REVSTAT – Statistical Journal
Volume 9, Number 2, June 2011, 181–198

GENERALIZED SUM PLOTS

Authors: J. Beirlant
– Department of Mathematics, Campus Kortrijk and Leuven Statistics Research Center, Katholieke Universiteit Leuven, Belgium
Jan.Beirlant@wet.kuleuven.be

E. Boniphace
– Department of Mathematics and Leuven Statistics Research Center, Katholieke Universiteit Leuven, Belgium
Edwin.Boniphace@wis.kuleuven.be

G. Dierckx
– Department of Mathematics and Statistics, Hogeschool-Universiteit Brussel, Belgium
Goedele.Dierckx@hubrussel.be

Received: September 2010 Revised: May 2011 Accepted: May 2011

Abstract:

- Sousa and Michailidis (2004) developed the sum plot based on the Hill (1975) estimator as a diagnostic tool for selecting the optimal k when the distribution is heavy tailed. We generalize their method to any consistent estimator with any tail type (heavy, normal and light tail). We illustrate the method associated to the generalized Hill estimator and the moment estimator.

  As an attempt to reduce the bias of the generalized Hill estimator, we propose new estimators based on the regression model which are based on the estimates of the generalized Hill estimator. Here weighted least squares and weighted trimmed least squares is proposed. The bias and the mean squared error (MSE) of the estimators is studied using a simulation study. A few practical examples are proposed.

Key-Words:

- *sum plot; generalized sum plot; extreme value analysis; generalized quantile plot; weighted regression model.*

AMS Subject Classification:

- 62G32, 62J05, 62F35, 62P05, 62P12.

1. INTRODUCTION

In order to estimate a tail index using k upper order statistics, one needs to determine an appropriate value of k. There exist a variety of diagnostic plots and adaptive estimation methods that assist in threshold selection. The list of plots includes Zipf, Hill, empirical mean-excess and sum plots. Adaptive selection procedures are listed for instance in Beirlant *et al.* (2005). The aim of this paper is to generalize the graphical tool developed by Sousa and Michailidis (2004) assisting in choosing a sensible estimate or a value of k. Their sum plot is based on the assumption that the distribution is heavy tailed. We extend the approach to all estimators which use a set of extreme order statistics in the estimation of a real valued extreme value index. Here we illustrate the approach using the generalized Hill estimator introduced in Beirlant *et al.* (1996b) and the moment estimator proposed by Dekkers *et al.* (1989).

In this paper we also propose new estimators of the extreme value index based on the regression associated to the estimates of the (generalized) Hill estimator for various $k = 1, ..., K$ for some K.

The article is organized as follows. In section 2, we first specify the original sum plot in subsection 2.1. Then we generalize it using the generalized Hill estimator in subsection 2.2 and using the moment estimator in subsection 2.3. In subsection 2.4 we illustrate the method with some simulation results. The new estimators based on regression models are introduced in section 3, first for the original Hill sum plot in subsection 3.1 and then for the generalized Hill sum plot in 3.2. Finally, some simulations and practical examples are presented in subsections 3.3 and 3.4.

2. SUM PLOTS

The sum plot by Sousa and Michailidis (2004) and Henry III (2009), are examples of the following principle. Let $\hat{\gamma}_{k,n}$ (which uses k upper order statistics from the total sample of size n) be a consistent estimator of γ as $k, n \to \infty$ and $k/n \to 0$. Assume first that $\hat{\gamma}_{k,n}$ is an unbiased estimator i.e. $\mathbf{E}\hat{\gamma}_{k,n} = \gamma$. Define the random variables S_k, for $k = 1, 2, ..., n-1$, by

$$S_k := k\,\hat{\gamma}_{k,n} \tag{2.1}$$

then $\mathbf{E}S_k = k\,\gamma$. Therefore the plot (k, S_k) is approximately linear for the range of k where $\hat{\gamma}_{k,n} \approx \gamma$, i.e. $\hat{\gamma}_{k,n}$ is constant in k. The slope of the linear part of the graph (k, S_k) can then be used as an estimator of γ. Assume now that $\hat{\gamma}_{k,n}$ is a consistent estimator but biased, that is $\mathbf{E}\hat{\gamma}_{k,n} = \gamma + (\text{bias})$, then $\mathbf{E}S_k =$

$k\,\gamma + k(\text{bias})$. If the bias is constant in k then (k, S_k) is again linear with the slope equal to $\gamma + (\text{bias})$. Typically though the bias is not constant in k and hence the path of (k, S_k) will depend on the non constant function in k defining the bias.

The sum plot introduced in Sousa and Michailidis (2004) is based on the Hill estimator (Hill, 1975). The sum plot by Henry III (2009) is based on a harmonic moment estimator. Both proposals were limited to the family of Pareto-type distributions.

So given $\hat{\gamma}_{k,n}$, any consistent estimator of γ based on k top order statistics, we propose a sum plot (k, S_k) based on $\hat{\gamma}_{k,n}$ with S_k defined in (2.1). The only strong assumption on $\hat{\gamma}_{k,n}$ is consistency which is a natural requirement on any estimator. This plot could be helpful in identifying an appropriate region of k, the number of order statistics to be used in $\hat{\gamma}_{k,n}$. One could argue that the plots (k, S_k) and $(k, \hat{\gamma}_{k,n})$ are statistically equivalent. The sum plot naturally leads to the estimation of the slope whereas $(k, \hat{\gamma}_{k,n})$ leads to horizontal plots and hence estimation of the intercept. Here we consider the case of a real-valued γ and hence increasing or decreasing sum plots allow to assess the sign of γ.

Since each estimator will have its own sum plot, we hereafter name the associated sum plot along the name of the estimator. For example the sum plot based on the Hill estimator is named the Hill sum plot.

In the following subsections we illustrate the proposed sum plot principle using the Hill, the generalized Hill and the moment estimator. We also illustrate the performance of these sum plots on simulated data and on some real data sets.

2.1. The Hill sum plot

Let $X_{1,n} < X_{2,n} < \cdots < X_{n,n}$ denote the order statistics of a random sample $(X_1, X_2, ..., X_n)$ from a heavy tailed distribution F with

$$1 - F(x) = x^{-1/\gamma}\, l_F(x)\ , \qquad x > 0\ , \tag{2.2}$$

where l_F is a slowly varying function at infinity satisfying

$$l_F(\lambda x)/l_F(x) \to 1 \quad \text{when} \quad x \to \infty\,, \qquad \text{for all} \quad \lambda > 0\ .$$

Let the random variables $S^H_{k,n}$ $(k = 1, ..., n)$ be defined as

$$S^H_{k,n} = \sum_{j=1}^{k} Z_j := \sum_{j=1}^{k} j \log \frac{X_{n-j+1,n}}{X_{n-j,n}}\ . \tag{2.3}$$

Sousa and Michailidis (2004) introduced the diagnostic plot $(k, S^H_{k,n})$, the sum plot for estimating the tail index γ. This plot is called the Hill sum plot since

the Hill (1975) estimator $H_{k,n}$ satisfies

$$H_{k,n} \,=\, \frac{1}{k}\sum_{j=1}^{k} \log X_{n-j+1,n} - \log X_{n-k,n} \,=\, \frac{1}{k}\, S_{k,n}^{H} \;. \tag{2.4}$$

To understand the behavior of the Hill sum plot we rely on a representation of the variables Z_j from (2.3) ($j = 1, ..., n$) provided in Beirlant *et al.* (2001). We remind that the model (2.2) is well-known to be equivalent to

$$U(x) \,=\, x^{\gamma}\, l_U(x) \;, \tag{2.5}$$

where $U(x) = \inf\{y\colon F(y) \geq 1 - 1/x\}$ $(x > 1)$ and with l_U again a slowly varying function. Often, the following second order condition on l_U is assumed

$$\frac{l_U(tx)}{l_U(x)} \,=\, 1 + b(x)\,\frac{t^{\rho}-1}{\rho}\,\big(1+o(1)\big) \;,$$

where b is a rate function satisfying $b(x) \to 0$ as $x \to \infty$ and $\rho < 0$. Under this second order condition, Beirlant *et al.* (2001) have shown that

$$\left| Z_j \,-\, \left(\gamma + b_{n,k}\Big(\frac{j}{k+1}\Big)^{-\rho}\right) E_j \,+\, \beta_j \right| \,=\, o_P(b_{n,k}) \;, \tag{2.6}$$

uniformly in $j \in \{1, ..., k\}$, as $k, n \to \infty$ with $k/n \to 0$, where $(E_1, ..., E_k)$ is a vector of independent and standard exponentially distributed random variables, $b_{n,k} := b\big((n+1)/(k+1)\big)$, $2 \leq k \leq n-1$ and $\frac{1}{k}\sum_{j=1}^{k} \beta_j = o_P(b_{n,k})$.

Hence, for $S_k^H = \sum_{j=1}^{k} Z_j$,

$$\left| S_k^H \,-\, \left(k\,\gamma \,+\, k\, b_{n,k}/(1-\rho) \,+\, \gamma \sum_{j=1}^{k} (E_j - 1) \,+\, o_P(k\, b_{n,k}) \right) \right| \,=\, o_P(k\, b_{n,k})$$

since $\frac{1}{k}\sum_{j=1}^{k} \big(\frac{j}{k+1}\big)^{-\rho} \sim \frac{1}{1-\rho}$ as $k, n \to \infty$ with $k/n \to 0$, where as usual, $a_n \sim b_n$ is equivalent to $a_n/b_n \to 1$ as $n \to \infty$.

In the specific case where $b(x) = C x^{\rho}\big(1 + o(1)\big)$ for some real constant C (Hall, 1982), then we obtain

$$\left| S_k^H \,-\, \left(k\,\gamma + C\, n^{\rho} k^{1-\rho} + \gamma \sum_{j=1}^{k} (E_j - 1) \right) + o_P(k\, b_{n,k}) \right| \,=\, o_P(k\, b_{n,k}) \;. \tag{2.7}$$

Sousa and Michailidis (2004) only considered the case $C = 0$.

The Hill sum plot is a graphical tool in which one is searching for a range of k where the sum plot is linear, or equivalently where $H_{k,n}$ is constant in k, if such a behaviour becomes apparent. Whereas the Hill estimator can be seen as

an estimator of the slope in a Pareto quantile plot (see for instance Beirlant *et al.* (1996a) and Kratz and Resnick (1996)), the sum plot now can be viewed as a regression plot from which new estimators can be constructed by regression of $S^H_{k,n}$ on k, as suggested by (2.7). As the regression error will turn out smaller on the sums of noise variables $\gamma(E_j - 1)$ rather than on extreme log-data, regression on the sum plot appears to be an interesting alternative approach. In practice we put $\rho = -1$ so that in that case we will fit a quadratic regression model as discussed in Section 3. The second order parameter could be replaced by estimators such as discussed in Fraga Alves *et al.* (2003). In the simulation study the case of the Burr distribution with $\rho = -0.5$ gives an idea of the loss of accuracy by setting $\rho = -1$.

2.2. The generalized Hill sum plot

Using a similar approach we derive the generalized sum plot for $\gamma \in \mathbb{R}$ based on the generalized Hill estimator by Beirlant *et al.* (1996b). Here the underlying model is that the distribution belongs to a maximum domain of attraction: there exist sequences of constants $(a_n;\, a_n > 0)$ and (b_n) such that

$$\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \le x\right) = \exp\bigl(-(1+\gamma x)^{-1/\gamma}\bigr)\,, \qquad 1+\gamma x > 0\,.$$

Define the function UH as follows

$$UH := U(x)\,\mathbf{E}\Bigl(\log X - \log U(x) \;\Big|\; X > U(x)\Bigr)\,. \tag{2.8}$$

This function possesses the regular variation property for the full range of γ. The empirical counterpart of UH at $x = n/k$ is given by

$$UH_{k,n} := X_{n-k,n}\left(\frac{1}{k}\sum_{i=1}^{k} \log X_{n-i+1,n} - \log X_{n-k,n}\right) = X_{n-k,n}\, H_{k,n}\,. \tag{2.9}$$

Using the property of regular variation of UH, Beirlant *et al.* (1996b) proposed an estimator of $\gamma \in \mathbb{R}$ by fitting a constrained least-squares line to the points with coordinates $(-\log(j/n), \log UH_{j,n})$ $(j = 1, ..., k)$ to obtain the generalized Hill estimator $H^*_{k,n}$. Similarly as in (2.4), $H^*_{k,n}$ is given by

$$H^*_{k,n} = \frac{1}{k}\sum_{i=1}^{k}\left((i+1)\,\log\frac{UH_{i,n}}{UH_{i+1,n}} + \frac{i+1}{i} - (i+1)\log\frac{i+1}{i}\right). \tag{2.10}$$

Define random variables S^{UH}_k, for $k = 1, ..., n-2$, as

$$S^{UH}_k := \sum_{i=1}^{k}(i+1)\left(\log\left(\frac{X_{n-i,n}}{X_{n-i-1,n}}\,\frac{H_{i,n}}{H_{i+1,n}}\right) + \frac{1}{i} + \log\frac{i+1}{i}\right). \tag{2.11}$$

Since $S_k^{UH} = k\,H_{k,n}^*$, we obtain the generalized Hill sum plot (k, S_k^{UH}), and we expect that for the range of k where $H_{k,n}^*$ is constant (or stable) the plot will be linear. Note that the range of k where the Hill estimator is constant, the term $H_{j,n}/H_{j+1,n} \to 1$, and (2.11) is almost reduced to S_k^H, except that the term including the largest observation is deleted.

Under general second order regular variation conditions, in Dierckx (2000) it is shown that for $1 \le j \le k$, $2 \le k \le n-2$, it holds for

$$Z_j^* := (j+1)\left(\left(\log UH_{j,n} - \log UH_{j+1,n}\right) + \frac{1}{j} + \log\frac{j+1}{j}\right)$$

that

$$\left| Z_j^* - \left(\left(\gamma + \tilde{b}_{n,k}\left(\frac{j+1}{k+1}\right)^{-\rho}\right)E_{j+1} + \gamma\,(E_{j+1}-1) + (j+1)\left(\log\frac{\bar{E}_j}{\bar{E}_{j+1}} - \log\frac{j+1}{j} + \frac{1}{j}\right)\right) + \tilde{\beta}_j \right| = o_P\big(\tilde{b}_{n,k}\big) \tag{2.12}$$

as $k, n \to \infty$ with $k/n \to 0$, where $(E_1, ..., E_k)$ is a vector of independent and standard exponentially distributed random variables, $\bar{E}_j$ denotes the sample mean of $(E_1, ..., E_j)$, $\tilde{b}_{n,k}$ is some generic notation for a function decreasing to zero, $\rho < 0$ and $\frac{1}{k}\sum_{j=1}^{k}\tilde{\beta}_j = o_P(b_{n,k})$. Note also that for $\gamma < 0$, the above expression only holds for $j \to \infty$.

Let us denote $e_j := \gamma\,(E_{j+1}-1) + (j+1)\left(\log\frac{\bar{E}_j}{\bar{E}_{j+1}} - \log\frac{j+1}{j} + \frac{1}{j}\right)$. In Dierckx (2000) it is shown that

$$\begin{aligned} &E\,e_i = 0\ , \\ &\mathrm{Cov}(e_i, e_j) = \frac{\gamma}{j}\ , \qquad i < j\ , \\ &\mathrm{Var}(e_i) = (\gamma-1)^2 + \frac{1+2\,i}{i^2}\ . \end{aligned} \tag{2.13}$$

Model (2.12) is a direct generalization of the regression model (6) used in the Hill sum plot, leading to a generalized Hill sum plot regression approach. In practice we fit the regression model

$$S_k^{UH} = k\,\gamma + C_\rho k^{1-\rho} + \sum_{j=1}^{k} e_j\ . \tag{2.14}$$

In the simulations below we will replace ρ by the canonical choice -1 so that later on we fit a quadratic regression model to the responses S_k^{UH}, $k = 1, ..., K$ for some $K > 0$.

2.3. The moment sum plot

Let $H_{k,n}^{(2)}$ be defined as follows

$$H_{k,n}^{(2)} := \frac{1}{k}\sum_{i=1}^{k}\Big(\log X_{n-i+1,n} - \log X_{n-k,n}\Big)^2.$$

The moment estimator $M_{k,n}$ (Dekkers *et al.* (1989)) is given by

$$M_{k,n} := H_{k,n} + 1 - \frac{1}{2}\left(1 - \frac{H_{k,n}^2}{H_{k,n}^{(2)}}\right)^{-1} \tag{2.15}$$

where $H_{k,n}$ is the Hill estimator from (2.4).

Let the random variables S_k^M, for $k = 1, ..., n-1$, be defined as

$$S_k^M := \left(\sum_{i=1}^{k}\log X_{n-i+1,n} - \log X_{n-k,n}\right) + k - \frac{k}{2}\left(1 - \frac{H_{k,n}^2}{H_{k,n}^{(2)}}\right)^{-1}. \tag{2.16}$$

Definition (2.16) is equivalent to writing $S_k^M = k\,M_{k,n}$, hence (k, S_k^M) is the moment sum plot. However here a regression model has not been established to the best of our knowledge.

2.4. Simulation results

The different sum plots have been applied to some simulated data sets. Six distributions are considered:

- The strict Pareto distribution given by $F(x) = 1 - x^{-1/\gamma}$, $x > 1$, $\gamma > 0$. We have chosen $\gamma = 1$. Here $b(x) = 0$.
- The standard Fréchet distribution given by $F(x) = \exp(-x^{-1/\gamma})$, $x > 0$, $\gamma > 0$. We have chosen $\gamma = 1$. Here $\rho = -1$.
- The Burr distribution $F(x) = 1 - \Big(\frac{\eta}{\eta + x^{-\tau}}\Big)^{\lambda}$, $x > 0$, $\eta, \tau, \lambda > 0$. We have chosen $\eta = 1$, $\tau = 0.5$, $\lambda = 2$, such that $\gamma = 1$. Here $\rho = -1/\lambda = -0.5$.
- The gamma distribution $F(x) = \frac{1}{b^a\,\Gamma(a)}\int_0^x t^{a-1}\exp(-t/b)\,dt$, $x > 0$, $a, b > 0$. Here we have chosen $a = 2$, $b = 1$. Always, $\gamma = 0$.
- The uniform distribution $F(x) = x$ $(0 < x < 1)$. Here $\gamma = -1$.
- The reversed Burr distribution $F(x) = 1 - \Big(\frac{\beta}{\beta + (x_+ - x)^{-\tau}}\Big)^{\lambda}$, $x > 0$, $\eta, \tau, \lambda > 0$, x_+ denotes the right endpoint of the distribution. We have chosen $\eta = 1$, $\tau = 0.5$, $\lambda = 2$, $x_+ = 2$, such that $\gamma = -1$. Here $\rho = -1/\lambda = -0.5$.

The Hill sum plots, the generalized Hill sum plots and the moment sum plots of these distributions are shown in Figure 1.

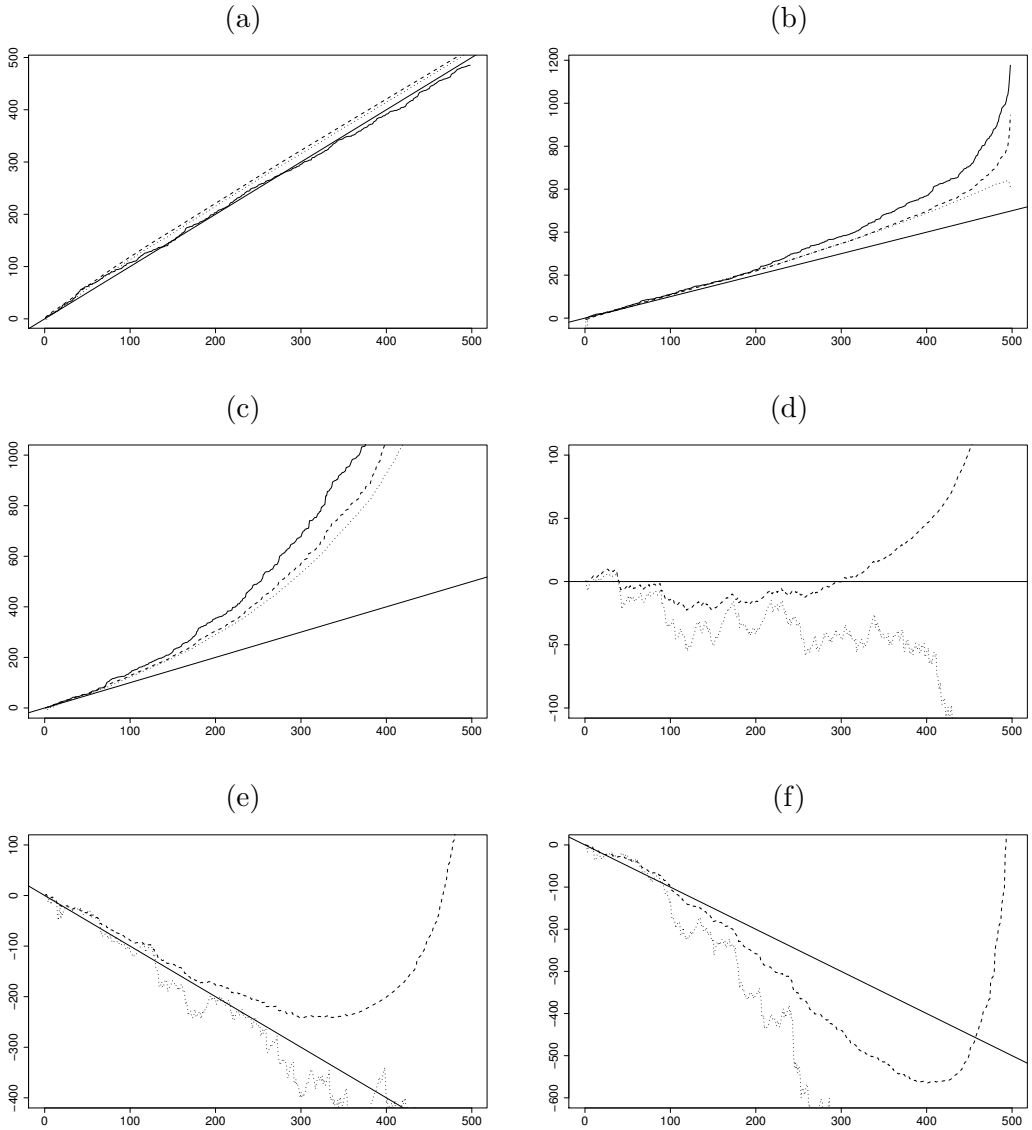

Figure 1: The Hill sum plot (full line); generalized Hill sum plot (dashed line) and the moment sum plot (dotted line) are plotted for simulated data sets of size n=500 from the: (a) strict Pareto; (b) Fréchet; (c) Burr; (d) gamma; (e) uniform; (f) reversed Burr distribution.

For $\gamma > 0$, the three sum plots are comparable for the linear parts of the plots. For $\gamma = 0$ and $\gamma < 0$, the generalized Hill sum plot and the moment sum plot are comparable on these particular data sets. However the generalized Hill sum plot seems to be less volatile. The sum plots can be used to identify the sign of γ

for a given data set: increase in k indicates $\gamma > 0$, a horizontal pattern indicates $\gamma = 0$ and decrease in k indicates $\gamma < 0$. In future work this could be used to test the domain of attraction condition. For an overview of this problem we can refer to Neves and Fraga Alves (2008). Moreover, the linear part of these generalized sum plots can be used to estimate the value of tail index.

3. REGRESSION ESTIMATORS

As indicated before, we propose regression estimators for the extreme value index γ based on the slope of the Hill and the Generalized Hill sum plots.

3.1. Hill sum plot estimators

Huisman *et al.* (2001) introduced a new estimator for $\gamma > 0$ based on the Hill sum plot which can be understood from (6). It indeed follows from (6) that for some constant D

$$\left| H_{k,n} - \left(\gamma + Dk^{-\rho} + \epsilon_k\right) \right| = o_P(b_{n,k}) \tag{3.1}$$

where $\epsilon_k = \gamma/k \sum_{j=1}^{k}(E_j - 1)$, leading to the regression model

$$H_{k,n} = \gamma + Dk^{-\rho} + \epsilon_k \ , \qquad k = 1, ..., K \ . \tag{3.2}$$

Since the variance of the error term $\mathrm{Var}(\epsilon_k) = \gamma^2/k$ is not constant, a weighted least squares regression is applied with a $K \times K$ diagonal weight matrix $W = \mathrm{diag}\sqrt{1}, ..., \sqrt{K}$. Note that in this way, Huisman *et al.* (2001), did not take into account that the error terms are not independent.

In practice, we put $\rho = -1$. Huisman *et al.* (2001) assumed that $\rho = -1/\gamma$ which is the case for an extreme value distribution.

Remark that when deleting the second order term $Dk^{-\rho}$ in the regression model, one obtains a simple average of K Hill estimators. Due to the volatile behaviour of Hill estimators $H_{k,n}$ as a function of k it is known that a robust average of Hill estimators provides better estimators. This will be discussed in more detail in case of the generalized Hill sum plot where we apply weighted trimmed least squares regression.

3.2. Generalized Hill sum plot estimators

In a similar way, a new estimator can be introduced for real valued γ based on the generalized Hill sum plot. Indeed from (2.14)

$$H^*_{k,n} \;=\; \gamma + Dk^{-\rho} + \tilde{\epsilon}_k \;, \qquad k = 1, ..., K \;. \tag{3.3}$$

with $\tilde{\epsilon}_k = \sum_{j=1}^{k} e_j/k$. The variance of $\tilde{\epsilon_k}$ is asymptotically equal to the asymptotical variance $\mathrm{AVar}(H^*_{k,n})$ which, according to Beirlant *et al.* (2005) is equal to

$$\begin{aligned} \mathrm{AVar}(H^*_{k,n}) \;&=\; \frac{1+\gamma^2}{k}; \quad \gamma \geqslant 0 \\ &=\; \frac{(1-\gamma)\,(1+\gamma+2\,\gamma^2)}{(1-2\,\gamma)\,k}; \quad \gamma < 0 \;. \end{aligned}$$

Since the variance of the error term $\mathrm{Var}(\epsilon_k) = C_\gamma/k$ is not constant, a weighted least square regression is applied with the same $K \times K$ weight matrix W as in case of the Hill sum plot. Here again we ignore the fact that the error terms are not independent. We also put $\rho = -1$.

We also apply weighted trimmed least squares regression minimizing the sum of the $\lfloor n/2 \rfloor + 1$ smallest squared residuals. For more information we refer to Rousseeuw and Leroy (1987).

3.3. Simulation results

In Figures 2 till 5 we show the simulation results we obtained concerning weighted least squares regression estimators, trimmed and non trimmed, for some of the distributions considered in Section 2.4. For each distribution 100 repetitions of samples of size $n = 500$ were performed.

Weighted trimmed least squares yields less bias but somewhat higher mean squared error compared with the non robust regression algorithm. In case $\gamma > 0$ we also show the results for the weighted least squares estimators based on the Hill sum plot. Hill sum plots then yield better results than the generalized Hill sum plot. Also the trimmed regression algorithm is typically better than the non-robust version in case of the generalized Hill sum plot.

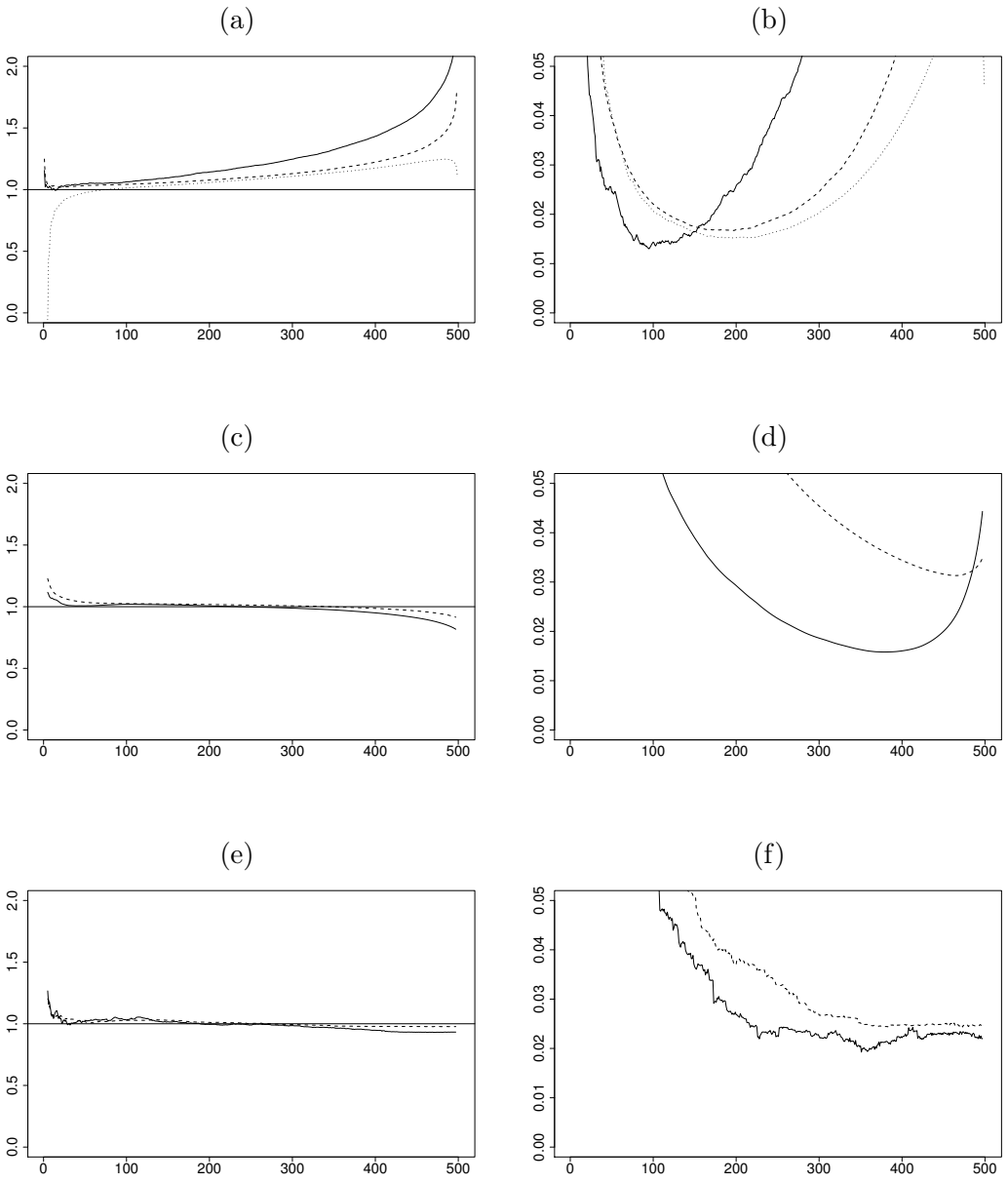

Figure 2: Fréchet distribution: (a) means of $H_{k,n}$ (full line), $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted line) as a function of k; (b) MSE of the estimators in (a); (c) means of weighted least squares estimators based on the regression model of the Hill sum plot (full line) and generalized Hill sum plot (dashed line); (d) MSE of the estimators in (c); (e) same as in (c), but now weighted trimmed least squares is used; (f) MSE of the estimators in (e).

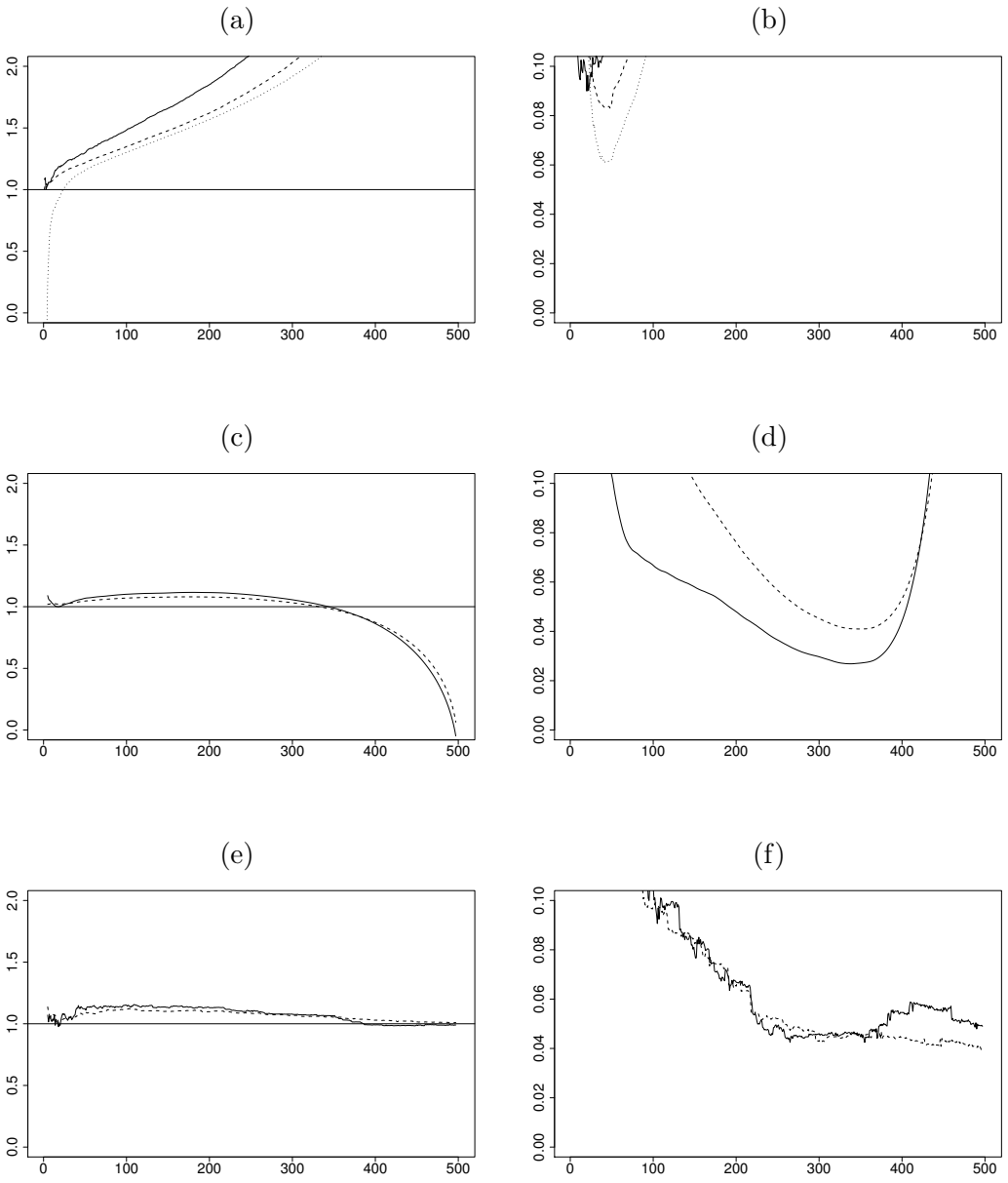

Figure 3: Burr distribution with $\rho = -0.5$: (a) means of $H_{k,n}$ (full line), $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted line) as a function of k. (b) MSE of the estimators in (a). (c) means of weighted least squares estimators based on the regression model of the Hill sum plot (full line) and generalized Hill sum plot (dashed line); (d) MSE of the estimators in (c); (e) same as in (c), but now weighted trimmed least squares is used; (f) MSE of the estimators in (e).

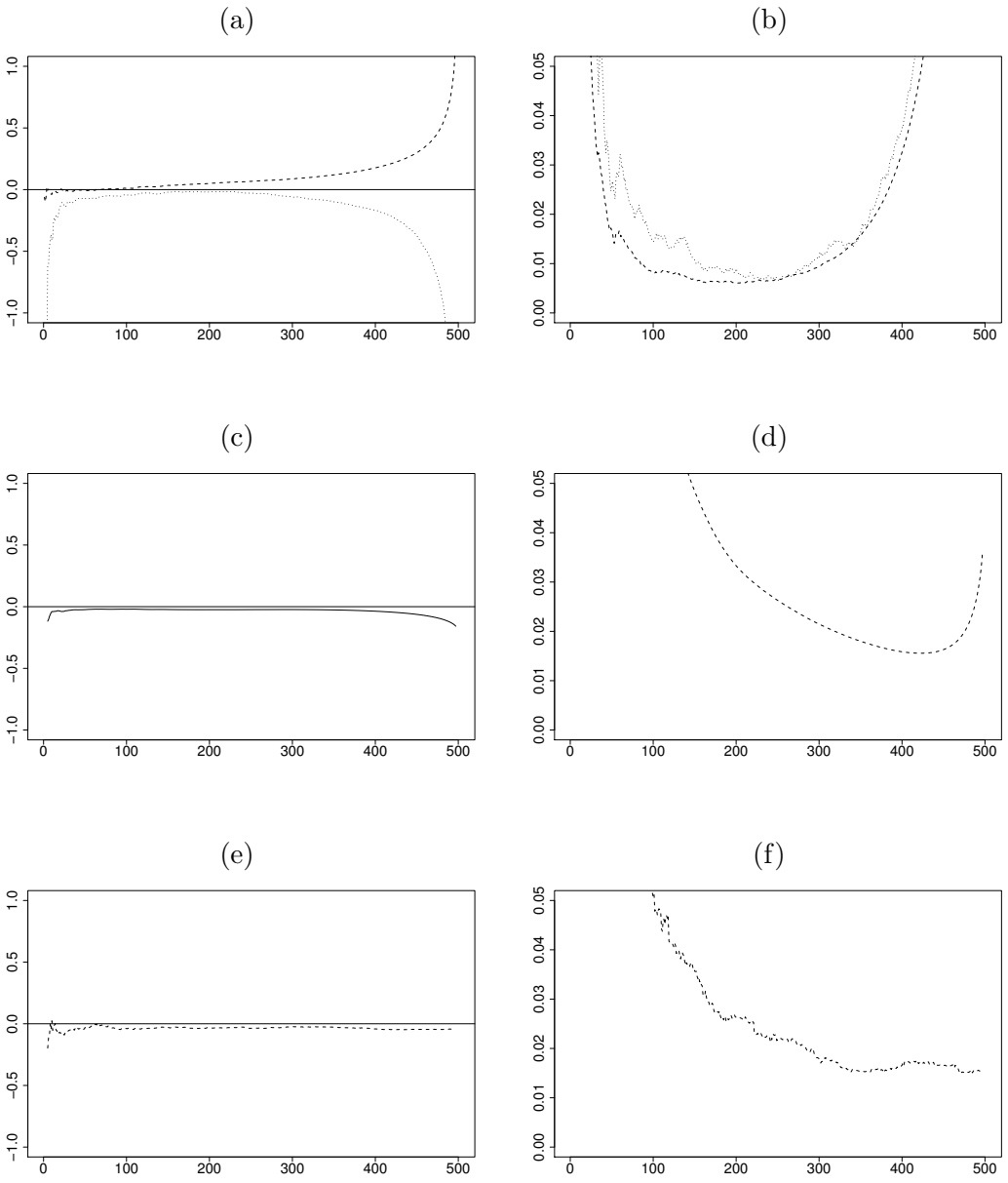

Figure 4: Gamma distribution: (a) means of $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted line) as a function of k; (b) MSE of the estimators in (a); (c) means of weighted least squares estimators based on generalized Hill sum plot (dashed line); (d) MSE of the estimators in (c); (e) same as in (c), but now weighted trimmed least squares is used; (f) MSE of the estimators in (e).

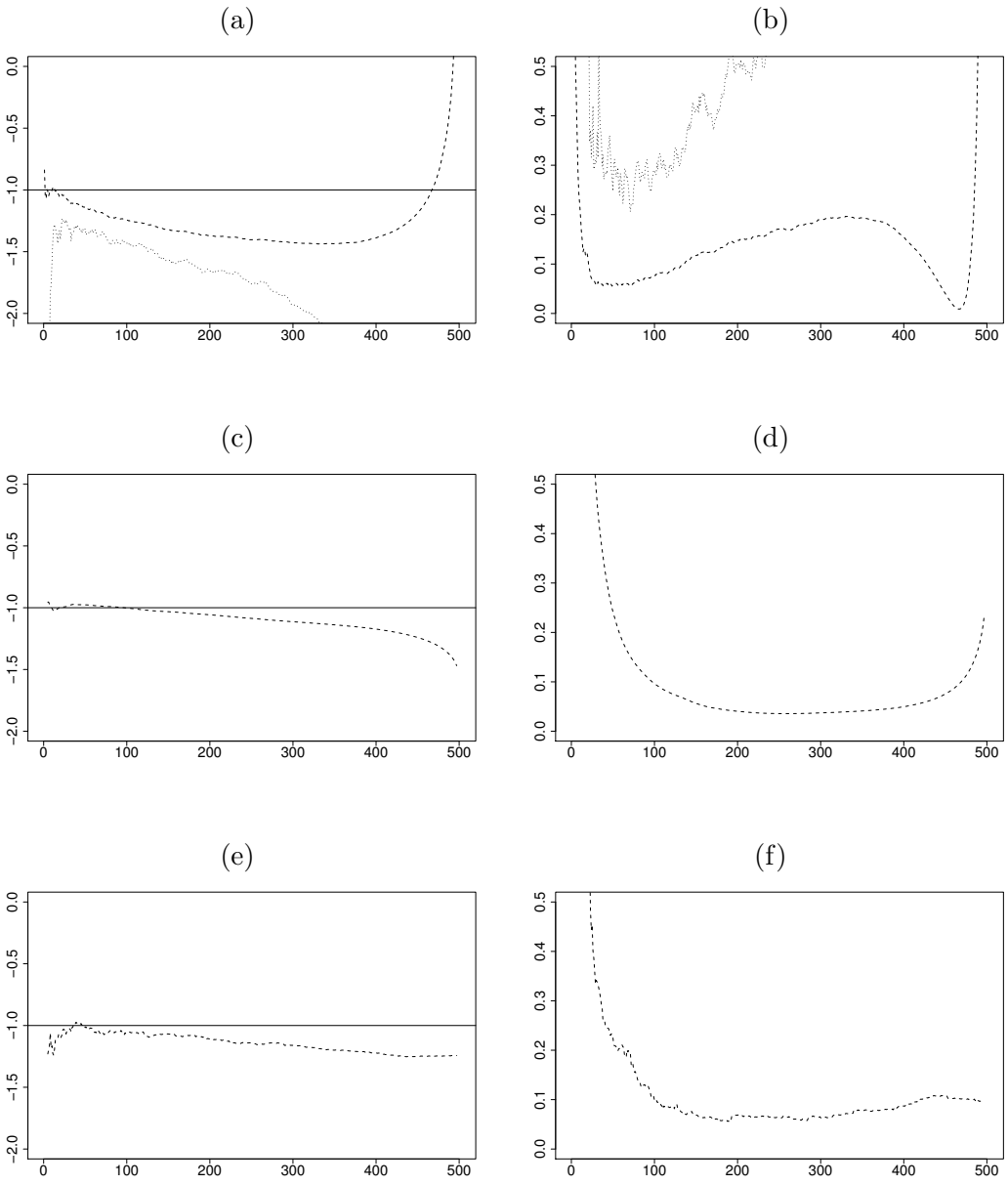

Figure 5: Reversed Burr: (a) means of $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted line) as a function of k; (b) MSE of the estimators in (a); (c) weighted least squares estimators based on the generalized Hill sum plot (dashed line); (d) MSE of the estimators in (c); (e) same as in (c), but now weighted trimmed least squares is used; (f) MSE of the estimators in (e).

3.4. Some practical examples

We end this paper showing the proposed methods into action. We apply the methods to two data sets proposed earlier in Beirlant *et al.* (2004). The data sets themselves can be found on `http://lstat.kuleuven.be/Wiley`.

The first data set contains daily maximal wind speeds at Brussels airport (Zaventem) from 1985 till 1992.

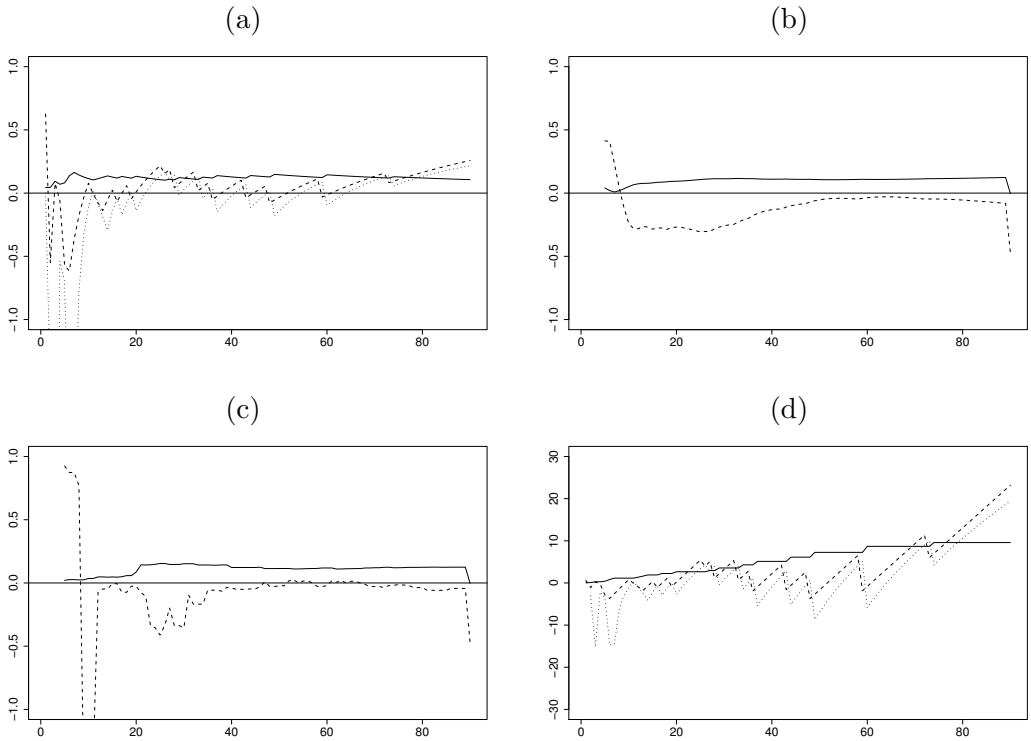

Figure 6: Zaventem daily maximum wind speed data: (a) $H_{k,n}$ (full line), $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted) as a function of k; (b) weighted least squares estimators based on the regression model of the Hill sum plot (full line) and generalized Hill sum plot (dashed line); (c) same as in (b), but now weighted trimmed least squares is used; (d) sum plots.

In Example 1.1 in Beirlant *et al.* (2004) the authors come to the conclusion that the data follow a simple exponential tail beyond 80 km/hr, and hence γ equals 0. The weighted (trimmed) least squares estimates based on the generalized Hill sum plot indeed indicates a zero valued extreme value index. Also the moment and generalized Hill sum plots indeed exhibit an overall horizontal behaviour for $K = 1, ..., 80$. The generalized Hill sum plot is less volatile however.

Finally we consider the AoN Re Belgium fire portfolio data introduced in section 1.3.3 in Beirlant *et al.* (2004). Here we omit the covariate information concerning sum insured and type of building. Here the estimate $\hat{\gamma} = 1$ follows from the weighted trimmed least squares regression analysis. These estimates are indeed quite stable over K-values compared to the non-robust version. The sum plots in Figure 7(d) are quite comparable for $K = 1, ..., 80$.

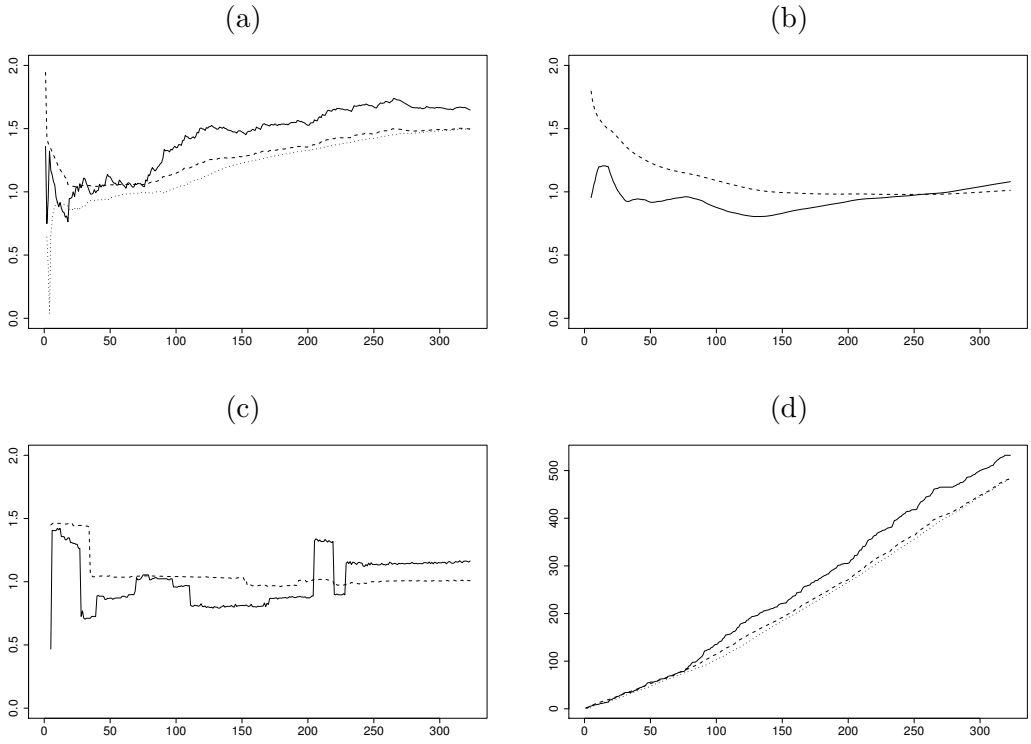

Figure 7: AoN claim size data: (a) $H_{k,n}$ (full line), $H^*_{k,n}$ (dashed line) and $M_{k,n}$ (dotted line) as a function of k; (b) weighted least squares estimators based on the regression model of the Hill sum plot (full line) and generalized Hill sum plot (dashed line); (c) same as in (b), but now weighted trimmed least squares is used; (d) sum plots.

ACKNOWLEDGMENTS

This research is sponsored by FWO grant G.0436.08N.

REFERENCES

[1] Beirlant, J.; Teugels, J.L. and Vynckier, P. (1996a). *Practical Analysis of Extreme Values*, Leuven University Press, Leuven.

[2] Beirlant, J.; Vynckier, P. and Teugels, J.L. (1996b). Excess functions and estimation of the extreme value index, *Bernoulli*, **2**, 293–318.

[3] Beirlant, J.; Dierckx, G.; Guillou, A. and Starica, C. (2001). On exponential Representations of Log-Spacings of Extreme Order Statistics, *Extremes*, **5**, 157–180.

[4] Beirlant, J.; Goegebeur, Y.; Segers, J. and Teugels, J.L. (2004). *Statistics of Extremes*, Wiley.

[5] Beirlant, J.; Dierckx, G. and Guillou, A. (2005). Estimation of the extreme-value index and generalized quantile plots, *Bernoulli*, **5**(6), 949–970.

[6] Dekkers, A.L.M.; Einmahl, J.H.J. and de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution, *Ann. Statist.*, **17**, 1833–1855.

[7] Dierckx, G. (2000). *Estimation of the Extreme Value Index*, Doctoral thesis, Katholieke Universiteit Leuven.

[8] Fraga Alves, M.I.; Gomes, M.I. and de Haan, L. (2003). A new class of semi-parametric estimators of the second order parameter, *Portugaliae Mathematica*, **60**, 193–213.

[9] Hall, P. (1982). On some simple estimates of an exponent of regular variation, *Journal of the Royal Statistical Society B*, **44**, 37–42.

[10] Huisman, R.; Koedijk, K.; Kool, C. and Palm, F. (2001). Tail-index estimates in small samples, *Journal of Business and Economic Statistics*, **19**(2), 208–216.

[11] Henry III, J.B. (2009). A harmonic moment tail index estimator, *Journal of Statistical Theory and Applications*, **8**(2), 141–162.

[12] Hill, B. (1975). A simple general approach to inference about the tail of a distribution, *Ann. Statist.*, **3**, 1163–1174.

[13] Kratz, M. and Resnick, S. (1996). The qq-estimator of the index of regular variation, *Communications in Statistics: Stochastic Models*, **12**, 699–724.

[14] Neves, C. and Fraga Alves, M.I. (2008). Testing extreme value conditions: an overview and recent approaches, *REVSTAT – Statistical Journal*, **6**, 83–100.

[15] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, Wiley.

[16] Sousa, B. de and Michailidis, G. (2004). A diagnostic plot for estimating the tail index of a distribution, *J. Comput. Graph. Statist.*, **13**(4), 974–1001.