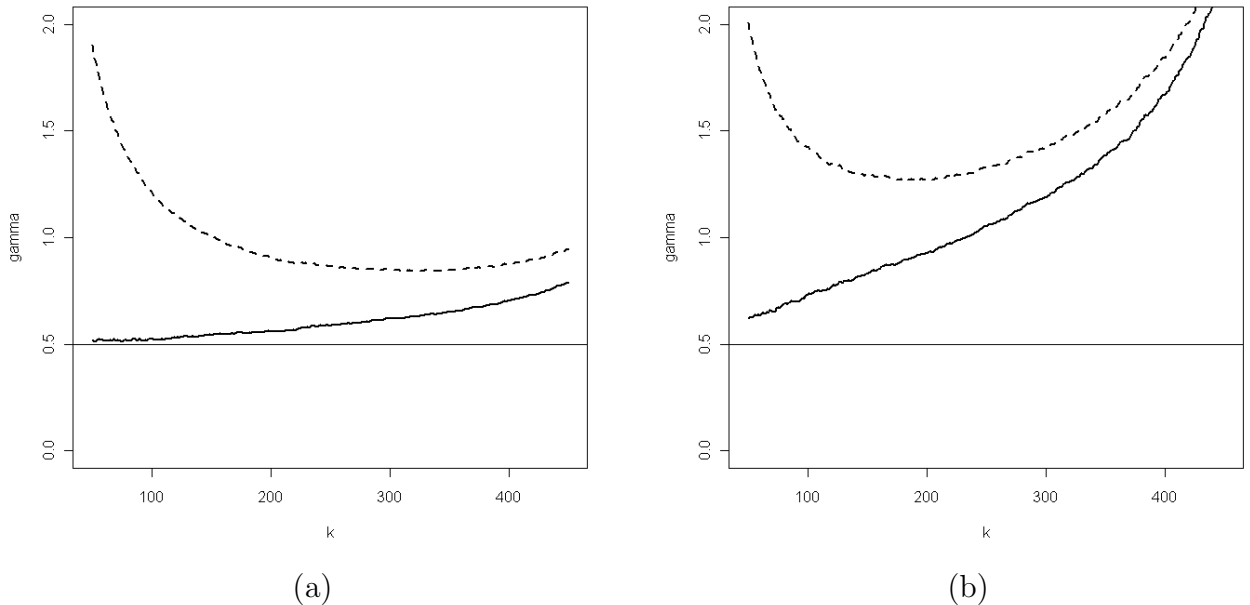


Figure 4: Medians over 50 Hill estimates for  $\gamma$  for data coming from a (a) Fréchet(2) distribution and (b) Burr(1,1,2) distribution with  $n = 500$  (full line) and with 10 contaminating points (dashed line).

### 3 A robust estimator for the extreme value index

To construct this robust estimator for  $\gamma$ , we prefer to use a known estimator of  $\gamma$  with reduced bias compared to the Hill estimator. Whereas the Hill estimator will show severe bias for many values of  $k$ , so that the choice of  $k$  is crucial, the maximum likelihood estimator in Beirlant *et al.* [1999] shows reduced bias for a larger set of  $k$ -values, making the result for  $\gamma$  less influenced by the choice of  $k$ . Note that robust estimators for  $\gamma$  with reduced bias have already been proposed, see e.g. Vandewalle *et al.* [2004, 2007], but since these methods do not provide us with an expression for the distribution of the estimator, and hence a cutoff value for the EIF, a new robust estimator is proposed here.

#### 3.1 The Maximum Likelihood estimator

Our robust estimator for  $\gamma$  will be based on a robust GLM-estimator proposed by Cantoni and Ronchetti [2001]. First, we derive an estimator for  $\gamma$  within the generalized linear model

framework, starting from the Maximum Likelihood estimator proposed in Beirlant *et al.* [1999]. In order to reduce the bias of the Hill estimator, a second order refinement was introduced, based on the scaled log-spacings  $Z_j := j(\log X_{n-j+1,n} - \log X_{n-j,n})$ . To this end a real constant  $\rho < 0$  and a rate function  $b$  satisfying  $b(x) \rightarrow 0$  for  $x \rightarrow \infty$  are introduced, assuming that

$$\frac{\ell_F(tx)}{\ell_F(x)} \sim 1 + b(x) \frac{t^\rho - 1}{\rho}.$$

Under this mild condition on the slowly varying function  $\ell_F$ , it was proven in Beirlant *et al.* [2002] that

$$\left| Z_j - \left( \gamma + b_{n,k} \left( \frac{j}{k+1} \right)^{-\rho} \right) f_j + \beta_j \right| = o_P(b_{n,k}), \quad (6)$$

uniformly in  $j \in \{1, \dots, k\}$ , as  $k, n \rightarrow \infty$  with  $k/n \rightarrow 0$ , where  $(f_1, \dots, f_k)$  is a vector of independent and standard exponentially distributed random variables,  $b_{n,k} := b\left(\frac{n+1}{k+1}\right)$ ,  $2 \leq k \leq n-1$  and  $\beta_j$  a vector of random variables for which it holds that  $\frac{1}{k} \sum_{j=1}^k \beta_j = o_P(b_{n,k})$ .

Therefore, the authors suggest to use the regression model

$$Z_j = \left( \gamma + b_{n,k} \left( \frac{j}{k+1} \right)^{-\rho} \right) f_j. \quad (7)$$

From the expression (7), maximum likelihood estimators for  $\gamma$ ,  $b_{n,k}$  and  $\rho$  can be derived. Note that if we ignore the bias terms  $b_{n,k}$ , the Maximum Likelihood estimator corresponds exactly to the Hill estimator, which can be written as  $\hat{\gamma}_{k,n}^H = \frac{1}{k} \sum_{j=1}^k Z_j$ . Alternatively, model (7) results in a generalized linear model, after plugging in a consistent estimator for  $\rho$ . Again a Maximum Likelihood estimator for  $\gamma$  (as well as a ML estimator for  $b_{n,k}$ ) can be computed, denoted as  $\hat{\gamma}_{k,n}^{ML}$ . However the estimation of  $\rho$  is known to be difficult. The variance of the proposed estimators is quite large. The approach in Fraga Alves *et al.* [2003] is probably one of the more promising ways to estimate the second-order shape parameter for heavy-tailed distributions. Hence, several authors (Matthys and Beirlant [2000], Vandewalle *et al.* [2004], Guillou and Hall [2001], Gomes and Oliveira [2003], Drees and Kaufmann [1998] amongst others) have proposed to set  $\rho = -1$  as a compromise between the bias and the

variance of resulting estimators. Drees and Kaufmann [1998] stated it as follows, “the root mean squared error of our estimator is lower using a fixed  $\rho$ -value rather than one of the consistent estimators, even if  $\rho$  is misspecified”. In Matthys and Beirlant [2000] it is shown that for most applications a  $\rho$ -value between -2 and 0 is most appropriate. Moreover, they show that specifically the estimates  $\gamma_{k,n}^{ML}$  are not highly influenced by a specific choice of  $\rho = -1$ .

In Figure 5 the plot of the ML estimator as a function of  $k$  is shown (full line) for the same data sets that were used in Figure 4. Again the medians over 50 data sets are shown for  $k$  between 1 and 499, in Figure 5(a) for data coming from a Fréchet(2) distribution, and in Figure 5(b) for data coming from a Burr(1,1,2) distribution. Compared to the Hill estimator, we see that the Maximum Likelihood estimator shows much less bias and the estimates of  $\gamma$  tend to be more stable around the true value of  $\gamma$  for a large set of  $k$ -values. Note that Figure 5(b) also illustrates that a misspecification of  $\rho = -1$  is not crucial, as the estimates for  $\gamma$  in case of no contamination are very accurate, even though the real  $\rho$ -value for the Burr(1,1,2) distribution is equal to  $-\frac{1}{\lambda} = -0.5$ . On the other hand we see that also this ML estimator is sensitive to contamination. Similar to the Hill estimator, we see from the dashed lines in Figure 5 that the medians of the ML estimates increase drastically when the data set contains outlying observations.

### 3.2 The robust GLM estimator

Generalized Linear Models [McCullagh and Nelder, 1983] are a tool to model the relationship between predictors  $\mathbf{u}_j$  and a function of the mean of the response variable  $Y_j, j = 1, \dots, n$ . The response variable  $Y_j$  should come from a distribution belonging to the exponential family with conditional mean  $E[Y_j|\mathbf{u}_j] = \mu_j$  and conditional variance  $V[Y_j|\mathbf{u}_j] = v(\mu_j)$ , a function of the mean, for  $j = 1, \dots, n$ . The relationship between the predictors  $\mathbf{u}_j \in \mathbb{R}^p$  and a function of the mean  $g(\mu_j)$ , with  $g(\cdot)$  the link function, is given by  $g(\mu_j) = \mathbf{u}_j^t \boldsymbol{\beta}$  where  $\boldsymbol{\beta} \in \mathbb{R}^p$  is the parameter vector to be estimated. The regression model stated in (7) can be

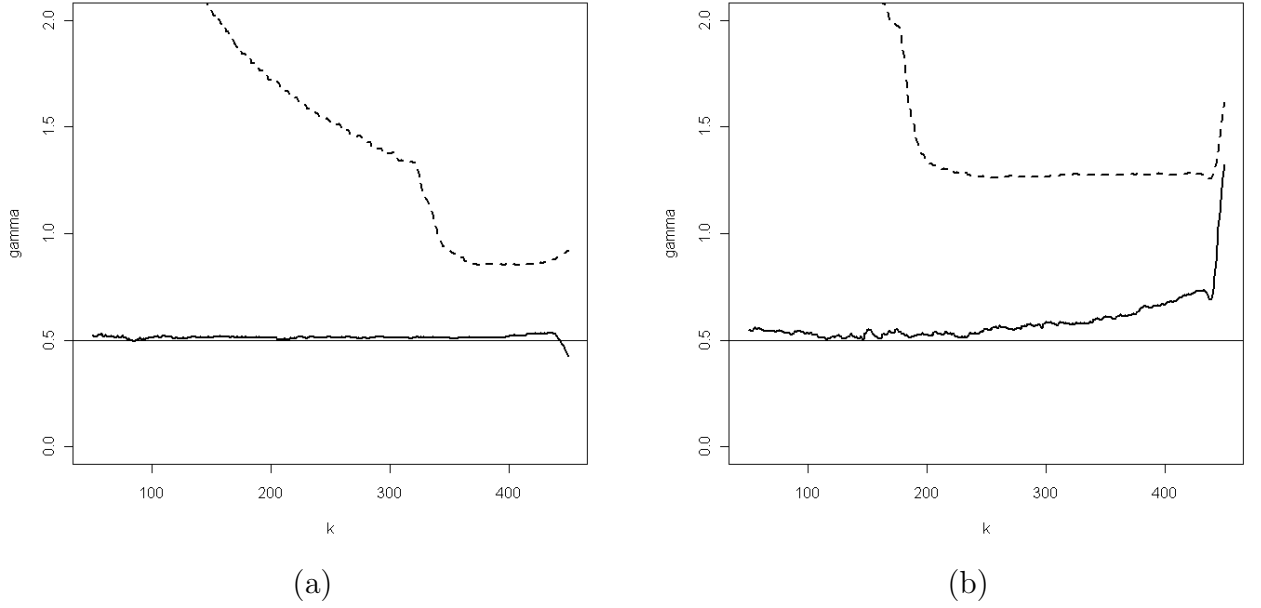


Figure 5: Medians over 50 Maximum Likelihood estimates for  $\gamma$  for data coming from a (a) Fréchet(2) distribution and (b) Burr(1,1,2) distribution with  $n = 500$  (full line) and with 10 contaminating data points (dashed line).

approximated by a Generalized Linear Model. Using a Taylor expansion of  $\exp(x)$  we find indeed that for  $k$  small, such that  $b_{n,k}$  is small,

$$\begin{aligned}
Z_j &\stackrel{\mathcal{D}}{\approx} \exp\left(\log \gamma + \frac{b_{n,k}}{\gamma} \left(\frac{j}{k+1}\right)^{-\rho}\right) f_j \\
&= \exp\left(\beta_0 + \beta_1 \left(\frac{j}{k+1}\right)^{-\rho}\right) f_j \\
&= \exp(\mathbf{u}_j^t \boldsymbol{\beta}) f_j
\end{aligned} \tag{8}$$

with  $\mathbf{u}_j = \left(1 \quad \left(\frac{j}{k+1}\right)^{-\rho}\right)^t$  for  $j = 1, \dots, k$  and

$$\boldsymbol{\beta} = \left(\beta_0 \quad \beta_1\right)^t = \left(\log \gamma \quad \frac{b_{n,k}}{\gamma}\right)^t. \tag{9}$$

This implies that  $Z_j$  is approximately exponentially distributed with mean  $\mu_j = \exp(\mathbf{u}_j^t \boldsymbol{\beta})$  and variance  $v(\mu_j) = \mu_j^2$ , so that the link function  $g$  of this generalized linear model for  $\mu_j$

is the natural logarithm. Note that the approximation in (8) is to be understood as in (6).

The classical quasi-likelihood estimator for  $\boldsymbol{\beta}$  is the solution of the estimating equations

$$\sum_{j=1}^k \frac{z_j - \mu_j}{\mu_j^2} \boldsymbol{\mu}'_j = \mathbf{0}, \quad (10)$$

with  $\mu_j = \exp(\mathbf{u}_j^t \boldsymbol{\beta})$  and  $\boldsymbol{\mu}'_j = \frac{\partial}{\partial \boldsymbol{\beta}} \mu_j$ . As pointed out in Cantoni and Ronchetti [2001], this quasi-likelihood estimator is not robust. A robust M-estimator  $\hat{\boldsymbol{\beta}}$  for Generalized Linear Models is proposed in Cantoni and Ronchetti [2001] and Cantoni and Ronchetti [2006] by solving

$$\sum_{j=1}^k \boldsymbol{\Psi}(z_j, \mu_j) = \mathbf{0} \quad (11)$$

where the score function  $\boldsymbol{\Psi}(z_j, \mu_j) = \boldsymbol{\Psi}(z_j, \exp(\mathbf{u}_j^t \boldsymbol{\beta}))$  is chosen as

$$\boldsymbol{\Psi}(z_j, \mu_j) = \left[ \psi_c(r_j) \frac{\boldsymbol{\mu}'_j}{\mu_j} - \alpha(\boldsymbol{\beta}) \right]. \quad (12)$$

In this expression,  $r_j = \frac{z_j - \mu_j}{\mu_j}$ ,  $j = 1, \dots, k$  are the Pearson residuals,  $\psi_c(r)$  is the Huber function:  $\psi_c(r) = r \min(1, c/|r|)$  and  $\alpha(\boldsymbol{\beta}) = \frac{1}{k} \sum_{j=1}^k E[\psi_c(r_j)] \frac{\boldsymbol{\mu}'_j}{\mu_j}$  a constant to ensure Fisher consistency. A weight on  $\mathbf{u}_j$  can also be included in this estimator, but this is not necessary in our setting. Note that the choice of  $c = \infty$  would result in the classical quasi-likelihood estimator described in (10).

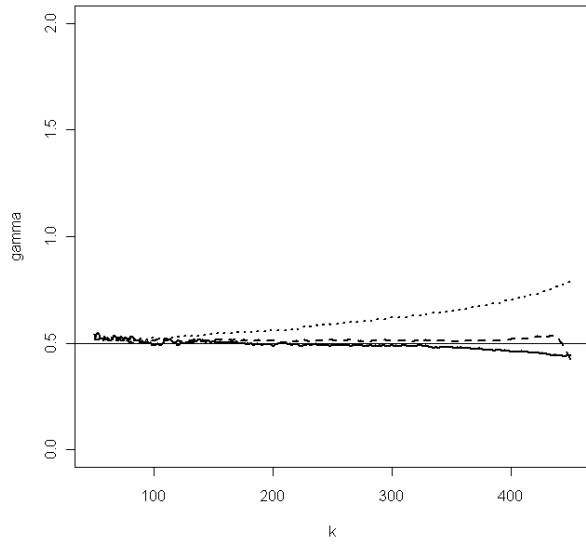
There is no known analytical solution for (11), but an approximation of the solution can be found iteratively e.g. using the Newton-Raphson method or iteratively reweighted least squares [Heritier *et al.*, 2009]. For simplicity reasons we used the first method. Since the ML estimator shows a severe bias for the largest values of  $k$  for some distributions, e.g. the Burr(1,1,2) distribution, and we don't want to base our estimates on too few data points, the M-estimator (11) is only calculated for a certain range of  $k$ -values. In our simulations we omit the 10% largest and 10% smallest values of  $k$ . Hence  $k_{max}$ , the maximum value for which

the estimator is calculated, is equal to  $\lfloor 0.9n \rfloor$ . This means that at least the 10% and at most the 90% largest data point are taken into account to calculate  $\hat{\gamma}_{k,n}^R$ . As a starting value in the Newton-Raphson method for  $k_{max}$ , we take  $\hat{\beta}_{k_{max}}^{(0)} = \left( \log \hat{\gamma}_{k_{max},n}^{ML}, \frac{b_{n,k_{max}}}{\hat{\gamma}_{k_{max},n}^{ML}} \right)^t$ , where  $\hat{\gamma}_{k_{max},n}^{ML}$  is the Maximum Likelihood estimator for  $k = k_{max}$ . For the other values of  $k$ , the result obtained for  $k + 1$  is used as a starting value. Calculations can be done until convergence, but since simulation results show that very good approximations are already obtained after one iteration step, we limit the number of iterations to one.

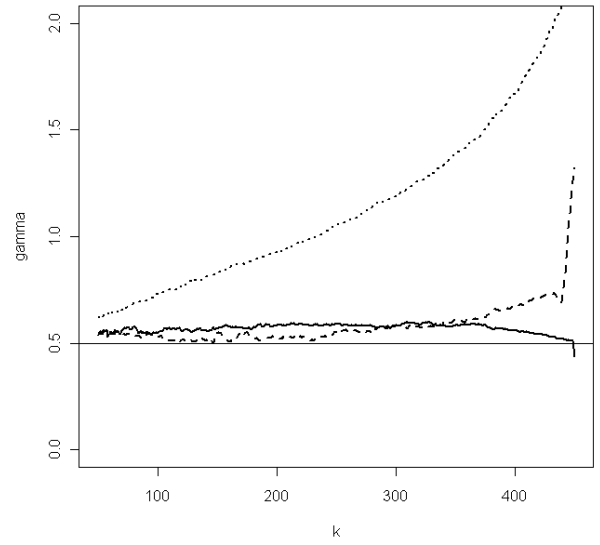
Once  $\hat{\beta}$  has been computed, it follows immediately from (9) that  $\exp(\hat{\beta}_0)$  can be used as an estimate for  $\gamma$ . We denote this estimator as  $\hat{\gamma}_{k,n}^R$ .

For the same data sets as used in Figures 4 and 5, the medians of the robust estimator  $\hat{\gamma}_{k,n}^R$  are calculated for a range of  $k$ -values when there is no contamination present, and in case there are ten contaminating data points. Here, we have set the constant  $c$  in the Huber function equal to 1.105. This yields an asymptotic relative efficiency of about 80%, as will be discussed in Section 3.4, Table 2. The results are compared to the ones found while using the Hill- and Maximum Likelihood estimator and presented in Figure 6. The upper left panel shows medians over 50 data sets of the Hill estimator (dotted line), the Maximum Likelihood estimator (dashed line) and the robust estimator (full line) for a Fréchet(2) distribution in case of no contamination. It can be seen that the median robust estimates stay close to the true value of  $\gamma = 0.5$ . In case of contamination, the lower left panel of Figure 6 shows clearly that the Hill and Maximum Likelihood estimator are much affected by the contamination (yielding estimates close to 1), whereas the robust estimator stays very close to the true value of 0.5.

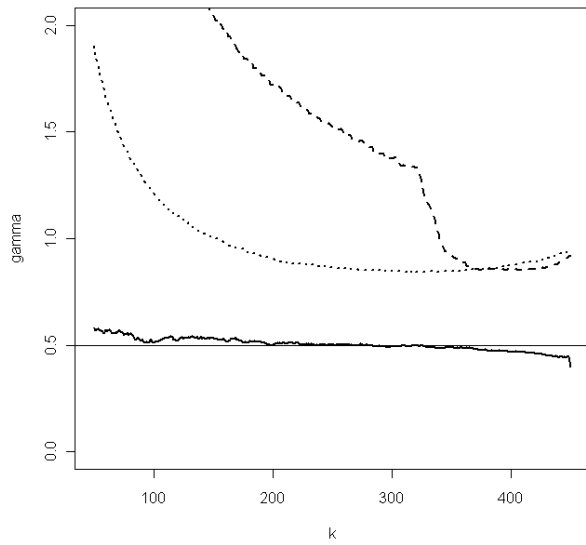
Very similar results are observed for a Burr(1,1,2) distribution in the upper right (no contamination) and lower right (2% contamination) panels of Figure 6. Here the overestimation of the Hill and Maximum Likelihood estimator becomes even more prominent in case of contamination, while the robust estimator only slightly overestimates the true value of  $\gamma = 0.5$ .



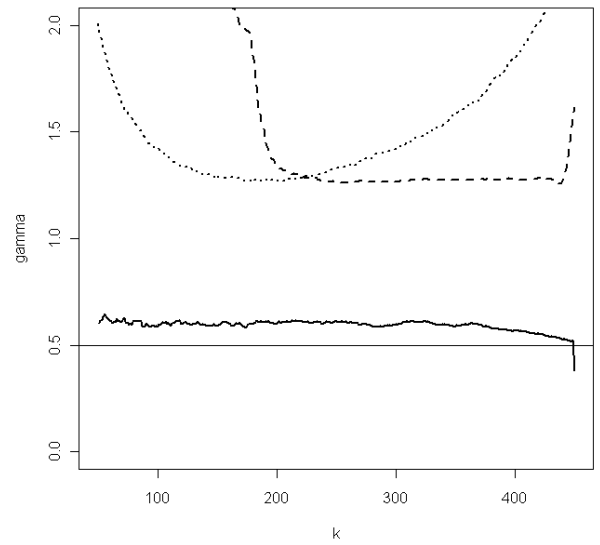
(a)



(b)



(c)



(d)

Figure 6: Medians over 50 Hill estimates (dotted line), Maximum Likelihood estimates (dashed line) and robust estimates (full line) with  $c = 1.105$  for  $\gamma$  for uncontaminated data with  $n = 500$  coming from a (a) Fréchet(2) distribution and (b) Burr(1,1,2) distribution and for data with 10 contaminating points ( $n = 500$ ) coming from a (c) Fréchet(2) distribution and (d) Burr(1,1,2) distribution.

### 3.3 Asymptotic normality

We derive the asymptotic normality of  $\hat{\gamma}_{k,n}^R$  under the assumption that the  $Z_j$  in (8) are exactly exponentially distributed. To simplify calculations, we first introduce some notations. The density function of  $Z_j|\mathbf{u}_j$  will be denoted by  $f_{Z_j}(z_j|\mathbf{u}_j, \mu_j) = 1/\mu_j \exp(-z_j/\mu_j)$  and the residuals of the model  $r_j = \frac{z_j - \mu_j}{\mu_j}$ . We also introduce the notations  $d_1 := \frac{1}{k} \sum_{j=1}^k \left(\frac{j}{k+1}\right)^{-\rho}$  and  $d_2 := \frac{1}{k} \sum_{j=1}^k \left(\frac{j}{k+1}\right)^{-2\rho}$ . Note that  $d_1 \rightarrow \frac{1}{1-\rho}$  and  $d_2 \rightarrow \frac{1}{1-2\rho}$  for  $k \rightarrow \infty$ .

From Cantoni and Ronchetti [2001] it immediately follows that the robust estimator  $\hat{\beta}$  that satisfies (11), is asymptotically normal with asymptotic variance  $M^{-1}QM^{-1}$ . With  $X$  being a  $k$  by 2 matrix where  $X(j, 1) = 1$  and  $X(j, 2) = \left(\frac{j}{k+1}\right)^{-\rho}$  for  $j = 1, \dots, k$ , we can write  $M = \frac{1}{k} X^t B X$  where  $B$  is a diagonal matrix with elements  $b_j = \mu_j E \left[ \psi_c(r_j) \frac{\partial}{\partial \mu_j} \log f_{Z_j}(z_j|\mathbf{u}_j, \mu_j) \right]$  and  $Q = \frac{1}{k} X^t A X - \alpha(\beta)\alpha(\beta)^t$  where  $A$  is a diagonal matrix with elements  $a_j = E[\psi_c^2(r_j)]$ .

In the Appendix (Section 7.2) it is shown that  $E[\psi_c(r_j)] = -e^{-(1+c)} := c_2$  for  $c \geq 1$ . This is in accordance with the values for  $c$  that will be proposed in Table 2, so we choose  $c \geq 1$  from now on. It is also shown that  $E[\psi_c^2(r_j)] = 1 - 2(1+c)e^{-(1+c)} := a$  and  $\mu_j E \left[ \psi_c(r_j) \frac{\partial}{\partial \mu_j} \log f_{Z_j}(z_j|\mathbf{u}_j, \mu_j) \right] = 1 - (2+c)e^{-(1+c)} := b$ . Putting all this together (the computation is provided in Appendix, Section 7.2) leads to the formulation of the asymptotic normality for  $\hat{\beta}$ :

$$\sqrt{k}(\hat{\beta} - \beta) \xrightarrow{d} N \left( 0, \frac{1}{b^2(d_2 - d_1^2)} \begin{pmatrix} (a - c_2^2)d_2 + d_1^2 c_2^2 & -ad_1 \\ -ad_1 & a \end{pmatrix} \right)$$

from which it follows that  $\sqrt{k}(\hat{\beta}_0 - \beta_0) \xrightarrow{d} N(0, \sigma_{\rho,c}^2)$  with  $\sigma_{\rho,c}^2 = \frac{(a - c_2^2)d_2 + d_1^2 c_2^2}{b^2(d_2 - d_1^2)}$ . Finally, it holds that  $\hat{\gamma}_{k,n}^R = \exp(\hat{\beta}_0)$  is asymptotical normally distributed [Serfling]. Hence  $\sqrt{k}(\hat{\gamma}_{k,n}^R - \gamma) \xrightarrow{d} N(0, \gamma^2 \sigma_{\rho,c}^2)$  such that the variance asymptotically behaves as  $\frac{\gamma^2 \sigma_{\rho,c}^2}{k}$ .

In the case of a Pareto type distribution, where the  $Z_j$  are only approximately exponentially distributed, this result will hold approximately. In order to compare the asymptotical and empirical variance of  $\hat{\gamma}_{k,n}^R$ , 100 data sets coming from an uncontaminated Burr(1,1,2)

distribution of different sizes  $n$  were simulated. The empirical and asymptotical variances for  $\hat{\gamma}_{k,n}^R$  for different values of  $k$  were calculated with the Huber constant  $c = 1.105$  as in Table 2, to ensure an ARE of about 80%. Although the  $Z_j$  are only approximately exponentially distributed, Table 1 shows that the empirical and asymptotical variance for each  $k$  coincide quite well. Moreover the difference between the two variances become smaller as  $n$  and/or  $k$  becomes larger.

$n$	$k$	asymptotical variance	empirical variance
500	50	0.0258	0.0284
	100	0.0127	0.0156
	150	0.00842	0.00915
1000	75	0.0170	0.0206
	150	0.00842	0.00978
	225	0.00560	0.00717
2000	100	0.0127	0.0146
	200	0.00630	0.00742
	300	0.00419	0.00500

Table 1: Asymptotical and empirical variance of  $\hat{\gamma}_{k,n}^R$  at a Burr(1,1,2) distribution.

### 3.4 Robustness properties

For a robust estimator, it is highly desirable to have a bounded influence function. The influence function of  $\hat{\beta}$  is proportional to its score function  $\Psi$ , i.e. it is given by [Hampel *et al.*, 1986]

$$IF(z; \hat{\beta}, G_j) = -(E[\frac{\partial}{\partial \beta} \Psi(z, \mu_j)])^{-1} \Psi(z, \mu_j)$$

where  $G_j$  behaves asymptotically as an exponential distribution with mean  $\mu_j = \exp(\mathbf{u}_j^t \beta)$ . From the choice of  $\Psi$  in (12), the bounded Huber function  $\psi_c$  and the fact that  $\mu_j'/\mu_j = \left(1 - \left(\frac{j}{k+1}\right)^{-\rho}\right)^t$  is bounded with respect to  $z$  (see Appendix, Section 7.2), we obtain a bounded influence function for  $\hat{\beta}$ , and hence for  $\hat{\beta}_0 = \log(\hat{\gamma}^R)$ .

The level of robustness can be controlled by the choice of  $c \geq 0$ , the constant in the

Huber function referred to as the Huber constant. A high  $c$  implies a high efficiency but low robustness, since the Pearson residuals are not very likely to be truncated by  $c$ . A smaller value of  $c$  makes the estimator more robust, but at the cost of a lower efficiency. Often  $c$  is chosen to ensure a certain level of efficiency with respect to a classical counterpart of the estimator, which is expressed in terms of the Asymptotic Relative Efficiency (ARE), see e.g. Victoria-Feser and Ronchetti [1994]. We already know that choosing  $c = \infty$  results in the classical quasi-likelihood estimator. At the end of Section 3.3 we derived how the asymptotic variance of  $\hat{\gamma}_{k,n}^R$  depends on  $c$  and  $\rho$ . Table 2 gives an overview of which values of  $c$  should be chosen to reach a certain ARE for the most common choices for  $\rho$ . From this Table we can see that  $c$  indeed increases with larger efficiency. Although we could opt for a large efficiency of, say 95%, this would increase the maxbias at contaminated data (see also Maronna *et al.* [2006], page 144). Hence, in our examples we have always set  $c = 1.105$  and  $\rho = -1$  to obtain an efficiency of 80% in case  $\rho$  is specified correctly. In case of a misspecification of  $\rho$ , an ARE around 80% can still be expected since for other  $\rho$ -values the required  $c$ -value to obtain an ARE of 80% does not deviate much from 1.105. Note that in case  $\rho < -1$ , the ARE will be larger than 80% with this choice for  $c$ .

	$\rho = -2$	$\rho = -1$	$\rho = -0.5$
ARE	$c$	$c$	$c$
80%	1.080	1.105	1.125
85%	1.380	1.400	1.410
90%	1.815	1.825	1.830
95%	2.555	2.560	2.565

Table 2: Values for  $c$  given a certain  $\rho$  and a certain ARE.

### 3.5 Choosing the value of $k$

A natural question that always arises in extreme value analysis is how to choose an appropriate value for  $k$ , the number of data points that are taken into account to estimate  $\gamma$ . One option is to look at a so called “stable region”, an interval of  $k$ -values in which the estimates

do not fluctuate a lot, and to pick a  $k$  in that interval. The drawback of this method is that it needs the intervention and skilled judgement of the user.

For the Hill estimator, a number of automatic methods to estimate  $k$  are available, an overview can be found in Beirlant *et al.* [2004], pages 123–129. A common approach consists of minimizing the asymptotic MSE of the Hill estimator, which uses estimates of  $b_{n,k}$  and  $\rho$  calculated by the ML estimator. Note that some of these methods rely on a value of the parameter  $\rho$  but the authors mention that a wrong specification of  $\rho$  does not seem to be a major problem. In practical applications, ad hoc methods are often applied, such as DuMouchel’s rule which consists in choosing  $k$  so that the 10% largest data points are used to calculate the Hill estimator [DuMouchel, 1983].

In order to set up an automatic method for our robust estimator, from a practical point of view we propose to use the median of the calculated  $\hat{\gamma}_{k,n}^R$  values over all  $k$ , which we denote as  $\hat{\gamma}_n^R$ . The median value is not always observed, therefore we define  $k_R$  as the minimum of all  $k$ -values for which the difference with  $\hat{\gamma}_n^R$  is the smallest. The robust estimator is then given by the corresponding  $\hat{\gamma}_{k_R,n}^R$ .

For comparison reasons, we apply our ad hoc rule (taking the median of the estimates for  $\gamma$  over all  $k$  and find the minimal  $k$ -value, denoted as  $k_{ML}$ , for which the difference of  $\hat{\gamma}_{k,n}^{ML}$  with the median is the smallest) also to the Maximum Likelihood estimator, to which we will refer to as  $\hat{\gamma}_{k_{ML},n}^{ML}$ .

Alternatively, one could estimate  $k$  and  $\gamma$  robustly as in Dupuis and Victoria-Feser [2006]. The value of  $k$  is estimated minimizing the robust prediction error criterion. This criterion depends on the choice of a robust estimator for  $\gamma$ . Dupuis and Victoria-Feser [2006] used the Weighted Maximum Likelihood Estimator (WMLE) [Dupuis and Morgenthaler, 2002]. The WMLE is an M-estimator as the solution of  $\sum_{j=1}^k \psi(x_{n-j+1}, \theta) = 0$ . Here  $\psi(x, \theta) = w(x; \theta) \frac{\partial}{\partial \theta} \log f(x; \theta)$  where Dupuis and Victoria-Feser [2006] used the density of a Pareto distribution with extreme value index  $\gamma = \frac{1}{\theta}$  for  $f(x)$ . Further  $w(x; \theta)$  is a weight function with values in  $[0, 1]$ . Since the result can be biased, a bias correction term should

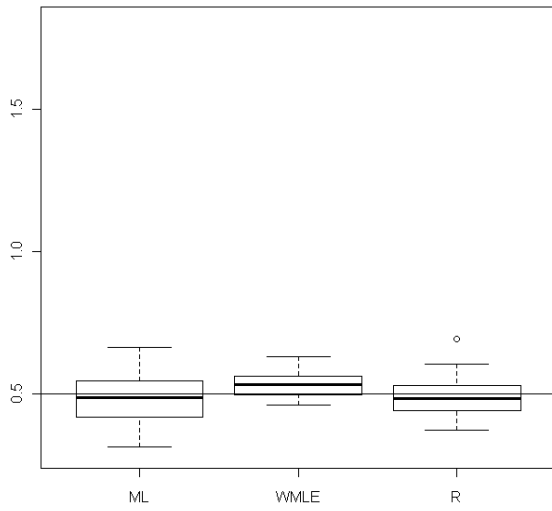
be added. For the weight function  $w(x; \theta)$ , Dupuis and Victoria-Feser [2006] propose to use  $w(x_{n-j+1}; \theta) = \psi_{c_w}(t_{n-j+1})/t_{n-j+1}$  where  $c_w$  is a tuning constant determining efficiency and robustness, and  $t_{n-j+1}$  are standardized residuals based on the data. We will refer to this estimator as  $\hat{\gamma}_n^{WMLE}$ .

To examine the performance of the ad hoc robust and non-robust estimators and compare them to the WMLE estimator, we calculated  $\hat{\gamma}_{k_{ML},n}^{ML}$ ,  $\hat{\gamma}_n^{WMLE}$  and  $\hat{\gamma}_{k_R,n}^R$  for the data sets and outlier configurations used in Figure 6. Again  $c$  was chosen equal to 1.105, and  $c_w = 1.25$  as suggested by Dupuis and Victoria-Feser [2006].

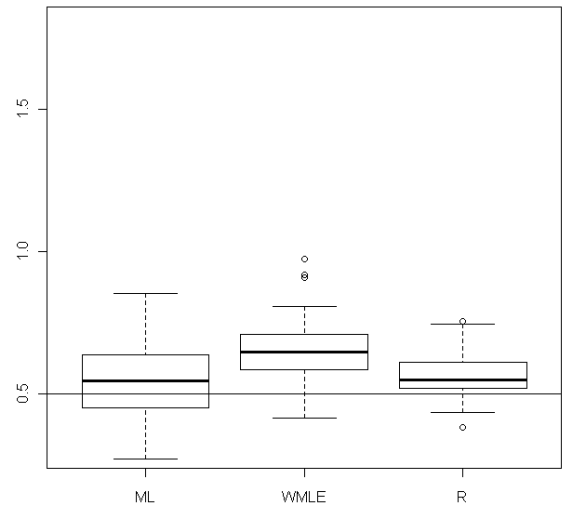
Figure 7(a) shows boxplots for  $\hat{\gamma}_{k_{ML},n}^{ML}$  (left),  $\hat{\gamma}_n^{WMLE}$  (middle) and  $\hat{\gamma}_{k_R,n}^R$  (right) for 50 data sets of size  $n = 500$  coming from a Fréchet(2) distribution without contamination. We see that the three estimators perform well, estimating the true value  $\gamma = 0.5$  quite accurately. The WMLE estimator has the smallest variance but overestimates the extreme value index in most of the cases. The bias of the ML and our new robust estimator  $\hat{\gamma}_{k_R,n}^R$  are comparable, the latter having a smaller variance. When the data come from a Burr(1,1,2) distribution without contamination, the three estimators often overestimate the true value  $\gamma = 0.5$  (as was already observed in Figure 6(b)). Here,  $\hat{\gamma}_{k_R,n}^R$  clearly has a smaller bias and variance than  $\hat{\gamma}_n^{WMLE}$ .

In Figures 7(c) and (d) 2% contamination was added as before. The ML estimator (left) shows a severe positive bias caused by the outliers, as expected, both for the Fréchet(2) (Figure 7(c)) and the Burr(1,1,2) (Figure 7(d)) distribution. On the other hand, both robust estimators,  $\hat{\gamma}_n^{WMLE}$  (middle) and  $\hat{\gamma}_{k_R,n}^R$  (right), give quite accurate values for  $\gamma$  in case of a contaminated Fréchet(2) distribution. Again it can be noticed in Figure 7(c) that the WMLE estimator has a smaller variance but larger bias than  $\hat{\gamma}_{k_R,n}^R$ . In case of a Burr(1,1,2) distribution with 2% contamination, we can see from Figure 7(d) that both robust estimators show a positive bias, which is however much smaller for  $\hat{\gamma}_{k_R,n}^R$  than for  $\hat{\gamma}_n^{WMLE}$ . Also its variance is much smaller.

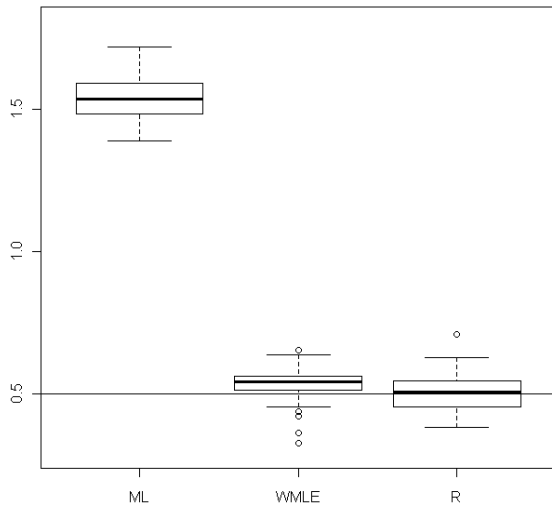
We conclude that  $\hat{\gamma}_{k_R,n}^R$ , where  $k_R$  is chosen by an ad hoc method, is not outperformed



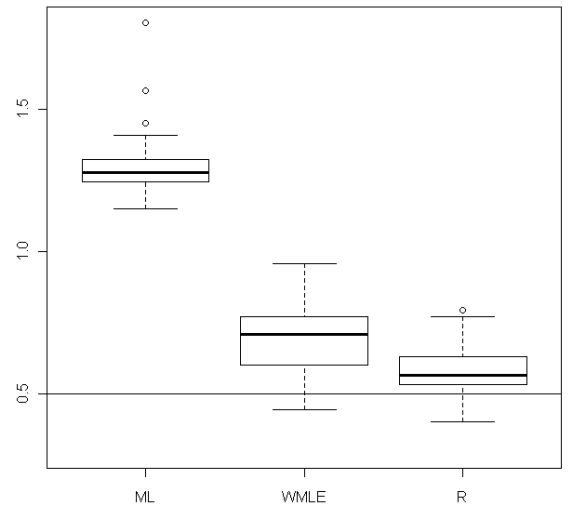
(a)



(b)



(c)



(d)

Figure 7: Boxplots of 50 Maximum Likelihood estimates (left), WMLE estimates (middle,  $c_w = 1.25$ ) and robust estimates (right,  $c = 1.105$ ) for  $\gamma$  for uncontaminated data with  $n = 500$  coming from a (a) Fréchet(2) distribution and (b) Burr(1,1,2) distribution and boxplots of 50 Maximum Likelihood estimates (left), WMLE estimates (middle,  $c_w = 1.25$ ) and robust estimates (right,  $c = 1.105$ ) for  $\gamma$  for data with 10 contaminating points ( $n = 500$ ) coming from a (c) Fréchet(2) distribution and (d) Burr(1,1,2) distribution.

by the WMLE-method, which also estimates  $k$ . Since the estimation of  $k$  (and hence of  $\gamma$ ) in the latter case seems to be affected by the choice of the upper and lower bound in the optimization algorithm, and the estimation of  $k$  uses a considerable amount of computation time, we choose to continue with the robust estimator  $\hat{\gamma}_{k_R, n}^R$ .

## 4 Flagging influential data points

Now that we found a robust estimator for  $\gamma$ , we can plug in this estimator in the empirical influence function (5) of the Hill estimator. Then we can flag influential data points, based on appropriate cutoff values for the EIF.

### 4.1 An empirical influence function plot

In practice, we will not look at all EIF's for all values of  $k$ , but concentrate on the EIF at the optimal  $k_R$  value, i.e.

$$EIF^R(y, \hat{\gamma}_{k_R, n}^H, \hat{F}_n) = \frac{n}{k_R} \left( \log \frac{y}{x_{n-k_R, n}} - \hat{\gamma}_{k_R, n}^R \right) I(y > x_{n-k_R, n}). \quad (13)$$

To measure the influence of a single observation, we can finally consider  $EIF^R(y, \hat{\gamma}_{k_R, n}^H, \hat{F}_n)$  for  $y = x_{n-k_R+1, n}, x_{n-k_R+2, n}, \dots, x_{n, n}$ . To display this information in a clear way, we define the **empirical influence function plot** (EIFP), which displays the  $EIF^R$  in the data points versus their rank number (where the largest data point has rank number 1, and the smallest rank number  $k_R$ ):

$$\left( j, EIF^R(x_{n-j+1, n}, \hat{\gamma}_{k_R, n}^H, \hat{F}_n) \right) \quad j = 1, \dots, k_R.$$

We illustrate this EIFP on the Hedge fund data consisting of 72 log-returns (in 100%) on alternative investments on a monthly basis between January 1997 and December 2002, which was described in detail in Perret-Gentil and Victoria-Feser [2003] and Dupuis and Victoria-

Feser [2006]. In this example, 100 minus the log-return is considered for the evaluation of the downside risk. It can be assumed that these data come from a Pareto-type distribution as can be seen from the Pareto QQ-plot in Figure 8, showing a more or less straight line from a certain point on. It can also be noticed that the three largest observations (plotted with a cross sign) do not seem to follow the model, since they deviate from the straight line formed by the other observations. We expect them to be highly influential for the Hill estimator.

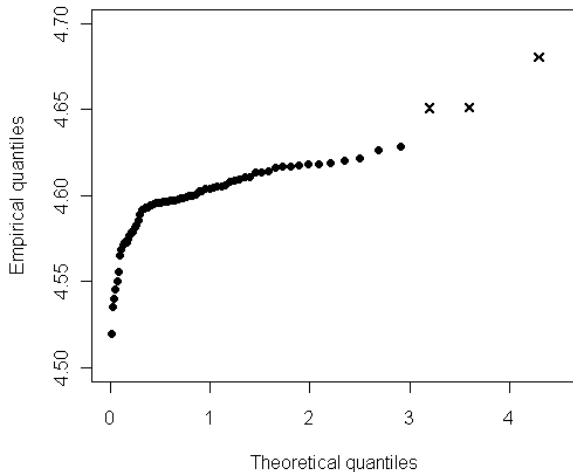


Figure 8: Pareto QQ-plot for the Hedge fund data set.

We first make a plot of  $\hat{\gamma}_{k,n}^R$  as a function of  $k$ , shown in Figure 9(a). To obtain an ARE around 80%, we have set  $c = 1.105$ . The full line in Figure 9(a) shows the robust estimator  $\hat{\gamma}_{k,n}^R$ . For illustrative purposes we have also added the Hill estimator (dotted line) and the maximum likelihood estimator (dashed line). The value for  $\hat{\gamma}_{k_R,n}^R$  is equal to 0.01365 ( $k_R = 25$ ), which is very comparable to  $\hat{\gamma}_n^{WMLE} = 0.01343$  (obtained at  $k = 22$ ) for  $c_w = 1.25$  as was shown in Dupuis and Victoria-Feser [2006]. The Maximum Likelihood estimator yields  $\hat{\gamma}_{k_{ML},n}^{ML} = 0.01531$  for  $k_{ML} = 36$  and the Hill estimator for a  $k_{opt} = 45$  that minimizes the Asymptotic Mean Squared Error (AMSE), has a value  $\hat{\gamma}_{k_{opt},n}^H = 0.01534$ .

The corresponding EIFP for the  $k_R = 25$  largest observations is shown in Figure 9(b). We see that the three largest cases have a considerably larger  $EIF^R$  value than the other observations. To decide whether they are more influential than we expect when all obser-

variations would come from a homogenous fat-tailed distribution, we need to derive a cutoff value on this EIFP. This will be outlined in the next section.

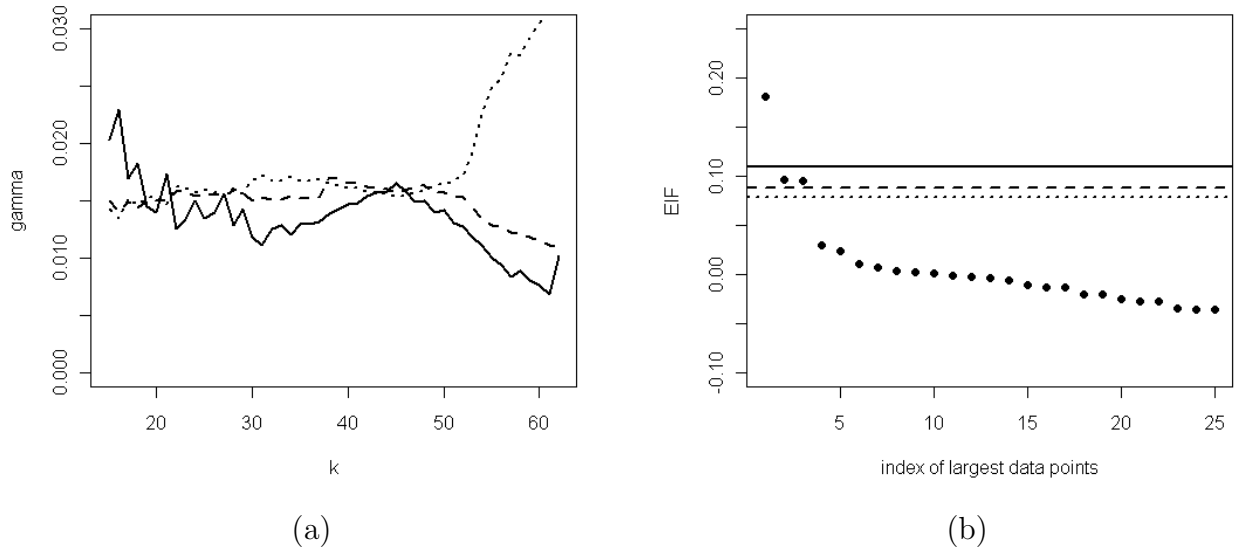


Figure 9: (a) Plot of  $\gamma$  as a function of  $k$  for the Hedge fund data for the Hill estimator (dotted line), Maximum Likelihood estimator (dashed line) and robust estimator (full line); (b) EIFP plot for the same data with 99% (dotted line), 99.5% (dashed line) and 99.9% (full line) cutoff value.

## 4.2 The asymptotic distribution of the robust empirical influence function

We can obtain a cutoff value on the EIFP by deriving the asymptotic distribution of the  $EIF^R$ . To this end, we need to know the distribution of the two components:  $\log Y/x_{n-k_R,n}$  and  $\hat{\gamma}_{k_R,n}^R$ .

In case  $y$  comes from the same Pareto-type distribution as the data,  $Y/x_{n-k,n}$  follows an asymptotic strict Pareto distribution with parameter  $\gamma$  (if the threshold  $x_{n-k,n}$  is chosen high enough), so that the distribution of  $\log Y/x_{n-k,n}$  is asymptotically exponential with mean  $\gamma$ . The asymptotic distribution of  $\hat{\gamma}_{k,n}^R$  was derived in Section 3.3.

Now that we know the asymptotic distribution of the two components  $\log Y/x_{n-k_R,n}$  and

$\hat{\gamma}_{k_R,n}^R$ , the distribution of  $EIF^R(y, \hat{\gamma}_{k_R,n}^H, \hat{F}_n) = \frac{n}{k_R} \left( \log \frac{y}{x_{n-k_R,n}} - \hat{\gamma}_{k_R,n}^R \right)$  can be derived. Note that the empirical influence function asymptotically equals  $EIF \sim \frac{n}{k_R} \left( \log \frac{Y}{U(n/k_R)} - \hat{\gamma}_{k_R,n}^R \right)$  with  $U$  as before the tail quantile function. Indeed from the asymptotical results for the empirical quantile function  $\sqrt{n} \left( \hat{F}^{-1}(p) - F^{-1}(p) \right) \sim N(0, p(1-p)/f^2(F^{-1}(p)))$  it follows that  $\frac{n}{k_R} \log \frac{U(n/k_R)}{X_{n-k_R}} \rightarrow_P 0$ . Because we know the asymptotic distributions of the two components  $\log Y/U(n/k_R)$  and  $\hat{\gamma}_{k_R,n}^R$ , the asymptotical distribution of the  $EIF^R$  can be derived.

Let us first consider the situation where  $y$  is not a data point, but it follows the same Pareto-type distribution as the data. In that case  $\log(Y/U(n/k_R))$  is asymptotically exponentially distributed, and  $\hat{\gamma}_{k_R,n}^R$  is approximately asymptotically normally distributed. Moreover they are independent from each other, which implies that the asymptotic distribution of  $\log Y/U(n/k_R) - \hat{\gamma}_{k_R,n}^R$  is approximately Exponential Gaussian. This distribution has probability density function

$$f(y; \mu, \sigma, \lambda) = \frac{1}{\lambda} \exp \left[ \frac{\mu}{\lambda} + \frac{\sigma^2}{2\lambda^2} - \frac{y}{\lambda} \right] \Phi \left( \frac{y - \mu - \frac{\sigma^2}{\lambda}}{\sigma} \right) \quad (14)$$

where  $\Phi$  denotes the distribution function of the standard normal distribution,  $\mu$  and  $\sigma$  are the mean and standard deviation respectively of the normal part of the distribution and  $\lambda$  the mean of the exponential part. Here,  $\lambda = \mu = \gamma$  and  $\sigma = \frac{\gamma \sigma_{\rho,c}}{\sqrt{k_R}}$ .

There is no known analytical expression for the distribution function and quantile function of an Exponential Gaussian distribution, but for a fixed probability  $p$ , a quantile can be calculated using (statistical) software. We used in R the package GAMLSS.DIST and the function `qexGAUS` to calculate the desired quantile. In the calculations  $\gamma$  is estimated by  $\hat{\gamma}_{k_R,n}^R$ , and the quantile found by using the statistical software needs to be multiplied by  $\frac{n}{k_R}$  to find the desired cutoff value for the  $EIF^R$ .

In case  $y$  is a certain data point  $x_{n-j,n}$ , with  $j < k_R$ , we can no longer assume independence between  $\log Y/x_{n-k_R,n}$  and  $\hat{\gamma}_{k_R,n}^R$ , because  $x_{n-j,n}$  was used to calculate  $\hat{\gamma}_{k_R,n}^R$ . This implies that the distribution of  $\log Y/x_{n-k_R,n} - \hat{\gamma}_{k_R,n}^R$  is no longer known. To overcome this problem, the data point  $x_{n-j,n}$  could be omitted in the calculations to estimate  $\gamma$ , which

would result in a jackknife estimator based on  $k_R - 1$  data points. This jackknife estimator also follows an asymptotically normal distribution, but with mean  $\gamma$  and variance  $\frac{\gamma^2 \sigma_{\rho,c}^2}{k_R - 1}$ , which leads to a slightly altered density function of  $\log Y/x_{n-k_R,n} - \hat{\gamma}_{k_R,n}^R$  and hence a slightly different cutoff value. To compute the desired quantile,  $\gamma$  should then be estimated by the jackknife estimator. Simulation results however show very little differences between the cutoff value when the jackknife estimator is used and the original one. This is in line with the expectations since we are working with a robust estimator, so we do not expect a large difference between the two estimators. To avoid long calculation times, we therefore work with the original  $\hat{\gamma}_{k_R,n}^R$  as a good approximation for the jackknife estimator. This implies that we take as cutoff value  $Q^R(p)$  the  $p$ -th quantile of an Exponential Gaussian distribution with density function (14) and parameters  $(\gamma, \frac{\gamma \sigma_{\rho,c}}{\sqrt{k_R}}, \gamma)$ , in which  $\gamma$  is estimated by  $\hat{\gamma}_{k_R,n}^R$ .

Summarized, to flag data points that are highly influential for the Hill estimator, we use the following procedure:

1. Find  $\hat{\gamma}_{k_R,n}^R$  as the estimate closest to the median over all  $\hat{\gamma}_{k,n}^R$  and the corresponding  $k_R$ .
2. Plot the  $EIF^R(x_{n-j+1,n}, \hat{\gamma}_{k_R,n}^H, \hat{F}_n)$  for the  $k_R$  largest data points  $x_{n-j+1,n}$  with  $j = 1, \dots, k_R$ .
3. Choose a high probability  $p$  (99%, 99.5%, 99.99%, ..., the larger the data set, the higher the probability should be) and compute the cutoff value  $Q^R(p)$  for the  $EIF^R$ .
4. If the  $EIF^R(x_{n-j+1,n}, \hat{\gamma}_{k_R,n}^H, \hat{F}_n) > Q^R(p)$ , flag  $x_{n-j+1}$  as highly influential.

## 5 Real data examples

### 5.1 Hedge fund data

We focus again on the Hedge fund data introduced in Section 4.1. The  $EIF^R$  for the  $k_R = 25$  largest data points was already shown in Figure 9(b). On this plot we have added the 99%

cutoff value for the  $EIF^R$ , which is equal to 0.07799, the 99.5% cutoff value being 0.08745 and the 99.9% cutoff value 0.1094. It is clear from the Figure that the largest data point has a value for the  $EIF^R$  that well exceeds these cutoff values, whereas the two next largest exceed the two lowest cutoff values, and therefore all three observations can be flagged as highly influential for the Hill estimator. Note that Dupuis and Victoria-Feser [2006] show that the WMLE estimator downweights the largest and third largest observation, but no indication is given on how much these observations deviate from the postulated model.

We could also define  $EIF^H$  and  $EIF^{ML}$  by replacing  $\hat{\gamma}_{k_R,n}^R$  in (13) with  $\hat{\gamma}_{k_{opt},n}^H$  and  $\hat{\gamma}_{k_{ML},n}^{ML}$  respectively. As both  $\hat{\gamma}_{k,n}^H$  and  $\hat{\gamma}_{k,n}^{ML}$  are known to be asymptotically normal with mean  $\gamma$  and variance  $\frac{\gamma^2}{k}$  and  $\left(\frac{1-\rho}{\rho}\right)^4 \frac{\gamma^2}{k}$  respectively [Beirlant *et al.*, 2004], we could also derive appropriate cutoff values from the Exponential Gaussian distribution. In Figure 10(a) we can see that when the Hill estimator is plugged in in the  $EIF$ , the 99.9% cutoff value becomes larger (0.1215), as well as the 99.5% (0.09677) and 99% (0.08614) cutoff values, and the  $EIF^H$  values show less variability. As a result none of the data points are flagged as influential for the 99.9% cutoff value, and only the largest data point exceeds the two other cutoff values. The  $EIF^{ML}$  shows a similar though slightly better result in Figure 10(b): the cutoff values become 0.1245 (99.9%), 0.09990 (99.5%) and 0.08929 (99%), and are slightly larger than the cutoff values found while using the Hill estimator, but so are the  $EIF^{ML}$  values, so that now the largest data point is flagged for all cutoff values (but less prominent than in the robust analysis). The other two data points however do not exceed any cutoff line.

To conclude, we see that it is important to avoid the masking effect of the Hill (and the ML estimator) by using a robust estimator for the extreme value index.

It was suggested by an anonymous referee that the robust thresholds should be compared to confidence bands for QQ-plots. In order to do so we show in Figure 11 a Pareto QQ-plot for the  $k_R = 25$  largest data points with a robust regression line and a 95% pointwise confidence band calculated by parametric bootstrap. It is clear that the three data points flagged as influential on the EIFP, are lying well within the confidence bands, so that this

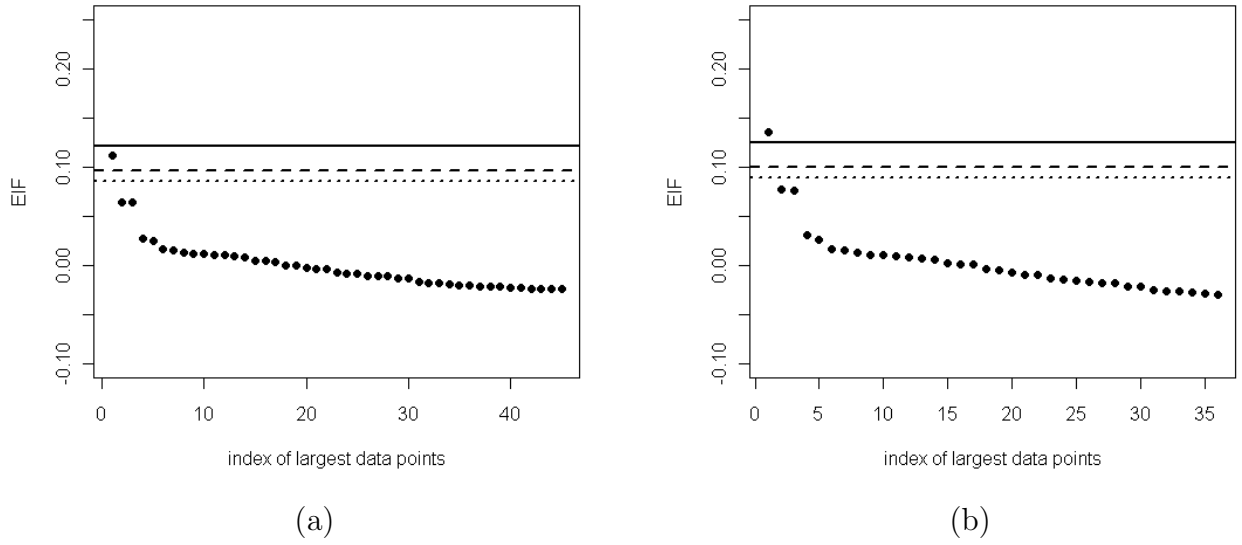


Figure 10: Plot of (a)  $EIF^H$  and (b)  $EIF^{ML}$  for the Hedge Fund data with corresponding cutoff values.

method can not be used to flag influential data points. Note that there are also no data points flagged as influential when a different threshold was chosen. The results are not shown here.

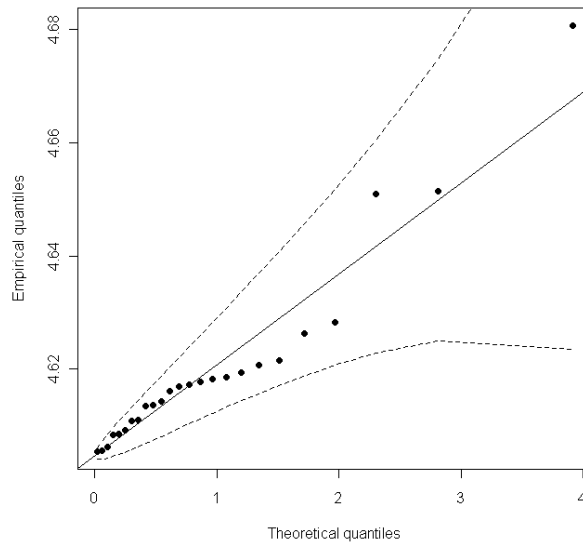


Figure 11: Pareto QQ-Plot for the  $k_R = 25$  largest observations in the Hedge fund data with robust regression line and 95% pointwise confidence bands.

## 5.2 Norwegian fire insurance data

The Norwegian fire insurance data, treated by Beirlant *et al.* [1996], consist of claim values ( $\times 1000$  NOK) and year of occurrence for the period 1972–1992. In this example, we will only consider the claims from 1987, yielding 767 observations. The Pareto QQ-plot in Figure 12(a) shows that the data are very likely to come from a Pareto-type distribution, and none of the observations seem to be outlying.

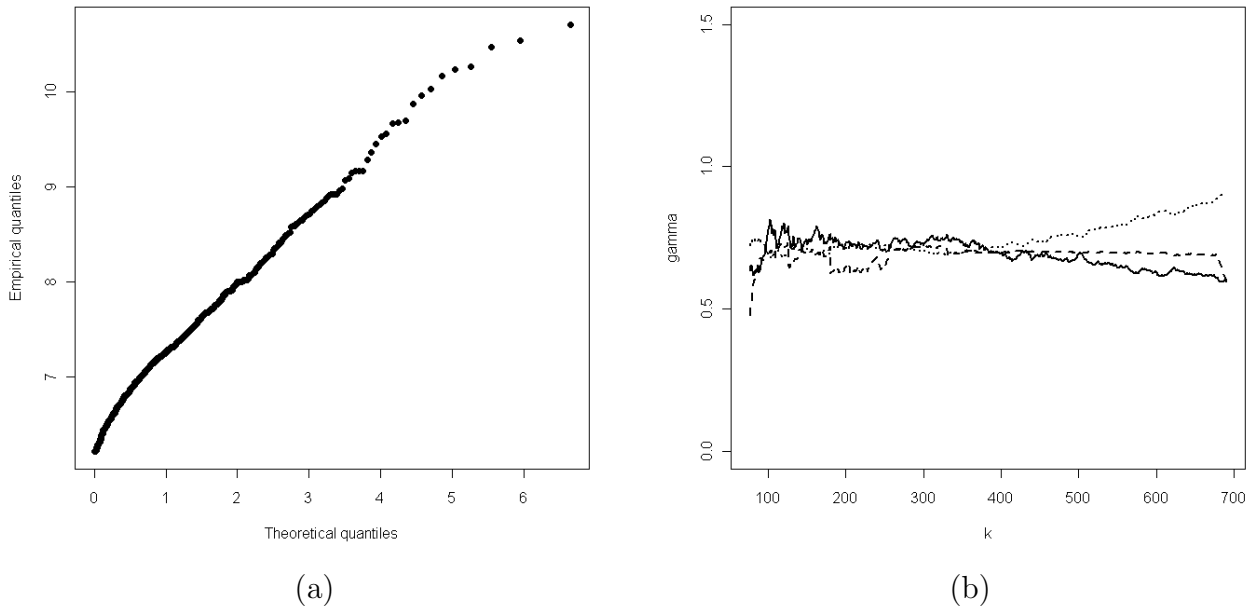


Figure 12: (a) Pareto QQ-plot of the 1987 Norwegian fire insurance data ; (b) plot of  $\gamma$  as a function of  $k$  for the Norwegian fire insurance data for the Hill estimator (dotted line), Maximum Likelihood estimator (dashed line) and robust estimator with  $c = 1.105$  (full line).

Since it seems unlikely to find influential data points, we expect the robust estimator to be close to the Hill and Maximum Likelihood estimator. It can already be seen from the plot of  $\gamma$  as a function of  $k$  in Figure 12(b) that there is very little deviation between the robust estimator (full line) and the Maximum Likelihood estimator (dashed line). Both estimators seem quite stable for all  $k$ -values. Also the Hill estimator (dotted line) returns values very comparable to the other two, but shows some bias for the larger  $k$ 's. For  $c = 1.105$ , by applying the ad hoc rule as before, the robust estimator yields  $\hat{\gamma}_{k_R, n}^R = 0.7055$

( $k_R = 384$ ) whereas the non-robust methods estimate  $\hat{\gamma}_{k_{ML},n}^{ML} = 0.6978$  (with the same ad hoc rule,  $k_{ML} = 165$ ) and  $\hat{\gamma}_{k_{opt},n}^H = 0.7300$  (for  $k_{opt} = 77$ ). Note that changing the value of  $c$  barely changes the value of  $\hat{\gamma}_{k_R,n}^R$  (e.g. for  $c = 1.825$ , which yields an ARE around 90%,  $k_R = 392$  and  $\hat{\gamma}_{k_R,n}^R = 0.6989$ ), which can be explained by the fact that if there are no influential observations present, there will be no need to downweight any of the log-spacings  $Z_j$ .

The absence of influential data points is also confirmed by the  $EIF^R$  of the 384 largest observations, which is plotted in Figure 13. Also the cutoff value for the 99.999% (full line), 99.995% (dashed line) and 99.99% (dotted line) quantiles are shown. Note that we need to look at high quantiles here, since the data set is rather large.

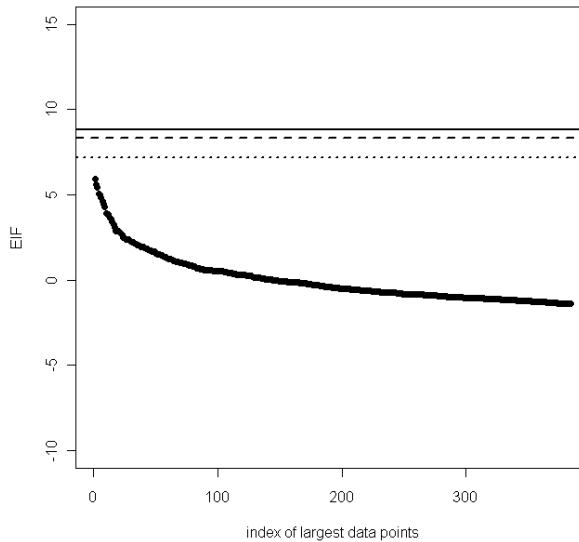


Figure 13:  $EIF^R$  plot for the Norwegian fire insurance data, with 99.999% cutoff value (full line), 99.995% cutoff value (dashed line) and 99.99% cutoff value (dotted line)

## 6 Conclusion

In this paper we have constructed a graphical tool to flag highly influential data points for the Hill estimator. This allows us to mark unusual extreme data points which can then be

investigated thoroughly. To obtain our results, we derived the influence function of the Hill estimator of Pareto-type distributions and a robust estimator for the extreme value index.

In further research we will investigate how these results can be used to construct a new boxplot for Pareto-type distributions, thereby extending the ideas of the adjusted boxplot for skewed distributions [Hubert and Vandervieren, 2008].

It will also be interesting to explore how our robust estimator and EIF with cutoff values could be extended to Weibull type distributions that satisfy  $\bar{F}(x) = \exp(-x^\alpha \ell(x))$ , with  $\alpha > 0$  and  $\ell(x)$  a slowly varying function, which is a subclass of the Gumbel class for which  $\gamma = 0$ . We could base our research on the non-robust, bias reduced estimator based on the mean excess function that was proposed by Dierckx *et al.* [2009].

### **Acknowledgements**

We would like to express our gratitude towards professor E. Cantoni for kindly providing us with the R-code for the robust GLM procedure, which we slightly modified. We are also very grateful to professor M.P. Victoria-Feser who kindly shared with us her R-code for calculating the WMLE estimator, as well as the Hedge fund data.

Mia Hubert acknowledges the financial support of the FWO research grant G.0436.08N.

## **7 Appendix**

### **7.1 Influence function of the Hill estimator**

First we calculate the exact influence function of the Hill estimator in general. Next we calculate an approximation using the fact that we are working with Pareto-type distributions.

The influence function of a functional  $T(F)$  can be calculated as  $\frac{\partial}{\partial \varepsilon} T(F_{\varepsilon, \Delta_y})|_{\varepsilon=0}$ , so we start with calculating  $T(F_{\varepsilon, \Delta_y})$  for the functional defined in (2). Denote with  $\alpha$  the percentage of the population that is used to estimate  $\gamma$ . It can easily be verified that for

$y \leq F^{-1}(1 - \alpha)$ , the influence function is equal to 0. For  $y > F^{-1}(1 - \alpha)$ , with  $\varepsilon$  sufficiently small, it holds that  $y > F_{\varepsilon, \Delta_y}^{-1}(1 - \alpha)$ , and  $F_{\varepsilon, \Delta_y}^{-1}(1 - \alpha) = F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)$ . Using this result we calculate

$$\begin{aligned} T(F_{\varepsilon, \Delta_y}) &= \frac{1}{\alpha} \int_{F_{\varepsilon, \Delta_y}^{-1}(1-\alpha)}^{+\infty} \log x dF_{\varepsilon, \Delta_y}(x) - \log F_{\varepsilon, \Delta_y}^{-1}(1 - \alpha) \\ &= \frac{1 - \varepsilon}{\alpha} \int_{F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)}^{+\infty} \log x dF(x) + \frac{\varepsilon}{\alpha} \log y - \log F^{-1}\left(\frac{1 - \alpha}{1 - \varepsilon}\right) \end{aligned}$$

and

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} T(F_{\varepsilon, \Delta_y}) &= \frac{\partial}{\partial \varepsilon} \left( \frac{1 - \varepsilon}{\alpha} \right) \int_{F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)}^{+\infty} \log x dF(x) + \frac{1 - \varepsilon}{\alpha} \frac{\partial}{\partial \varepsilon} \left( \int_{F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)}^{+\infty} \log x dF(x) \right) \\ &\quad + \frac{\partial}{\partial \varepsilon} \left( \frac{\varepsilon}{\alpha} \log y \right) - \frac{\partial}{\partial \varepsilon} \log F^{-1}\left(\frac{1 - \alpha}{1 - \varepsilon}\right) \\ &= -\frac{1}{\alpha} \int_{F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)}^{+\infty} \log x dF(x) + \frac{1 - \varepsilon}{\alpha} \frac{\partial}{\partial \varepsilon} \left( \int_{F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)}^{+\infty} \log x dF(x) \right) \\ &\quad + \frac{1}{\alpha} \log y - \frac{\partial}{\partial \varepsilon} \log F^{-1}\left(\frac{1 - \alpha}{1 - \varepsilon}\right). \end{aligned} \tag{15}$$

According to Leibniz' rule we find that

$$\frac{\partial}{\partial \varepsilon} \left( \int_{F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)}^{+\infty} \log x dF(x) \right) = -\log F^{-1}\left(\frac{1 - \alpha}{1 - \varepsilon}\right) \frac{1 - \alpha}{(1 - \varepsilon)^2},$$

and moreover

$$\frac{\partial}{\partial \varepsilon} \log F^{-1}\left(\frac{1 - \alpha}{1 - \varepsilon}\right) = \frac{1}{F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)} \frac{1}{F'\left(F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)\right)} \frac{1 - \alpha}{(1 - \varepsilon)^2}.$$

Putting everything together allows us to rewrite (15) in the following way:

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} T(F_{\varepsilon, \Delta_y}) &= -\frac{1}{\alpha} \int_{F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)}^{+\infty} \log x dF(x) - \frac{1 - \varepsilon}{\alpha} \log F^{-1}\left(\frac{1 - \alpha}{1 - \varepsilon}\right) \frac{1 - \alpha}{(1 - \varepsilon)^2} + \frac{1}{\alpha} \log y \\ &\quad - \frac{1}{F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)} \frac{1}{F'\left(F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right)\right)} \frac{1 - \alpha}{(1 - \varepsilon)^2}. \end{aligned}$$

Finally, considering that  $IF(y; T, F) = \frac{\partial}{\partial \varepsilon} T(F_{\varepsilon, \Delta_y}) |_{\varepsilon=0}$ , we obtain

$$IF(y; T, F) = \frac{-\frac{1}{\alpha} \int_{F^{-1}(1-\alpha)}^{+\infty} \log xdF(x) - \frac{1-\alpha}{\alpha} \log F^{-1}(1-\alpha) + \frac{1}{\alpha} \log y}{\frac{1-\alpha}{F^{-1}(1-\alpha)F'(F^{-1}(1-\alpha))}}. \quad (16)$$

For Pareto-type distributions,  $\bar{F}$  can be written in terms of a certain  $\gamma$  and a slowly varying function:  $\bar{F}(x) = x^{-1/\gamma} \ell(x)$ . The definition of slowly varying functions yields

$$\frac{\bar{F}(uy)}{\bar{F}(u)} = \frac{P(X > uy)}{P(X > u)} = \frac{(uy)^{-1/\gamma} \ell(uy)}{u^{-1/\gamma} \ell(u)} = y^{-1/\gamma} \frac{\ell(uy)}{\ell(u)} \sim y^{-1/\gamma} \text{ ( for } u \rightarrow +\infty),$$

from which it follows that  $\bar{F}(x) \sim \left(\frac{x}{u}\right)^{-1/\gamma} \bar{F}(u)$ . For the choice  $u = F^{-1}(1-\alpha)$  it holds that  $\bar{F}(u) = \alpha$ , so that

$$F(x) \sim 1 - \left(\frac{x}{F^{-1}(1-\alpha)}\right)^{-1/\gamma} \alpha \quad (17)$$

for  $\alpha \rightarrow 0$ . This expression for  $F$  can be used to rewrite the different terms of (16). For the first term, it follows directly from (2) that

$$\int_{F^{-1}(1-\alpha)}^{+\infty} \log xdF(x) \sim \alpha (\log F^{-1}(1-\alpha) + \gamma),$$

where  $T_\alpha(F)$  is approximated by  $\gamma$ .

Deriving (17) with respect to  $x$  leads to  $F'(x) \sim \frac{1}{\gamma} (F^{-1}(1-\alpha))^{1/\gamma} x^{-1/\gamma-1} \alpha$ , so that  $F'(F^{-1}(1-\alpha)) \sim \alpha / (\gamma F^{-1}(1-\alpha))$ . This allows us to rewrite the last term of (16):

$$\frac{1-\alpha}{F^{-1}(1-\alpha)F'(F^{-1}(1-\alpha))} \sim \frac{1-\alpha}{F^{-1}(1-\alpha) \frac{\alpha}{\gamma F^{-1}(1-\alpha)}} \sim \frac{1-\alpha}{\alpha} \gamma.$$

Combining these results provides us with an approximation for the influence function of the

Hill estimator for Pareto-type distributions:

$$\begin{aligned} IF(y, T, F) &\sim -\log F^{-1}(1 - \alpha) - \gamma - \frac{1 - \alpha}{\alpha} \log F^{-1}(1 - \alpha) + \frac{\log y}{\alpha} - \frac{1 - \alpha}{\alpha} \gamma \\ &= -\frac{1}{\alpha} (\log F^{-1}(1 - \alpha) - \log y + \gamma) \end{aligned}$$

for  $y > F^{-1}(1 - \alpha)$ .

## 7.2 Asymptotic normality of $\hat{\beta}$

In Cantoni and Ronchetti [2001] it is shown that  $\hat{\beta}$  is asymptotically normal with asymptotic variance  $M^{-1}QM^{-1}$ . Here,  $M = \frac{1}{k}X^tBX$  where  $B$  is a diagonal matrix with elements  $b_j = \mu_j E \left[ \psi_c(r_j) \frac{\partial}{\partial \mu_j} \log f_{Z_j}(z_j | \mathbf{u}_j, \mu_j) \right]$  and  $Q = \frac{1}{k}X^tAX - \alpha(\beta)\alpha(\beta)^t$  where  $A$  is a diagonal matrix with elements  $a_j = E[\psi_c^2(r_j)]$ . We now compute the different terms at the exponential distribution  $Z_j \sim \text{Exp}(\mu_j)$ .

### 1. Calculation of $E[\psi_c(r_j)]$ .

Define  $r_j = \frac{Z_j - \mu_j}{\mu_j}$  with  $Z_j \sim \text{Exp}(\mu_j)$  which means that  $f_{Z_j}(z_j) = \frac{1}{\mu_j} \exp(-\frac{z_j}{\mu_j})$ . The condition  $|\frac{z_j - \mu_j}{\mu_j}| \leq c$  in the Huber function is equivalent to  $\mu_j(1 - c) \leq z_j \leq \mu_j(c + 1)$ . Usually  $c$  is chosen  $\geq 1$ , so that the left side would be  $\leq 0$ , but since  $Z_j \sim \text{Exp}(\mu_j)$ ,  $z_j$  has to be  $\geq 0$ , which leads to the conclusion that  $|\frac{z_j - \mu_j}{\mu_j}| \leq c$  is equal to  $0 \leq z_j \leq \mu_j(c + 1)$  if  $c \geq 1$ .

Knowing this we can write

$$\begin{aligned} E[\psi_c(r_j)] &= \int_0^{+\infty} \psi_c(r_j) f_{Z_j}(z_j) dz_j \\ &= \int_0^{\mu_j(c+1)} \frac{z_j - \mu_j}{\mu_j} \frac{1}{\mu_j} e^{-z_j/\mu_j} dz_j + c \int_{\mu_j(c+1)}^{+\infty} \frac{1}{\mu_j} e^{-z_j/\mu_j} dz_j = -e^{-(c+1)}. \end{aligned}$$

### 2. Calculation of $\alpha(\beta)$ .

Remembering that  $\mu_j = \exp(\mathbf{u}_j^t \boldsymbol{\beta}) = \exp(\beta_0 + (\frac{j}{k+1})^{-\rho} \beta_1)$ , we easily derive that

$$\frac{\boldsymbol{\mu}'_j}{\mu_j} = \frac{1}{\mu_j} \begin{pmatrix} \frac{\partial}{\partial \beta_0} \mu_j \\ \frac{\partial}{\partial \beta_1} \mu_j \end{pmatrix} = \begin{pmatrix} 1 \\ (\frac{j}{k+1})^{-\rho} \end{pmatrix}$$

from which it follows that

$$\alpha(\boldsymbol{\beta}) = \frac{1}{k} \sum_{j=1}^k E[\psi_c(r_j)] \frac{\boldsymbol{\mu}'_j}{\mu_j} = - \begin{pmatrix} 1 \\ d_1 \end{pmatrix} e^{-(c+1)}.$$

3. Calculus yields that

$$\begin{aligned} E[\psi_c^2(r_j)] &= \int_0^{+\infty} \psi_c^2(r_j) f_{Z_j}(z_j) dz_j \\ &= \int_0^{\mu_j(c+1)} \left( \frac{z_j - \mu_j}{\mu_j} \right)^2 \frac{1}{\mu_j} e^{-z_j/\mu_j} dz_j + c^2 \int_{\mu_j(c+1)}^{+\infty} \frac{1}{\mu_j} e^{-z_j/\mu_j} dz_j \\ &= 1 - 2(c+1)e^{-(c+1)}. \end{aligned}$$

4. Calculation of  $E[\psi_c(r_j) \frac{\partial}{\partial \mu_j} \log f_{Z_j}(z_j)]$ .

First note that  $\log f_{Z_j}(z_j) = -\frac{z_j}{\mu_j} - \log \mu_j$  and  $\frac{\partial}{\partial \mu_j} \log f_{Z_j}(z_j) = \frac{z_j}{\mu_j^2} - \frac{1}{\mu_j} = \frac{z_j - \mu_j}{\mu_j} \frac{1}{\mu_j} = \frac{1}{\mu_j} r_j$ ,

so that

$$\begin{aligned} E[\psi_c(r_j) \frac{\partial}{\partial \mu_j} \log f_{Z_j}(z_j)] &= \frac{1}{\mu_j} E[\psi_c(r_j) r_j] \\ &= \frac{1}{\mu_j} \int_0^{\mu_j(c+1)} \left( \frac{z_j - \mu_j}{\mu_j} \right)^2 \frac{1}{\mu_j} e^{-z_j/\mu_j} dz_j + \frac{c}{\mu_j} \int_{\mu_j(c+1)}^{+\infty} \frac{z_j - \mu_j}{\mu_j} \frac{1}{\mu_j} e^{-z_j/\mu_j} dz_j \\ &= \frac{1}{\mu_j} (1 - (2+c)e^{-(c+1)}). \end{aligned}$$

5. Calculation of the asymptotic variance of  $\hat{\boldsymbol{\beta}}$ .

With  $X$  being a  $k$  by 2 matrix where  $X(j, 1) = 1$  and  $X(j, 2) = (\frac{j}{k+1})^{-\rho}$  for  $j = 1, \dots, k$  and  $B$  being a diagonal  $k$  by  $k$  matrix with elements  $b$ , it is easy to verify that

$$M = \frac{1}{k}X^tBX = b \begin{pmatrix} 1 & d_1 \\ d_1 & d_2 \end{pmatrix}.$$

Similar calculations can be done to show that  $\frac{1}{k}X^tAX = a \begin{pmatrix} 1 & d_1 \\ d_1 & d_2 \end{pmatrix}$  with  $a$  the elements of the  $k$  by  $k$  diagonal matrix  $A$ , which yields

$$\begin{aligned} Q &= \frac{1}{k}X^tAX - \alpha(\boldsymbol{\beta})\alpha(\boldsymbol{\beta})^t \\ &= \begin{pmatrix} a - e^{-2(c+1)} & d_1(a - e^{-2(c+1)}) \\ d_1(a - e^{-2(c+1)}) & ad_2 - d_1^2e^{-2(c+1)} \end{pmatrix}. \end{aligned}$$

The asymptotic variance of  $\hat{\boldsymbol{\beta}}$  can now be calculated: Computing  $M^{-1}QM^{-1}$  finally gives

$$\frac{1}{b^2(d_2 - d_1^2)} \begin{pmatrix} d_2(a - c_2^2) + d_1^2c_2^2 & -ad_1 \\ -ad_1 & a \end{pmatrix}.$$

## References

- J. Beirlant, P. Vynckier, and J.L. Teugels. *Practical Analysis of Extreme Values*. Leuven University Press, 1996.
- J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys. Tail index estimation and an exponential regression model. *Extremes*, 2:177–200, 1999.
- J. Beirlant, G. Dierckx, A. Guillou, and C. Stărică. On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5(2):157–180, 2002.
- J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of Extremes: Theory and Applications*. Wiley, Chichester, 2004.
- V. Brazauskas and R.J. Serfling. Robust and efficient estimation of the tail index of a single-parameter Pareto distribution. *North American Actuarial Journal*, 4:12–27, 2000.

- E. Cantoni and E. Ronchetti. Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96:1022–1030, 2001.
- E. Cantoni and E. Ronchetti. A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditure. *Journal of Health Economics*, 25:198–213, 2006.
- M. Debruyne, M. Hubert, and J. Van Horebeek. Detecting influential observations in kernel PCA. *Computational Statistics and Data analysis*, 54:3007–3019, 2009.
- R. Dell’Aquila and P. Embrechts. Extremes and robustness: a contradiction? *Financial Markets and Portfolio Management*, 20:103–118, 2006.
- G. Dierckx, J. Beirlant, D. De Waal, and A. Guillou. A new estimation method for Weibull-type tails based on the mean excess function. *Journal of Statistical Planning and Inference*, 139:1905–1920, 2009.
- H. Drees and E. Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75(2):149–172, 1998.
- W.H. DuMouchel. Estimating the stable index  $\alpha$  in order to measure tail thickness: a critique. *The Annals of Statistics*, 11(4):1019–1031, 1983.
- D.J. Dupuis and C.A. Field. Robust estimation of extremes. *The Canadian Journal of Statistics*, 26(2):199–215, 1998.
- D.J. Dupuis and S. Morgenthaler. Robust weighted likelihood estimators with an application to bivariate extreme value problems. *The Canadian Journal of Statistics*, 30(1):17–36, 2002.
- D.J. Dupuis and M.-P. Victoria-Feser. A robust prediction error criterion for Pareto modelling of upper tails. *The Canadian Journal of Statistics*, 34(4):639–658, 2006.

- R.A. Fisher and L.H.C. Tippett. On the estimation of the frequency distributions of the largest or smallest members of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190, 1928.
- M.I. Fraga Alves, M.I. Gomes, and L. de Haan. A new class of semi-parametric estimators of the second order parameter. *Portugaliae Mathematica*, 60(2):193–213, 2003.
- B.V. Gnedenko. Sur la distribution limite du terme maximum d’une série aléatoire. *Annals of Mathematics*, 44:423–453, 1943.
- M.I. Gomes and O. Oliveira. Maximum likelihood revisited under a semi-parametric context—estimation of the tail index. *Journal of Statistical Computation and Simulation*, 73(4):285–301, 2003.
- A. Guillou and P. Hall. A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society B*, 63:293–305, 2001.
- F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York, 1986.
- S. Heritier, E. Cantoni, S. Copt, and M.-P. Victoria-Feser. *Robust methods in Biostatistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2009.
- B.M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.
- M. Hubert and E. Vandervieren. An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52(12):5186–5201, 2008.
- L.A. Jaeckel. The infinitesimal jackknife. *Bell Laboratories Memorandum MM*, 72–1215–11, 1972.
- S.F. Juárez and W.R. Schucany. Robust and efficient estimation for the generalized Pareto distribution. *Extremes*, 7(3):237–251, 2004.

- C.L. Mallows. On some topics of robustness. *Technical Bell Memorandum, Bell Telephone Laboratories, Murray Hill, NJ*, 1975.
- R.A. Maronna, D.R. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, New York, 2006.
- G. Matthys and J. Beirlant. Adaptive threshold selection in tail index estimation. In *Extremes and Integrated Risk Management*, pages 37–49, UBS Warburg, 2000. Ed. Paul Embrechts, Risk Books.
- P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, London, 1983.
- C. Perret-Gentil and M.P. Victoria-Feser. Robust mean-variance portfolio selection. *Cahiers du Département d’Économétrie*, 2003.3, 2003.
- G. Pison and S. Van Aelst. Diagnostic plots for robust multivariate methods. *Journal of Computational and Graphical Statistics*, 13:310–329, 2004.
- G. Pison, P.J. Rousseeuw, P. Filzmoser, and C. Croux. Robust factor analysis. *Journal of Multivariate Analysis*, 84:145–172, 2003.
- R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley, New York.
- B. Vandewalle, J. Beirlant, and M. Hubert. A robust estimator of the tail index based on an exponential regression model. In M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, editors, *Theory and Applications of Recent Robust Methods*, pages 367–376, Basel, 2004. Statistics for Industry and Technology, Birkhauser.
- B. Vandewalle, J. Beirlant, A. Christmann, and M. Hubert. A robust estimator for the tail index of Pareto-type distributions. *Computational Statistics and Data Analysis*, 51:6252–6268, 2007.
- M.-P. Victoria-Feser and E. Ronchetti. Robust methods for personal-income distribution models. *The Canadian Journal of Statistics*, 22:247–258, 1994.