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

Mia Hubert^{a,*}, Goedele Dierckx^{a,b}, Dina Vanpaemel^a

^a*KU Leuven, Department of Mathematics and Leuven Statistics Research Center (LStat), Celestijnenlaan 200B, BE-3001 Heverlee, Belgium.*

^b*Hogeschool-Universiteit Brussel, Stormstraat 2, BE-1000 Brussel, Belgium.*

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. The empirical influence function plot, which displays the influence each data point has on the Hill estimator, is introduced. To avoid a masking effect, the empirical influence function is based on a new robust GLM estimator for γ . This robust GLM estimator is used to determine high quantiles of the data generating distribution, allowing to flag data points as unusually large if they exceed this high quantile.

Keywords: Pareto-type distribution, extreme value index, tail index estimation, influential data points, robustness.

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 those sequences a_n and b_n , the limit distribu-

*Corresponding author at: KU Leuven, Department of Mathematics and Leuven Statistics Research Center (LStat), Celestijnenlaan 200B, BE-3001 Heverlee, Belgium. Tel.: +32 16 32 20 23. Fax: +32 16 32 79 98

Email address: Mia.Hubert@wis.kuleuven.be (Mia Hubert)

tion H is of the extreme value type:

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

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_γ has only one parameter, γ , 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.

Our focus will be on these Pareto-type distributions. A very popular estimator of γ 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 a widely used estimator to estimate the tail index (see e.g. Cai et al. (2011), Ferreira and Canto e Castro (2010), McElroy and Agnieszka (2012)). 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, along with a method of flagging data points as “unusual”. Points that are flagged as unusual and that are highly influential for the Hill estimator deserve more attention, as they might be responsible for 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), which still depends on the unknown extreme value index γ . To avoid the masking effect, we then robustly estimate γ . Our robust GLM estimator will be derived from the robust M-estimator for Generalized Linear Models of Cantoni and Ronchetti (2001), and will allow us to calculate high quantiles of the data generating distribution in a robust way, based on a quantile estimator proposed by Weissman (1978). A data point that exceeds

this high quantile is most likely an unusual observation, eventually coming from another distribution. Next we look at a diagnostic plot of the EIF to see whether these unusual observations also have an abnormal large influence on the Hill estimator.

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 the 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 value. Whether these unusual points concern some outliers eventually coming from another distribution or not, often will then be decided on by specialists in the field of the data.

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). In their words “... robust methods can improve the quality of extreme value data analysis by providing information on influential observations, deviating substructures and possible misspecifications of a model ...”.

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 GLM estimator for γ , derive its asymptotic distribution, study some robustness properties, propose a robust criterion to select the optimal number of data points that should be used to compute the estimator and illustrate the performance of our robust GLM estimator in

a small simulation study. In Section 4 we derive high quantiles for the data generating distribution in a robust way and combine this with the empirical influence function on a diagnostic plot to visualize which data points are unusually large and are highly influential for the Hill estimator. 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.

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 ℓ_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 $U(x) = F^{-1}(1 - 1/x)$ for $x > 1$ with F^{-1} the quantile function of F , i.e. $U(x) = x^\gamma \ell_U(x)$ with ℓ_U again a slowly varying function.

For strict Pareto distributions with $\bar{F}(x) = x^{-1/\gamma}$ for $\gamma > 0$, $\log(X)$ follows an exponential distribution with mean γ . 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 γ 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(θ) distribution is given by $\bar{F}(x) = 1 - \exp(-x^{-\theta})$ for $x > 0, \theta > 0$, and $\gamma = \frac{1}{\theta}$.

Since the slope of the Pareto QQ-plot approximates the unknown parameter γ , an

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).

estimate of the slope can be used to estimate γ . 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 γ . 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 (IF) 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 $\lim_{\varepsilon \rightarrow 0} \frac{T(F_{\varepsilon, \Delta_y}) - T(F)}{\varepsilon}$ (if the limit exists) with $F_{\varepsilon, \Delta_y} = (1 - \varepsilon)F + \varepsilon\Delta_y$ the contaminated distribution where ε represents

the percentage of contamination and Δ_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 ε .

To derive the IF of the Hill estimator at any absolutely continuous distribution function F , 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 γ . From this functional form (2) the following expression for the influence function of the Hill estimator can be derived (see Appendix):

$$IF(y, \gamma_\alpha^H, F) = \begin{cases} -T_\alpha(F) + \frac{1}{\alpha} \log \frac{y}{F^{-1}(1-\alpha)} - \frac{1-\alpha}{F^{-1}(1-\alpha)F'(F^{-1}(1-\alpha))} & \text{for } y > F^{-1}(1-\alpha) \\ -T_\alpha(F) + \frac{\alpha}{F^{-1}(1-\alpha)F'(F^{-1}(1-\alpha))} & \text{for } 0 < y \leq F^{-1}(1-\alpha). \end{cases} \quad (3)$$

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 γ can be approximated by

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

Note that this is in accordance with the influence function for the Hill estimator derived in Dupuis and Victoria-Feser (2006) for strict Pareto distributions. Figure 2 plots the exact and approximated influence function of the Hill estimator at a Fréchet(2) distribution for y -values between 0 and 10. We see that the approximated influence function lies close to the exact influence function. The approximation is best at the smallest values of α . Moreover we see that the influence function is monotone increasing for $y > F^{-1}(1-\alpha)$. This corresponds with our intuition that the largest observations have the largest influence on the estimation of the tail behavior of the distribution.

Figure 2: Exact (full line) and approximated (dotted line) influence function for y between 0 and 10 with (a) $\alpha = 0.1$, (b) $\alpha = 0.5$ and (c) $\alpha = 0.8$.

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 γ 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 is known to be very biased by highly influential data points (see e.g. Dupuis and Victoria-Feser (2006) and Vandewalle et al. (2004)). Its estimated values for γ will be considerably larger than the true value in case of contamination. This will have an impact on the EIF, which will underestimate the real influences of the observations so that those highly influential data points will not be detected as such. In robust statistics this is called a masking effect. Hence, it is important to plug in an estimator for γ 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 Debruyne et al. (2010) in the context of canonical correlation analysis and principal component analysis.

3. A robust estimator for the extreme value index

Our robust GLM estimator for γ will be based on a robust M-estimator for Generalized Linear Models proposed by Cantoni and Ronchetti (2001). First, we derive an estimator for γ within the generalized linear model framework, starting from the Maximum Likelihood estimator proposed in Beirlant et al. (1999).

3.1. The Maximum Likelihood estimator

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 for all $t > 0$: $\frac{\ell_U(tx)}{\ell_U(x)} \sim 1 + b(x) \frac{t^\rho - 1}{\rho}$. For e.g. a Fréchet(θ) distribution, it is easy to verify that $\rho = -1$ and $b(x) \sim \frac{1}{2\theta x}$ for $x \rightarrow \infty$. For a Burr(η, τ, λ) distribution with tail distribution function $\bar{F}(x) = \left(\frac{\eta}{\eta + x^\tau}\right)^\lambda$ ($x > 0; \eta, \tau, \lambda > 0$), it can be shown that $\rho = -\frac{1}{\lambda}$, $b(x) \sim \frac{1}{\lambda\tau} x^{-\frac{1}{\lambda}}$ as $x \rightarrow \infty$ (Beirlant et al., 1999; Gomes et al., 2007) and the extreme value index γ equals $\frac{1}{\lambda\tau}$. Under this mild condition on the slowly varying function ℓ_U , 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 β_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 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 γ , $b_{n,k}$ and ρ can be derived. Note that if we ignore the bias terms $b_{n,k}$, the Maximum Likelihood estimator for γ 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

ρ . Again a Maximum Likelihood estimator for γ (as well as a ML estimator for $b_{n,k}$) can be computed, denoted as $\hat{\gamma}_{k,n}^{ML}$. However the estimation of ρ 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 ρ -value rather than one of the consistent estimators, even if ρ is misspecified”. In Matthys and Beirlant (2000) it is shown that for most applications a ρ -value between -2 and 0 is most appropriate. Moreover, they show that specifically the estimates $\gamma_{k,n}^{ML}$ are not highly influenced by the specific choice of $\rho = -1$, which we will always use.

Although the ML estimator performs very well in case of uncontaminated data, it shows, just like the Hill estimator, a severe bias when contamination is present (simulation results are shown in Figure 3). To robustly estimate γ , we need a different estimator, which is based on a Generalized Linear Model (GLM).

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 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,

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

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

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

This implies that Z_j is approximately exponentially distributed with mean $\mu_j = \exp(\mathbf{u}_j^t \boldsymbol{\beta}_k)$ and variance $v(\mu_j) = \mu_j^2$, so that the link function g of this generalized linear model for μ_j is the natural logarithm. Note that the approximation in (8) is to be understood as in (6).

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

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

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

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

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

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

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 $\alpha(\beta_k) = \frac{1}{k} \sum_{j=1}^k E[\psi_c(r_j)] \frac{\mu'_j}{\mu_j}$ 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 points are taken into account to calculate the solution of the M-estimator. 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 robust 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 $\hat{\beta}_k$ has been computed, it follows immediately from (9) that $\exp(\hat{\beta}_0)$ can be used as an estimate for γ . We denote this robust GLM estimator as $\hat{\gamma}_{k,n}^{RGLM}$.

3.3. A small simulation study of the robust GLM estimator

To study the behavior of the GLM estimator and compare it to the Hill- and ML estimator, we have generated 50 data sets containing 500 data points from a Fréchet(2) distribution for the first example. We did the same for the second example, but now with data coming from a Burr(1,1,2) distribution. Remember that for both distributions $\gamma = 0.5$. In 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).

For each estimator, the medians are calculated for a range of k -values when there is no contamination present (Figure 3(a) and 3(b)), and in case there are ten contaminating data points (Figure 3(c) and 3(d)). For our robust GLM estimator, we have set the constant c in the Huber function equal to 1.105. This yields an asymptotic relative efficiency (ARE) of about 80%, as will be discussed in Section 3.5, Table 1. We also show results obtained by the robust 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 $\hat{\gamma}_{k,n}^{WMLE}$. Note that using the Pareto density function to calculate $\psi(x, \theta)$ restricts the WMLE estimator to a well-specified parametric model, whereas our robust approach only assumes a generalized Pareto distribution as a limiting distribution.

The upper left panel in Figure 3 shows for a Fréchet(2) distribution medians over 50 data sets of the Hill estimator (dotted line), the Maximum Likelihood estimator (dashed line), the WMLE estimator (dash-dotted line) with tuning constant $c_w = 2$ which ensures an ARE of about 80% (Dupuis and Victoria-Feser, 2006) and our robust GLM estimator (full line) with $c = 1.105$ yielding around the same ARE. It can be seen that the median ML estimates and the robust GLM estimates stay close to the true value of $\gamma = 0.5$, whereas the WMLE and the Hill estimator show some bias which increases with k . In case of contamination, the lower left panel of Figure 3 shows clearly that the Hill and Maximum Likelihood estimator are much affected by the contamination (yielding estimates close to 1), whereas the robust GLM estimator stays very close to the true value of 0.5. The WMLE estimates γ fairly accurate, but again shows a bias when k becomes large.

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 3. In case of no contamination both the ML and the robust GLM estimator show a very small bias, whereas the

WMLE and the Hill estimator give very poor estimates for almost all values of k . The overestimation of the Hill and Maximum Likelihood estimator in case of contamination becomes even more prominent, while the robust GLM estimator only slightly overestimates the true value of $\gamma = 0.5$. The WMLE has again a strongly increasing bias when k increases. Note that Figure 3(b) also illustrates that a misspecification of $\rho = -1$ is not crucial, as the ML and GLM estimates for γ in case of a Burr(1,1,2) distribution without contamination are very accurate for a large range of k -values, even though the real ρ -value for the Burr(1,1,2) distribution is equal to -0.5.

3.4. Asymptotic distribution of $\hat{\gamma}_{k,n}^{RGLM}$

We use the results of Cantoni and Ronchetti (2001) to derive the asymptotic distribution of $\hat{\gamma}_{k,n}^{RGLM}$ 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 $\hat{\beta}_k$ that satisfies (11), is asymptotically normal with asymptotical 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 $Q = \frac{1}{k} X^t A X - \alpha(\beta_k) \alpha(\beta_k)^t$ where A is a diagonal matrix with elements $a_j = E[\psi_c^2(r_j)]$.

It can be verified 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 1, so we choose $c \geq 1$ from now on. It is also easy to verify 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 leads to the formulation of the asymptotic normality for $\hat{\beta}_k$:

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

from which it follows that $\sqrt{k}(\hat{\beta}_0 - \beta_0) \xrightarrow{d} N(0, \sigma_{\rho,c}^2)$ with $\sigma_{\rho,c}^2 = \frac{(a - c_2^2)d_2 + d_1^2 c_2^2}{b^2(d_2 - d_1^2)}$. Using

Figure 3: Medians over 50 Hill estimates (dotted line), ML estimates (dashed line), WMLE estimates (dash-dotted line) with $c_w = 2$ and RGLM estimates (full line) with $c = 1.105$ for γ for uncontaminated data with $n = 500$ coming from a (a) Fréchet(2) distribution; (b) Burr(1,1,2) distribution; (c) Fréchet(2) distribution with 10 contaminating points and (d) Burr(1,1,2) distribution with 10 contaminating points.

$\hat{\gamma}_{k,n}^{RGLM} = \exp(\hat{\beta}_0)$, we can write this as $\sqrt{k}(\log \hat{\gamma}_{k,n}^{RGLM} - \log \gamma) \xrightarrow{d} N(0, \sigma_{\rho,c}^2)$, which implies that the distribution of $\hat{\gamma}_{k,n}^{RGLM}$ can be approximated by a lognormal distribution (for $k, n \rightarrow \infty$) with variance $\gamma^2 \exp(\frac{\sigma_{\rho,c}^2}{k} - 1) \exp(\frac{\sigma_{\rho,c}^2}{k})$.

In the case of a Pareto-type distribution, where the Z_j are only approximately exponentially distributed, this result will hold approximately.

3.5. Robustness properties

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

$$IF(y; \hat{\beta}_k, G_j) = - \left(E \left[\frac{\partial}{\partial \beta_k} \Psi(y, \mu_j) \right] \right)^{-1} \Psi(y, \mu_j),$$

where G_j behaves asymptotically as an exponential distribution with mean $\mu_j = \exp(\mathbf{u}_j^t \beta_k)$. From the choice of Ψ in (12), the bounded Huber function ψ_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 y , we obtain a bounded influence function for $\hat{\beta}_k$, and hence for $\hat{\beta}_0 = \log(\hat{\gamma}^{RGLM})$. The influence function for $\hat{\gamma}^{RGLM}$ equals

$$\begin{aligned} IF(y, \hat{\gamma}^{RGLM}, G_j) &= \gamma IF(y, \hat{\beta}_0, G_j) \\ &= \frac{\gamma \left(d_2 - \left(\frac{j}{k+1} \right)^{-\rho} d_1 \right)}{b(d_2 - d_1^2)} \psi_c \left(\frac{y - \mu_j}{\mu_j} \right) - \frac{\gamma}{b} \exp(-(c+1)). \end{aligned}$$

This influence function is also bounded and illustrated in Figure 4.

The level of robustness can be controlled by the choice of $c \geq 1$, 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 3.4 we derived how the asymptotic variance of $\hat{\gamma}_{k,n}^{RGLM}$ depends on c and ρ . Table 1 gives an overview of which values of c should be chosen to reach a certain ARE with respect to the ML estimator $\hat{\gamma}_{k,n}^{ML}$.

Figure 4: Influence function for $\hat{\gamma}^{RGLM}$ at a Fréchet(2) distribution.

(which has a variance equal to $\frac{\gamma^2}{k} \left(\frac{1-\rho}{\rho} \right)^2$) for the most common choices for ρ . 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 bias at contaminated data. Hence, in our examples we have always set $c = 1.105$ and $\rho = -1$ to obtain an ARE of 80% in case ρ is specified correctly. In case of a misspecification of ρ , an ARE around 80% can still be expected since for other ρ -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 .

ARE	c		
	$\rho = -2$	$\rho = -1$	$\rho = -0.5$
80%	1.080	1.105	1.125
85%	1.280	1.400	1.410
90%	1.815	1.825	2.830
95%	2.555	2.560	2.565

Table 1: Values for the Huber constant c to obtain a required ARE, for several values of ρ .

3.6. 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 γ . 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 ρ calculated by the ML estimator. The resulting k -value will be denoted by k_H . Note that some of these methods rely on a value of the parameter ρ but the authors mention that a wrong specification of ρ 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 selecting the optimal k -value for $\hat{\gamma}_{k,n}^{RGLM}$, denoted by k_{RGLM} , we propose to minimize a mean weighted squared error of the Z_j , given by

$$\begin{aligned} \text{MWSE} &= \frac{1}{k} \sum_{j=1}^k w_j \left(\frac{Z_j - \widehat{E[Z_j]}}{\hat{\sigma}_j} \right)^2 \\ &= \frac{1}{k} \sum_{j=1}^k w_j r_j^2, \end{aligned}$$

with $\widehat{E[Z_j]} = \hat{\mu}_j = \exp(\mathbf{u}_j^t \hat{\beta}_k)$, $\hat{\sigma}_j^2 = \hat{\mu}_j^2$, $r_j = (z_j - \hat{\mu}_j)/\hat{\mu}_j$ as before and $w_j = \psi_c(r_j)/r_j$. This means that k_{RGLM} and $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$ are estimated simultaneously. Note that a more refined criterion like the robust prediction error criterion proposed by Dupuis and Victoria-Feser (2006) could be used, which consists of adding a correction factor for out of sample correction, but this falls outside the scope of this paper.

For comparison reasons, we would also like to have an automatic method to select an optimal k -value for the ML estimator, denoted as k_{ML} . Since for uncontaminated data the ML estimator typically has a large stable region, we believe that the median of all $\hat{\gamma}_{k,n}^{ML}$ values is a good candidate for the optimal $\hat{\gamma}_{k_{ML},n}^{ML}$. The median value is not always observed, therefore we define k_{ML} as the minimum of all k -values for which the difference with $\hat{\gamma}_{k_{ML},n}^{ML}$ is the smallest. Alternatively, one could estimate k_{ML} and γ using the classical prediction error criterion as in Dupuis and Victoria-Feser (2006). For the WMLE estimator, a robust

prediction error criterion has been used to simultaneously estimate k_{WMLE} and $\hat{\gamma}_{k_{WMLE},n}^{WMLE}$.

To examine the performance of the estimators when applying the criteria described above to select an appropriate value of k_{ML} , k_{WMLE} and k_{RGLM} , we computed $\hat{\gamma}_{k_{ML},n}^{ML}$, $\hat{\gamma}_{k_{WMLE},n}^{WMLE}$ and $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$ for the data sets and outlier configurations used in Figure 3. To ensure an ARE of around 80%, c was chosen equal to 1.105 for the robust GLM estimator, and for the WMLE we set $c_w = 1.25$ as this yields a good trade-off between robustness and efficiency (yielding an ARE around 60%). We also did the simulations with $c_w = 2$, which yields around the same ARE as the RGLM estimator with $c = 1.105$, but those results were not as good as the ones shown here.

Figure 5(a) shows boxplots for $\hat{\gamma}_{k_{ML},n}^{ML}$ (left), $\hat{\gamma}_{k_{WMLE},n}^{WMLE}$ (middle) and $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$ (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 the robust GLM estimator $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$ are comparable, the latter having a slightly larger variance. When the data come from a Burr(1,1,2) distribution without contamination, depicted in Figure 5(b), the three estimators often overestimate the true value $\gamma = 0.5$ (as was already observed in Figure 3(b)). Here, $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$ has a smaller bias than $\hat{\gamma}_{k_{ML},n}^{ML}$ and a comparable variance. The WMLE has the largest bias and the smallest variance.

In Figures 5(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) and the Burr(1,1,2) distribution. On the other hand, both robust estimators, $\hat{\gamma}_{k_{WMLE},n}^{WMLE}$ and $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$, give quite accurate values for γ in case of a contaminated Fréchet(2) distribution. The WMLE overestimates the true value by a little bit, whereas the RGLM estimator slightly underestimates γ . It can be noticed in Figure 5(c) that $\hat{\gamma}_{k_{WMLE},n}^{WMLE}$ has a slightly larger variance than $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$. In case of a Burr(1,1,2) distribution with 2% contamination, we can see from Figure 5(d) that both robust estimators show a positive bias, which is however much smaller for $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$ than for $\hat{\gamma}_{k_{WMLE},n}^{WMLE}$. Also its variance is much smaller.

We conclude that $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$, where k_{RGLM} is chosen by an MSE-method, slightly outperforms the WMLE-method, which also uses a data-based estimation procedure to estimate

Figure 5: Boxplots of 50 ML estimates (left), WMLE estimates (middle, $c_w = 1.25$) and RGLM estimates (right, $c = 1.105$) for γ for data sets with (a) 500 uncontaminated observations from a Fréchet(2) distribution and (b) Burr(1,1,2) distribution; (c) 490 observations from a Fréchet(2) and (d) Burr(1,1,2) distribution and 10 contaminating observations.

k . Moreover the estimation of k (and hence of γ) in the latter case appears to be affected by the choice of the upper and lower bound in the optimization algorithm of $\hat{\gamma}_{k_{WMLE},n}^{WMLE}$, and the estimation of k uses a considerable amount of computation time.

4. Detecting influential data points

4.1. A robust high quantile estimate

We can use $\hat{\gamma}_{k_{RGLM}}^{RGLM}$ to determine a high quantile of the data generating distribution. This high quantile can be considered as the threshold which separates the “normal” from the “unusual” data points. Based on the Hill estimator, Weissman (1978) proposed for Pareto-type distributions an estimate for a large quantile q_p by

$$\hat{q}_{k,p}^H = x_{n-k,n} \left(\frac{k+1}{(n+1)p} \right)^{\hat{\gamma}_{k,n}^H}, \quad (13)$$

where k should be chosen as the optimal k_H -value and with p the exceedance probability. Since the Hill estimator shows a severe bias for k large, Beirlant et al. (2004) proposed to replace $\hat{\gamma}_{k,n}^H$ with $\hat{\gamma}_{k,n}^{ML}$, which yields the quantile estimate $\hat{q}_{k,p}^{ML}$. In order to obtain a robust estimate for q_p , we now replace the Hill estimator in (13) by $\hat{\gamma}_{k,n}^{RGLM}$, and set $k = k_{RGLM}$, which yields $\hat{q}_{k,p}^{RGLM}$. For an appropriate choice of p , unusual data points can now be detected. In the next Section we propose a visual tool to determine if those unusual data points also have a high influence on the Hill estimator. To do so, we plug in $\hat{\gamma}_{k_{RGLM}}^{RGLM}$ in the empirical influence function (5) of the Hill estimator.

4.2. 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_{RGLM} value, i.e.

$$EIF^{RGLM}(y, \hat{\gamma}_{k_{RGLM},n}^H, \hat{F}_n) = \frac{n}{k_{RGLM}} \left(\log \frac{y}{x_{n-k_{RGLM},n}} - \hat{\gamma}_{k_{RGLM},n}^{RGLM} \right) I(y > x_{n-k_{RGLM},n}). \quad (14)$$

To measure the influence of a single observation, we can finally consider $EIF^{RGLM}(y, \hat{\gamma}_{k_{RGLM},n}^H, \hat{F}_n)$ for $y = x_{n-k_{RGLM}+1,n}, x_{n-k_{RGLM}+2,n}, \dots, x_{n,n}$. This leads to the **empirical influence func-**

tion plot (EIFP), which displays the EIF^{RGLM} in the data points versus their value:

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

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 (2005) 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 6(a), 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 largest observations.

Figure 6: (a) Pareto QQ-plot for the Hedge fund data set with regression lines based on the Hill estimator (dotted line), the ML estimator (dashed line), the WMLE (dash-dotted line) and the RGLM estimator (full line); (b) detail of (a).

We first make a plot of $\hat{\gamma}_{k,n}^{RGLM}$ as a function of k , shown in Figure 7(a). To obtain an ARE around 80%, we have set $c = 1.105$. The full line in Figure 7(a) shows our robust GLM estimator $\hat{\gamma}_{k,n}^{RGLM}$. For comparison purposes we have also added the Hill estimator (dotted line), the ML estimator (dashed line) and the WMLE estimator (dash-dotted line) with $c_w = 1.25$ yielding an ARE around 60%. The value for $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$ with $c = 1.105$ is

equal to 0.01295 ($k_{RGLM} = 52$), which is comparable to $\hat{\gamma}_{k_{WMLE},n}^{WMLE} = 0.01343$ (obtained at $k = 22$) for $c_w = 1.25$ as was shown in Dupuis and Victoria-Feser (2006). Note that the choice $c_w = 2$ would be more natural to use here since this implies an ARE around 80%, but the results in this case are much less robust. The Maximum Likelihood estimator yields $\hat{\gamma}_{k_{ML},n}^{ML} = 0.01532$ for $k_{ML} = 37$ and the Hill estimator equals $\hat{\gamma}_{k_H,n}^H = 0.01534$ at $k_H = 45$. On the Pareto QQ-plot in Figure 6 we have added regression lines whose slope is given by the Hill estimator (dotted line), the ML estimator (dashed line), the WMLE (dash-dotted line) and the RGLM estimator (full line). It can be noticed that the regression lines based on the Hill and ML estimator are drawn towards the three largest observations, whereas the regression lines in which $\hat{\gamma}_{k_{WMLE}}^{WMLE}$ and $\hat{\gamma}_{k_{RGLM}}^{RGLM}$ are used, tend to follow the bulk of the data points. Figure 6(b) shows a detail of the Pareto QQ-plot.

Figure 7: (a) Plot of γ as a function of k for the Hedge fund data for the Hill estimator (dotted line), ML estimator (dashed line), WMLE estimator (dash-dotted line) and RGLM estimator (full line); (b) EIFP for the same data with quantiles for $p = 1/n$ (dotted line), $p = 0.01$ (dashed line) and $p = 0.005$ (full line).

The corresponding EIFP for the $k_{RGLM} = 52$ largest observations is shown in Figure 7(b). We see that the three largest cases have a considerably larger EIF^{RGLM} value than the other observations. We also added quantile estimates $\hat{q}_{k_{RGLM},p}^{RGLM}$ for $p = 0.005$ (full line), $p = 0.01$ (dashed line) and $p = 1/n \approx 0.014$ (dotted line) and observe that for $p = 0.005$ only the largest data point, which also has the largest EIFP, is detected as unusual, whereas for the two other values of p all three largest data points exceed the threshold line (defined

by the quantile estimate). Note that Dupuis and Victoria-Feser (2006) show that the WMLE downweights the largest and third largest observation (which could explain a slightly higher estimate for γ), but no indication is given on how much these observations deviate from the postulated model.

We could also define EIF^H , EIF^{ML} and EIF^{WMLE} by replacing $\hat{\gamma}_{k_{RGLM},n}^{RGLM}$ in (14) with $\hat{\gamma}_{k_H,n}^H$, $\hat{\gamma}_{k_{ML},n}^{ML}$ and $\hat{\gamma}_{k_{WMLE},n}^{WMLE}$ respectively. In Figure 8(a) we can see that when the Hill estimator is plugged in into the EIF , the quantile estimates (now derived from the Hill estimator as is (13)) become larger so that for $p = 0.005$, $p = 0.01$ and $p = 1/n$ only the largest data point is detected as unusual. The EIF^{ML} shows a very comparable result in Figure 8(b), where the quantile estimates are now computed using the ML estimator for γ . When the WMLE is used, the EIF^{WMLE} shows satisfying results in Figure 8(c). For $p = 0.005$ only one data point exceeds the threshold, when $p = 0.01$ the second and third largest are very close to the quantile, and for $p = 1/n$, all three largest points can be clearly defined as unusual.

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.

Figure 8: Plot of (a) EIF^H ; (b) EIF^{ML} and (c) EIF^{WMLE} for the Hedge Fund data with corresponding quantile estimates.

It was suggested by an anonymous referee that the robust threshold values should be compared to confidence bands for QQ-plots. In order to do so we show in Figure 9 a Pareto QQ-plot for the $k_{RGLM} = 52$ 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 unusual on the EIFP, are lying well within the confidence bands, so that this method can not be used to flag unusual data points.

Figure 9: Pareto QQ-Plot for the $k_{RGLM} = 52$ largest observations in the Hedge fund data with robust regression line and 95% pointwise confidence bands.

5. Another data example: Norwegian fire insurance data from 1984

The Norwegian fire insurance data, treated by Beirlant et al. (2004), 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 1984, yielding 557 observations. The Pareto QQ-plot in Figure 10(a) shows that the data are very likely to come from a Pareto-type distribution, except for the two largest observations (indicated with a cross sign) who seem to deviate a bit from the straight line formed by the other data points.

In Figure 10(b) a plot of γ as a function of k is shown for the Hill estimator (dotted line), the ML estimator (dashed line) and the robust GLM estimator (full line). The Hill estimator shows a slight bias for larger k -values, whereas the ML and robust GLM estimator are quite stable in case k is not too small. For most k -values, the robust estimates for γ are smaller than the ML estimates. The robust MSE-criterion selects $\hat{\gamma}_{k_{RGLM},n}^{RGLM} = 0.6086$ at $k_{RGLM} = 345$. This is somewhat smaller than $\gamma_{k_{ML},n}^{ML} = 0.6812$ at $k_{ML} = 301$, and than $\hat{\gamma}_{k_H,n}^H = 0.7078$ for $k_H = 239$.

The effect of the different estimates of γ on the quantile estimates and the EIF becomes quite visible in Figure 11. Since this data set contains roughly 10 times more observations

Figure 10: (a) Pareto QQ-plot of the 1984 Norwegian fire insurance data ; (b) plot of γ as a function of k for the Norwegian fire insurance data for the Hill estimator (dotted line), ML estimator (dashed line) and RGLM estimator with $c = 1.105$ (full line).

than the Hedge fund data, we now compute quantiles for $p = 0.0005$ (full line), $p = 0.001$ (dashed line) and $p = 1/n \approx 0.0018$ (dotted line). When the Hill estimator is used to calculate the EIF and the quantile estimators, it can be seen on Figure 11(a) that, although the two largest data points seem to have a much larger influence on the Hill estimator than the other data points, only the largest observation is indicated as unusual, and only for the two smallest quantiles. The EIF based on the ML estimator (Figure 11(b)) also indicates a large influence on the Hill estimator for the two largest data points. Both are flagged as unusual when the smallest quantile is used, for the middle quantile only the largest point is flagged, and when using the largest quantile, none of them seem unusually large. When using the robust GLM estimator, the largest data point can clearly be considered as unusual when the largest quantile is used in Figure 11(c), and for the other quantiles the two largest are flagged. Moreover we notice that those two data points have a large influence on the Hill estimator as their EIF value is much larger than those of the other data points.

6. Conclusion

In this paper we have constructed a graphical tool to detect highly influential data points for the Hill estimator. We can use this in combination with a robust quantile estimate to

Figure 11: Plot of (a) EIF^H ; (b) EIF^{ML} and (c) EIF^{RGLM} for the Norwegian fire insurance data with corresponding quantile estimates.

mark data points with an unusual high value which can then be investigated thoroughly. To obtain our results, we derived the influence function of the Hill estimator of Pareto-type distributions and we developed 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 GLM estimator and hence the EIF and quantile estimates 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. We are also very grateful to professor M.-P. Victoria-Feser who kindly shared with us her R-code for computing the WMLE estimator, as well as the Hedge fund data.

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

Appendix: Influence function of the Hill estimator

First we compute the influence function of the Hill estimator at any distribution F . Next we derive an approximation at 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 $T_{\alpha}(F_{\varepsilon, \Delta_y})$ for the functional defined in (2), with α the percentage of the population that is used to estimate γ .

For $y > F^{-1}(1 - \alpha)$, with ε 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

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

and

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

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 (.1) in the following way:

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} T_\alpha(F_{\varepsilon, \Delta_y}) &= -\frac{1}{\alpha} \int_{F^{-1}(\frac{1-\alpha}{1-\varepsilon})}^{+\infty} \log x dF(x) - \frac{1-\varepsilon}{\alpha} \log F^{-1}\left(\frac{1-\alpha}{1-\varepsilon}\right) \frac{1-\alpha}{(1-\varepsilon)^2} + \frac{1}{\alpha} \log y \\ &\quad - \frac{1}{F^{-1}(\frac{1-\alpha}{1-\varepsilon})} \frac{1}{F'(F^{-1}(\frac{1-\alpha}{1-\varepsilon}))} \frac{1-\alpha}{(1-\varepsilon)^2}. \end{aligned}$$

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

$$\begin{aligned} IF(y, T_\alpha(F), F) &= -\frac{1}{\alpha} \int_{F^{-1}(1-\alpha)}^{+\infty} \log x dF(x) - \frac{1-\alpha}{\alpha} \log F^{-1}(1-\alpha) + \frac{1}{\alpha} \log y \\ &\quad - \frac{1-\alpha}{F^{-1}(1-\alpha)F'(F^{-1}(1-\alpha))} \\ &= -T_\alpha(F) + \frac{1}{\alpha} \log \frac{y}{F^{-1}(1-\alpha)} - \frac{1-\alpha}{F^{-1}(1-\alpha)F'(F^{-1}(1-\alpha))}. \quad (.2) \end{aligned}$$

For Pareto-type distributions, $\bar{F}(x) = x^{-1/\gamma} \ell(x)$. 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

$$F(x) \sim 1 - \left(\frac{x}{F^{-1}(1-\alpha)}\right)^{-1/\gamma} \alpha, \quad (.3)$$

for $\alpha \rightarrow 0$. Deriving (.3) 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 (.2):

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

Combining this result with the fact that $T_\alpha(F) \sim \gamma$ for $\alpha \rightarrow 0$ provides us with an approxi-

mation for the influence function of the Hill estimator for Pareto-type distributions:

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

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

Using the same type of calculations, it can easily be verified that for $0 < y \leq F^{-1}(1-\alpha)$, the influence function is equal to $-T_\alpha(F) + \frac{\alpha}{F^{-1}(1-\alpha)F'(F^{-1}(1-\alpha))}$, and since $T_\alpha(F) \sim \gamma$ and $\frac{\alpha}{F^{-1}(1-\alpha)F'(F^{-1}(1-\alpha))} \sim \gamma$ for Pareto-type distributions, $IF(y, T_\alpha(F), F) \sim 0$ in case of a Pareto-type distribution.

References

- Beirlant, J., Dierckx, G., Goegebeur, Y., Matthys, G. Tail index estimation and an exponential regression model. *Extremes*. 2:177–200, 1999.
- Beirlant, J., Dierckx, G., Guillou, A., Stărică, C. On exponential representations of log-spacings of extreme order statistics. *Extremes*. 5(2):157–180, 2002.
- Beirlant, J., Goegebeur, Y., Teugels, J.L., Segers, J. *Statistics of extremes: Theory and applications*. Wiley, Chichester, 2004.
- Brazauskas, V., Serfling, R.J. Robust and efficient estimation of the tail index of a single-parameter Pareto distribution. *North American Actuarial Journal*. 4:12–27, 2000.
- Cai, J.-J., Einmahl, J., de Haan, L. Estimation of extreme risk regions under multivariate regular variation. *The Annals of Statistics*. 39:1803–1826, 2011.
- Cantoni, E., Ronchetti, E. Robust inference for generalized linear models. *Journal of the American Statistical Association*. 96:1022–1030, 2001.
- Cantoni, E., Ronchetti, E. A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditure. *Journal of Health Economics*. 25:198–213, 2006.

- Debruyne, M., Hubert, M., Van Horebeek, J. Detecting influential observations in kernel PCA. *Computational Statistics and Data Analysis*. 54:3007–3019, 2010.
- Dell’Aquila, R., Embrechts, P. Extremes and robustness: a contradiction? *Financial Markets and Portfolio Management*. 20:103–118, 2006.
- Dierckx, G., Beirlant, J., De Waal, D., Guillou, A. A new estimation method for Weibull-type tails based on the mean excess function. *Journal of Statistical Planning and Inference*. 139:1905–1920, 2009.
- Drees, H., Kaufmann, E. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*. 75(2):149–172, 1998.
- DuMouchel, W.H. Estimating the stable index α in order to measure tail thickness: a critique. *The Annals of Statistics*. 11(4):1019–1031, 1983.
- Dupuis, D.J., Field, C.A. Robust estimation of extremes. *The Canadian Journal of Statistics*. 26:199–215, 1998.
- Dupuis, D.J., Morgenthaler, S. Robust weighted likelihood estimators with an application to bivariate extreme value problems. *The Canadian Journal of Statistics*. 30(1):17–36, 2002.
- Dupuis, D.J., Victoria-Feser, M.-P. A robust prediction error criterion for Pareto modelling of upper tails. *The Canadian Journal of Statistics*. 34:639–658, 2006.
- Ferreira, M., Canto e Castro, L. Modeling rare events through a p RARMAX process. *Journal of Statistical Planning and Inference*. 140:3552–3566, 2010.
- Fisher, R., Tippett, L. 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.
- Fraga Alves, M.I., Gomes, M.I., de Haan, L. A new class of semi-parametric estimators of the second order parameter. *Portugaliae Mathematica*. 60(2):193–213, 2003.
- Gnedenko, B. Sur la distribution limite du terme maximum d’une série aléatoire. *Annals of Mathematics*. 44:423–453, 1943.

- Gomes, M.I., Oliveira, O. Maximum likelihood revisited under a semi-parametric context - estimation of the tail index. *Journal of Statistical Computation and Simulation*. 73(4): 285–301, 2003.
- Gomes, M.I., Martins, M.J., Neves, M. Improving second order reduced bias extreme value index estimation. *REVSTAT Statistical Journal* 5(2):177–207, 2007.
- Guillou, A., Hall, P. A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society B*. 63:293–305, 2001.
- Hampel, F., Ronchetti, E., Rousseeuw, P., Stahel, W. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York, 1986.
- Heritier, S., Cantoni, E., Copt, S., Victoria-Feser, M.-P. *Robust methods in biostatistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2009.
- Hill, B. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*. 3:1163–1174, 1975.
- Hubert, M., Vandervieren, E. An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*. 52 (12):5186–5201, 2008.
- Juárez, S.F. and Schucany, W.R. Robust and efficient estimation for the generalized Pareto distribution. *Extremes*. 7:237–251, 2004.
- Matthys, G., Beirlant, J. Adaptive threshold selection in tail index estimation. *Extremes and Integrated Risk Management*, Ed. Paul Embrechts, Risk Books. 37-49, UBS Warburg, 2000.
- McCullagh, P., Nelder, J.A. *Generalized Linear Models*. Chapman and Hall, London, 1983.
- McElroy, T., Agnieszka J. Tail index estimation in the presence of long-memory dynamics. *Computational Statistics and Data Analysis*. 56:266–282, 2012.
- Perret-Gentil, C., Victoria-Feser, M.-P. Robust mean-variance portfolio selection. *FAME Research Paper Series rp140*, International Center for Financial Asset Management and Engineering, 2005.

- Pison, G., Rousseeuw, P.J., Filzmoser, P., Croux, C. Robust factor analysis. *Journal of Multivariate Analysis*. 84:145–172, 2003.
- Pison, G., Van Aelst, S. Diagnostic plots for robust multivariate methods. *Journal of Computational and Graphical Statistics*. 13:310–329, 2004.
- Vandewalle, B., Beirlant, J., Hubert, M. A robust estimator of the tail index based on an exponential regression model. In *Theory and Applications of Recent Robust Methods* (M. Hubert, G. Pison, A. Struyf and S. Van Aelst, eds.), *Statistics for Industry and Technology*, Birkhauser, Basel, 367-376, 2004.
- Vandewalle, B., Beirlant, J., Christmann, A., Hubert, M. A robust estimator for the tail index of Pareto-type distributions. *Computational Statistics and Data Analysis*. 51:6252–6268, 2007.
- Victoria-Feser, M.-P., Ronchetti, E. Robust methods for personal-income distribution models. *The Canadian Journal of Statistics*. 22:247–258, 1994.
- Weissman, I. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*. 73:812–812, 1978.