

A Robust Estimator of the Tail Index based on an Exponential Regression Model

B. Vandewalle, J. Beirlant, and M. Hubert

Abstract. The objectives of a robust statistical analysis and of an extreme value analysis apparently are contradictory. Where the extreme data are downweighted in robust statistics, these observations receive most attention in an extreme value approach. The most prominent extreme value methods however are constructed on maximum likelihood estimates based on specific parametric models which are fitted to exceedances over large thresholds. So within an extreme value framework some robust algorithms replacing the maximum likelihood part of this methodology can be of use leading to estimates which are less sensitive to few particular observations. This study is motivated by a soil database quality management project, where in the background of Pareto-type tails, automatic identification of suspicious data is needed.

1. Introduction

In agriculture, a new concept of crop management has emerged, permitting within-field variation of crop techniques as, for instance, the adjustment of fertilizer inputs on the basis of soil sampling and analysis. The development of these techniques has greatly increased the demand for soil data and Laboratories are burdened with large data sets, which inevitably cause concern about outliers and quality of the information. Therefore, automatic outlier detection methods have become a necessity in the database management in order to provide high quality data to Laboratories. The present paper studies Ca and pH records from the Condroz region in Belgium (1505 observations, Goegebeur *et al.* (2002)). The Ca distribution, conditional on pH-level, appears to be right-skewed and long-tailed, resulting in rather frequent large Ca measurements, as can be seen in the scatterplot of Ca versus pH (given in Figure 1). Robust statistical procedures which assume that the regular data points are sampled from a normal distribution will flag too many large observations as outliers. Such long tailed data can be analysed more efficiently in the framework of extreme value theory. We will present a method which, in the

Received by the editors May 1, 2003.

1991 *Mathematics Subject Classification.* Primary 62G35; Secondary 62G32.

Key words and phrases. Extreme value statistics, Robustness.

context of a model with heavy tails, allows to point out potential outliers which need to be investigated before further analysis can be done.

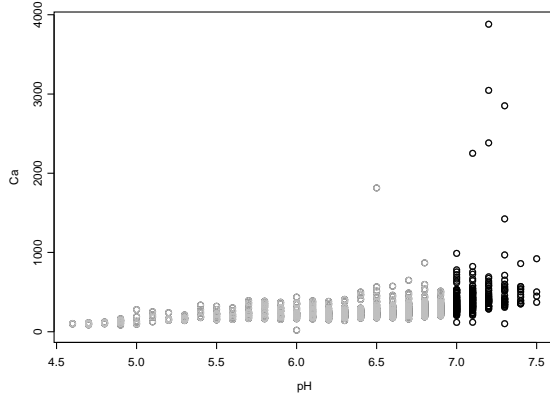


FIGURE 1. Scatterplot of Ca versus pH for one of the communities of the Condroz region in Belgium.

2. Extreme value statistics

In extreme value statistics the extreme value index (or tail index), denoted by γ , is used to characterize the tail behaviour of a distribution. This real-valued parameter helps to indicate the size and frequency of certain extreme events under a given probability distribution: the larger γ , the heavier the tail.

Consider X_1, \dots, X_n independent and identically distributed (i.i.d.) random variables with common cumulative distribution function F and quantile function Q . Denote the corresponding order statistics by $X_{1,n} \leq \dots \leq X_{n,n}$ and suppose that the properly centred and normed sample maxima $X_{n,n} = \max\{X_1, \dots, X_n\}$ converge in distribution, for $n \rightarrow \infty$, to a non-degenerate limit. This limit distribution is necessarily of extreme value type. Indeed, sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$ can then be found such that

$$(2.1) \quad \lim_{n \rightarrow \infty} P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) = H_\gamma(x),$$

with

$$(2.2) \quad H_\gamma(x) = \begin{cases} \exp\left(-(1 + \gamma x)^{-\frac{1}{\gamma}}\right), & 1 + \gamma x > 0, \gamma \neq 0, \\ \exp\left(-\exp(-x)\right), & x \in \mathbb{R}, \gamma = 0. \end{cases}$$

If (2.1) is satisfied, then F is said to belong to the maximum domain of attraction of H_γ , denoted as $F \in \mathcal{D}(H_\gamma)$.

Distributions F for which $\gamma > 0$ are called Pareto-type (or heavy-tailed) distributions, i.e. $\bar{F}(x) = x^{-1/\gamma}l_F(x)$ with l_F a slowly varying function. The Gumbel class of distributions with $\gamma = 0$ is a quite extensive class of distributions, mainly with exponential decreasing tails. The Weibull class, with $\gamma < 0$ consists of distributions with a finite right endpoint x_+ for which $\bar{F}(x_+ - 1/x) = x^{1/\gamma}l_F(x)$ with l_F again a slowly varying function. A general reference on extreme value statistics is Embrechts *et al.* (1998).

We will concentrate on Pareto type distributions, i.e. distributions for which there exists a $\gamma > 0$ such that $\bar{F}(x) = x^{-1/\gamma}l_F(x)$ or, $U(x) = x^\gamma l_U(x)$ with l_F, l_U slowly varying functions and $U(x) = Q(1 - 1/x)$ with Q the quantile function of F . As for slowly varying functions $l_F(tx)/l_F(t) \rightarrow 1$ for all $x > 0$ and $t \rightarrow \infty$, the conditional distribution of relative excesses $P(\frac{X}{t} > x \mid X > t)$ converges to $x^{-1/\gamma}$ for all $x > 1$.

A graphical tool for checking Pareto type behaviour is the Pareto quantile plot. As log-transformed Pareto distributed random variables are exponentially distributed, the hypothesis of strict Pareto behaviour can be verified by looking at an exponential quantile plot based on the log-transformed data, leading to a Pareto quantile plot

$$(2.3) \quad \left(\log \left(\frac{n+1}{j} \right), \log x_{n-j+1,n} \right), \quad j = 1, \dots, n.$$

For Pareto type distributions, since $U(x) = x^\gamma l_U(x)$, it follows that

$$(2.4) \quad \log U(x) = \gamma \log x + \log l_U(x).$$

Since $\log l_U(x)/\log x \rightarrow 0$ as $x \rightarrow \infty$ we have that $\log U(x) \sim \gamma \log x$ as $x \rightarrow \infty$, which implies that for Pareto type distributed data, the Pareto quantile plot, ultimately for the smaller j -values, shows a linear behaviour, with slope γ .

In Figure 2(a), the Pareto quantile plot is shown for the variable Ca for one of the communities in the Condroz database. The last seven observations that do not follow the ultimate linearity of the rest of the Pareto quantile plot can be considered as outliers with respect to the Pareto model. As can be seen from the Ca versus pH scatterplot given in Figure 1, extreme Ca measurements tend to occur more often at higher pH levels, justifying a tail analysis conditional on higher pH levels. In Figure 2(b), the Pareto quantile plot is shown for the variable Ca, conditional on pH-levels from 7 up to 7.5 (428 observations). Now, six observations can be indicated as possible outliers with respect to the Pareto model. We will use the data set with pH-levels from 7 up to 7.5 in the sequel.

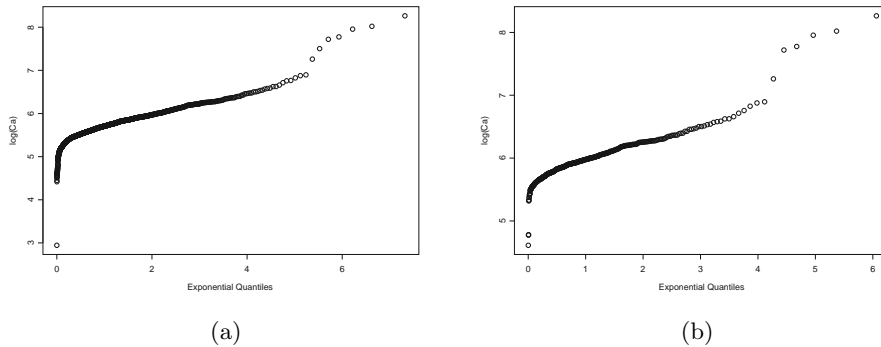


FIGURE 2. Pareto quantile plot of Ca for the Condroz data (a) not taking into account pH and (b) conditional on pH-levels from 7 up to 7.5.

The problem of estimation of the extreme value index, and of extreme quantiles and small exceedance probabilities, in case the distribution is of Pareto type, has been studied in great detail in the recent literature. Hill (1975) introduced the estimator

$$(2.5) \quad \hat{\gamma}_{k,H} = \frac{1}{k} \sum_{j=1}^k \log x_{n-j+1,n} - \log x_{n-k,n}$$

or, equivalently

$$(2.6) \quad \hat{\gamma}_{k,H} = \frac{1}{k} \sum_{j=1}^k j(\log x_{n-j+1,n} - \log x_{n-j,n}).$$

This estimator has received much attention in the literature. As the Hill estimator measures the average increase of the Pareto quantile plot above an anchor point $(\log(\frac{n+1}{k+1}), \log x_{n-k,n})$ it can be interpreted as a slope estimator of the linear part of the Pareto quantile plot. More recently several authors have recognized and exploited the potential of quantile plots in estimating $\gamma > 0$ (Beirlant *et al.* (1996), Schultze and Steinebach (1996) and Kratz and Resnick (1996)).

Recently, in Beirlant *et al.* (1999), Feuerverger and Hall (1999) and Beirlant *et al.* (2002), it was proven that under some suitable conditions on the slowly varying function l_F and for k relatively small with respect to n (i.e. $k/n = o(1)$ as $k, n \rightarrow \infty$), the scaled log-spacings

$$(2.7) \quad y_j = j(\log x_{n-j+1,n} - \log x_{n-j,n}), \quad 1 \leq j \leq k \leq n,$$

approximately follow an exponential regression model

$$(2.8) \quad y_j \sim_d \left(\gamma + b_{n,k} \left(\frac{j}{k+1} \right)^{-\rho} \right) g_j,$$

with the g_j denoting i.i.d. standard exponential random variables, $b_{n,k} \rightarrow 0$ as $k, n \rightarrow \infty$, and $\rho < 0$. From this model γ can be estimated jointly with $b_{n,k}$ and ρ using the maximum likelihood method. In comparison to the Hill estimator, the maximum likelihood estimator for γ based on (2.8) is typically more stable over k and performs better with respect to bias.

Figure 3(a) shows the Hill (solid line) and maximum likelihood (broken line) estimates for the tail index of the conditional distribution of Ca as a function of k . On this plot, we clearly see the influence of the outlying Ca measurements on the γ -estimates. Considering decreasing k , then when k becomes smaller than 100 the estimates first increase drastically and then suddenly decrease for the smallest k -values. Figure 3(b) shows estimates of the asymptotic mean squared error of the Hill estimator

$$(2.9) \quad \widehat{AMSE}(\hat{\gamma}_{k,H}) = \frac{\hat{\gamma}_{k,ML}^2}{k} + \left(\frac{\hat{b}_{n,k,ML}}{1 - \hat{\rho}_{k,ML}} \right)^2,$$

which, as in Beirlant *et al.* (1999), can be used as a criterion to find an optimal sample fraction k_{opt} for the Hill estimator. Here, k_{opt} is found to be 249 (vertical line) leading to an estimate $\hat{\gamma}_{k_{opt},H} = 0.298$.

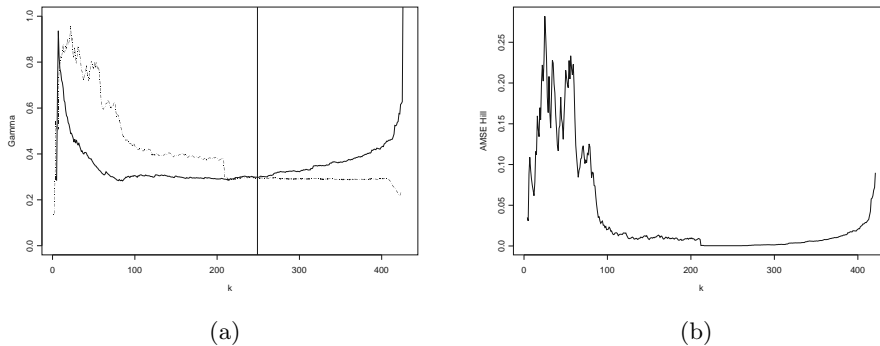


FIGURE 3. (a) Hill (solid line) and maximum likelihood (broken line) estimates of the conditional Ca measurements and (b) AMSE estimates for the Hill estimator as a function of k .

3. Robust estimation of the tail index.

To obtain a robust estimate of γ , we start by transforming the exponential regression model (2.8) to its linearized form, given by

$$(3.1) \quad y_j \sim_d \gamma + b_{n,k} \left(\frac{j}{k+1} \right)^{-\rho} + \gamma e_j, \quad 1 \leq j \leq k \leq n,$$

where $e_j = g_j - 1$. Further we specify a canonical value for ρ . In Matthys and Beirlant (2000) it is shown that for most applications a ρ value between -2 and 0 is most appropriate. Moreover, estimates of γ and $b_{n,k}$ are not highly influenced by a specific choice of ρ . Hence, they recommend to use $\rho = -1$ or $\rho = -0.5$. For a fixed value of k , expression (3.1) becomes a linear regression model with asymmetric errors of the form

$$(3.2) \quad y_j = \theta_1 + \theta_2 t_j + \sigma e_j, \quad j = 1, \dots, k,$$

with $t_j = \left(\frac{j}{k+1} \right)^{-\rho}$, $e_j = g_j - 1$, $\theta_1 = \gamma$, $\theta_2 = b_{n,k}$ and $\sigma = \gamma$.

In Marazzi and Yohai (2001) high efficiency and high breakdown point estimators were proposed for this type of regression models. In general, consider k observations (\mathbf{t}_j, y_j) with $j = 1, \dots, k$ and $\mathbf{t}_j = (1, t_{2j}, \dots, t_{pj})'$ that satisfy the linear relationship

$$(3.3) \quad y_j = \theta_1 + \theta_2 t_{2j} + \dots + \theta_p t_{pj} + \sigma e_j, \quad j = 1, \dots, k,$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)' \in \mathbf{R}^p$ is a vector of regression parameters and σ is a scale parameter. Here, the error terms e_j are assumed to be i.i.d. as a random variable e with cdf $F_{0,1}$, which is the standard element of a parametric location-scale family of asymmetric distributions with density function $f_{\lambda,\sigma}$ and cdf $F_{\lambda,\sigma}(z) = F_{0,1}((z - \lambda)/\sigma)$.

As a robust initial estimator for the general model (3.3) with asymmetric errors, a corrected S-estimator for regression is suggested. The biweight S-estimate $(\hat{\boldsymbol{\theta}}^*, \hat{\sigma}^*)$ of $(\boldsymbol{\theta}, \sigma)$ with 50% breakdown value is defined by $\hat{\boldsymbol{\theta}}^* = \arg \min_{\boldsymbol{\theta}} S_{m_0}(\boldsymbol{\theta})$ and $\hat{\sigma}^* = S_{m_0}(\hat{\boldsymbol{\theta}}^*)$. For a given $\boldsymbol{\theta}$, $S_{m_0}(\boldsymbol{\theta})$ is an M-estimator of scale of the residuals, as the solution of

$$(3.4) \quad \frac{1}{k-p} \sum_{j=1}^k \chi_{m_0}((y_j - \mathbf{t}'_j \boldsymbol{\theta})/S) = 0.5$$

with respect to S (Rousseeuw and Yohai (1984)). The function χ_m and the constants a_0 and m_0 are defined by

$$(3.5) \quad \chi_m(z) = \begin{cases} 3(z/m)^2 - 3(z/m)^4 + (z/m)^6 & \text{if } |z| \leq m, \\ 1 & \text{if } |z| > m, \end{cases}$$

and satisfy $\int \psi_{m_0}(z - a_0) f_{0,1}(z) dz = 0$ and $\int \chi_{m_0}(z - a_0) f_{0,1}(z) dz = 0.5$, where $\psi_m = \frac{d}{dz} \chi_m$. A consistently corrected S-estimate for $\boldsymbol{\theta}$ and σ is then given by

$$\hat{\theta}_1^0 = \hat{\theta}_1^* - \hat{\sigma}^* a_0, \hat{\theta}_j^0 = \hat{\theta}_j^* \text{ for } j = 2, \dots, p \text{ and } \hat{\sigma}^0 = \hat{\sigma}^*.$$

Next, a reweighting step is applied, as proposed in Rousseeuw and Leroy (1987). Observations whose standardized residual $r_j = (y_j - \mathbf{t}'_j \hat{\boldsymbol{\theta}}^0) / \hat{\sigma}^0$ has a large negative log-likelihood $\rho_j = \rho(r_j) = -\log(f_{0,1}(r_j))$ have a small likelihood under the model $F_{0,1}$ and can be flagged as outliers. They are given a weight $w_j = I_{(\rho_j < \eta)}$, where I denotes the indicator function, and η is a large quantile of the cdf of $\rho(e)$. Finally, the maximum likelihood estimates of $(\boldsymbol{\theta}, \sigma)$ in model (3.3) are computed on the data points with $w_j = 1$.

When we apply this robust procedure to (3.2) we obtain preliminary estimates $(\hat{\theta}_1^0, \hat{\theta}_2^0), \hat{\sigma}_1^0$ and standardized residuals

$$(3.6) \quad r_j = \frac{\left(y_j - \hat{\theta}_1^0 - \hat{\theta}_2^0 \left(\frac{j}{k+1} \right)^{-\rho} \right)}{\hat{\sigma}^0}.$$

Note that the constants a_0 and m_0 that are used in the definition of the S-estimators are found to be $a_0 = -0.5700$ and $k_0 = 0.9466$. From the model cdf $F_{0,1}(z) = 1 - e^{-(z+1)}$ for $z \geq -1$, it follows that $\rho(e)$ is standard exponentially distributed, hence we set $\eta = -\log(1 - 0.95)$, the 0.95-quantile of the standard exponential distribution. Of course, also other quantiles could be considered as well. We now indicate the zero weight log-spacings y_j as outliers and remove the corresponding x_i from our data set. More precisely, let y_J be the outlying log-spacing with largest index, then all x_i with $n - J + 1 \leq i \leq n$ are flagged as outliers under the Pareto model.

4. Application

The results after applying our proposal to the Condroz data, are depicted in Figure 4. Note that we have set $\rho = -1$ in (3.1). We have also tried some other values of ρ between -2 and 0 but this did not change the result. The broken black line shows the maximum likelihood estimates after removal of the outliers that were found for each k , whereas the solid black line exposes the Hill estimator on the same reduced data sets. On this plot we have also superimposed (in grey) the ML and Hill curves of Figure 3 that are based on the full data set. We see that both robust curves yield much lower estimates for the tail index. Moreover they are rather constant for intermediate values of k , which gives support to the Pareto model assumption. For very small values of k , the estimates are however still not stable. This is due to the small number of observations in the regression model (3.2), where even an estimator with 50% breakdown is biased in the presence of outliers.

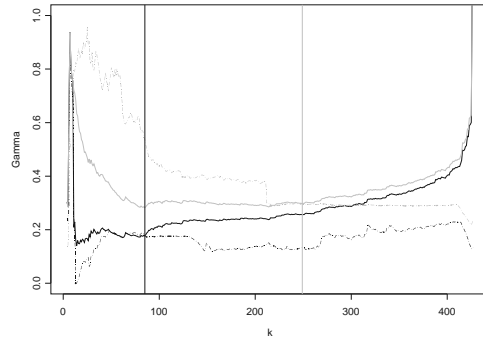


FIGURE 4. Hill (solid line) and maximum likelihood (broken line) estimates of the conditional Ca measurements after rejection of the outliers found for each k , together with the Hill (solid grey line) and maximum likelihood (broken grey line) estimates before rejection.

It is observed that our method found six outliers for most k values, except for the smaller ones. The dark vertical line in Figure 4 is set at $k_{opt} = 85$. To obtain this optimal sample fraction for the Hill estimator, we computed its AMSE as in (2.9). Since not all the ML estimates were based on the same data set for all k , here the ML estimators were computed on the data set after the six largest observations were removed. The AMSE curve, shown in Figure 5, attains its minimum at $k_{opt} = 85$, from which $\hat{\gamma}_{k_{opt},H} = 0.177$ follows.

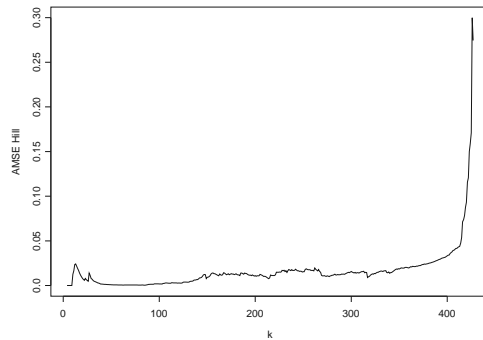


FIGURE 5. AMSE estimates for the Hill estimator after rejection of the six largest observations.

Figure 6 shows the $\rho_j = \rho(r_j)$ for the regression data (t_j, y_j) from model (3.2) for $k = k_{\text{opt}} = 85$, together with the cut-off line that separates the outliers from the regular observations. We see that t_6 is the observation with largest index that exceeds the cut-off, hence six observations are flagged out as being unlikely under the Pareto model. This confirms our findings from the Pareto quantile plot in Figure 2(b).

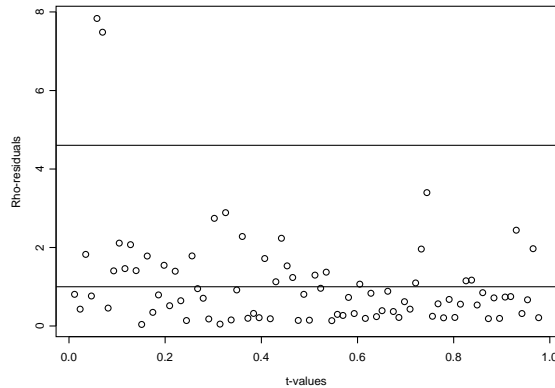


FIGURE 6. The $\rho(r_j)$ for the regression data (t_j, y_j) from model (3.2) with $k = 85$ together with the expected value and cut-off line that separates the outliers from the regular observations.

5. Conclusions and outlook

In this paper we have introduced a new robust estimator of the tail index of Pareto type distributions. It is obtained by applying a robust regression estimator for a linear model with asymmetric errors to scaled log-spacings. When we applied this method to Ca and pH measurements from the Condroz region in Belgium, we could easily identify the outliers that were also seen on a Pareto quantile plot, and we obtained a much lower estimate of the tail index.

In our further research we will study this robust estimator in more detail. We will investigate its breakdown value, its influence function and its performance at simulated data sets.

As a different approach, we will also study the application of a robust regression estimator to the Pareto quantile plot data (2.3). In particular we will consider

the deepest regression method (Rousseeuw and Hubert (1999)) as it yields consistent estimates under heteroscedastic error distributions. Nonconstant variances are likely to occur in quantile plots.

References

- [1] J. Beirlant, G. Dierckx, Y. Goegebeur and G. Matthys (1999), *Tail index estimation and an exponential regression model*. *Extremes*, **2**(2), 177-200.
- [2] J. Beirlant, G. Dierckx, A. Guillou and C. Stărică (2002), *On exponential representations of log-spacings of extreme order statistics*. *Extremes*, **5**, 157-180.
- [3] J. Beirlant, P. Vynckier and J.L. Teugels (1996), *Tail index estimation, Pareto quantile plots, and regression diagnostics*. *J. Am. Statist. Ass.*, **91**, 1659-1667.
- [4] P. Embrechts, C. Klüppelberg, T. Mikosch (1998), *Modelling extremal events for insurance and finance*. Springer, New York.
- [5] A. Feuerverger, P. Hall (1999). *Estimating a tail exponent by modelling departure from a Pareto distribution*. *Ann. Statist.*, **27**, 760-781.
- [6] Y. Goegebeur, V. Planchon, J. Beirlant and R. Oger (2002), *Quality Assessment of Petrochemical Data Using Extreme Value Methodology*. Technical Report. [Full text (PDF)] in http://www.kuleuven.ac.be/ucs/technical_reports.htm.
- [7] B.M. Hill (1975), *A simple general approach to inference about the tail of a distribution*. *Ann. Statist.*, **3**, 1163-1174.
- [8] M. Kratz and S. Resnick (1996), *The qq-estimator of the index of regular variation*. *Communications in Statistics: Stochastic Models*, **12**, 699-724.
- [9] A. Marazzi and V. Yohai (2002), *Adaptively truncated maximum likelihood regression with asymmetric errors*. *Journal of Statistical Planning and Inference*, in press.
- [10] G. Matthys and J. Beirlant (2000), *Adaptive threshold selection in tail index estimation*. In: *Extremes and Integrated Risk Management*, Ed. Paul Embrechts, Risk Books, UBS Warburg, 37-49.
- [11] P.J. Rousseeuw and M. Hubert (1999), *Regression depth*. *J. Am. Statist. Ass.*, **94**, 388-402.
- [12] P.J. Rousseeuw, A. Leroy (1987), *Robust Regression and Outlier Detection*. John Wiley, New York.
- [13] P.J. Rousseeuw, V. Yohai (1984), *Robust regression by means of S estimators*. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist.*, **26**, 256-272. Springer, New York.
- [14] J. Schultze and J. Steinebach (1996), *On least squares estimates of an exponential tail coefficient*. *Statistics and Decisions*, **14**, 353-372.

Department of Mathematics and University Center of Statistics, Katholieke Universiteit Leuven, de Croylaan 54, B-3001 Heverlee, Belgium

E-mail address: Bjorn.Vandewalle@wis.kuleuven.ac.be

Department of Mathematics and University Center of Statistics, Katholieke Universiteit Leuven, de Croylaan 54, B-3001 Heverlee, Belgium

E-mail address: Jan.Beirlant@wis.kuleuven.ac.be

Department of Mathematics and University Center of Statistics, Katholieke Universiteit Leuven, de Croylaan 54, B-3001 Heverlee, Belgium

E-mail address: Mia.Hubert@wis.kuleuven.ac.be