REVSTAT – Statistical Journal
Volume 6, Number 1, March 2008, 51–69

LINKING PARETO-TAIL KERNEL GOODNESS-OF-FIT STATISTICS WITH TAIL INDEX AT OPTIMAL THRESHOLD AND SECOND ORDER ESTIMATION

Authors: Yuri Goegebeur
– Department of Statistics,
University of Southern Denmark,
J.B. Winsløws Vej 9B, 5000 Odense C, Denmark
yuri.goegebeur@stat.sdu.dk

Jan Beirlant
– Department of Mathematics and Leuven Statistics Research Center,
Catholic University of Leuven,
Celestijnenlaan 200B, 3001 Heverlee, Belgium
jan.beirlant@wis.kuleuven.be

Tertius de Wet
– Department of Statistics and Actuarial Science,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa
tdewet@sun.ac.za

Abstract:

- In this paper the relation between goodness-of-fit testing and the optimal selection of the sample fraction for tail estimation, for instance using Hill's estimator, is examined. We consider this problem under a general kernel goodness-of-fit test statistic for assessing whether a sample is consistent with the Pareto-type model. The derivation of the class of kernel goodness-of-fit statistics is based on the close link between the strict Pareto and the exponential distribution, and puts some of the available goodness-of-fit procedures for the latter in a broader perspective. Two important special cases of the kernel statistic, the Jackson and the Lewis statistic, will be discussed in greater depth. The relationship between the limiting distribution of the Lewis statistic and the bias-component of the asymptotic mean squared error of the Hill estimator is exploited to construct a new tail sample fraction selection criterion for the latter. The methodology is illustrated on a case study.

Key-Words:

- *extreme value statistics; Pareto-type distribution; goodness-of-fit; threshold selection.*

AMS Subject Classification:

- 62G32, 62G30, 62E20.

1. INTRODUCTION

Extreme value theory focuses on characteristics related to the tail of a distribution function such as indices describing tail decay, extreme quantiles and small tail probabilities. In the process of making inferences about the far tail of a distribution function, it is necessary to extend the empirical distribution function beyond the available data. This is typically done by only considering the upper k order statistics, which then entails the issue of how to select a good, or, if possible, an optimal, k-value. Many proposals to tackle this issue have been made in the literature, see for instance Drees and Kaufmann (1998), Danielsson *et al.* (2001), Guillou and Hall (2001), and Beirlant *et al.* (2002). In this paper we use recently introduced kernel goodness-of-fit statistics for Pareto-type behavior as a basis for proposing a new procedure for selecting k.

Consider random variables $X_1, ..., X_n$ independent and identically distributed (i.i.d.) according to some distribution function F and let $X_{1,n} \le ... \le X_{n,n}$ denote the corresponding ascending order statistics. If for sequences of constants $(a_n > 0)_n$ and $(b_n)_n$

$$\lim_{n\to\infty} P\left(\frac{X_{n,n} - b_n}{a_n} \le x\right) = \lim_{n\to\infty} F^n(b_n + a_n x) = G(x) \tag{1.1}$$

at all continuity points of G, for G some non-degenerate distribution function, then G has to be of the generalized extreme value (GEV) type:

$$G_\gamma(x) = \begin{cases} \exp\bigl(-(1+\gamma x)^{-1/\gamma}\bigr), & 1+\gamma x > 0, \quad \gamma \neq 0 \ , \\ \exp\bigl(-\exp(-x)\bigr), & x \in \mathbb{R}, \quad \gamma = 0 \ . \end{cases} \tag{1.2}$$

Note that the behavior of this distribution function is governed by the single parameter γ, called the extreme value index. If F satisfies (1.1)–(1.2), then it is said to belong to the max-domain of attraction of G_γ, denoted $F \in \mathcal{D}(G_\gamma)$. An important subclass of the max-domain of attraction of the GEV distribution is the class of the Pareto-type models. These are characterized by heavy tailed distribution functions with infinite right endpoints, having $\gamma > 0$.

For Pareto-type distributions the *first order condition* (1.1) can be expressed in an equivalent way in terms of the survival function $1-F$:

$$1 - F(x) = x^{-1/\gamma}\, \ell_F(x) \ , \qquad x > 0 \ , \tag{1.3}$$

where ℓ_F denotes a slowly varying function at infinity, i.e.

$$\frac{\ell_F(\lambda x)}{\ell_F(x)} \to 1 \quad \text{as } x \to \infty \qquad \text{for all } \lambda > 0 \ . \tag{1.4}$$

In terms of the tail quantile function U, defined as $U(x) = \inf\{y \colon F(y) \geq 1-1/x\}$, $x > 1$, we then have that

$$U(x) = x^{\gamma}\, \ell_U(x) \,, \tag{1.5}$$

where ℓ_U again denotes a slowly varying function at infinity (Gnedenko, 1943). Pareto-type tails are systematically used in certain branches of non-life insurance, as well as in finance (stock returns), telecommunication (file sizes, waiting times), geology (diamond values, earthquake magnitudes), and many others. In the analysis of heavy tailed distributions the estimation of γ, and the subsequent estimation of extreme quantiles, assume a central position. Several estimators for γ have been proposed in the literature, and their asymptotic distributions established, usually under a *second order condition* on the tail behavior (see e.g. Beirlant *et al.*, 2004, and de Haan and Ferreira, 2006). This condition specifies the rate of convergence of ratios of the form $\ell(\lambda x)/\ell(x)$, with ℓ a slowly varying function, to their limit (see Bingham *et al.*, 1987).

Second order condition $(\mathcal{R}_\ell)$**.** *A slowly varying function ℓ satisfies a second order condition if there exists a real constant $\rho \leq 0$ and a rate function b satisfying $b(x) \to 0$ as $x \to \infty$, such that for all $\lambda \geq 1$, as $x \to \infty$,*

$$\frac{\ell(\lambda x)}{\ell(x)} - 1 \;\sim\; b(x)\, \frac{\lambda^{\rho} - 1}{\rho} \;.$$

In the context of estimation of γ, it is then typically assumed that the slowly varying function ℓ_U in (1.5) satisfies a second order condition. Of interest for the subsequent development of a procedure for selecting a threshold, is testing of the hypothesis that the underlying distribution is of Pareto-type together with a second order condition holding. Formally, this hypothesis can be stated as

$$H_0 \colon\; F \text{ is of Pareto-type with } \ell_U \text{ satisfying } \mathcal{R}_\ell \;. \tag{1.6}$$

It is well known that the log-transform of a (strict) Pareto random variable has an exponential distribution. Our approach to testing H_0 is to exploit this fact by considering goodness-of-fit tests for exponentiality as possible test statistics. The literature on goodness-of-fit tests for the exponential distribution is quite elaborate, see e.g. Henze and Meintanis (2005) for a recent overview of this literature. Such tests often take the form of the ratio of two estimators for the exponential scale parameter. In a similar way, one can construct test statistics as ratios of two estimators for the extreme value index γ.

Of course it is intuitively clear that goodness-of-fit procedures should enable one to choose an appropriate threshold $X_{n-k,n}$ for tail index estimation. Hill (1975) already recognized this idea, see also Beirlant *et al.* (1996). Typically, however, goodness-of-fit based procedures are too conservative with respect to the null hypothesis, leading to too high values of k (or equivalently too low thresholds) with respect to the asymptotic mean squared error (AMSE) criterion. Based on the limiting distribution of our kernel goodness-of-fit statistic, we propose an estimator for the bias component of the AMSE of the Hill estimator, yielding an alternative method to select the threshold $X_{n-k,n}$.

The remainder of this paper is organized as follows. In Section 2 we introduce a general kernel goodness-of-fit statistic for assessing whether a sample is consistent with the Pareto-type model, and state its main properties. Section 3 deals with the link between goodness-of-fit testing and the selection of the optimal tail sample fraction, for instance when using the Hill estimator. In Section 4 we illustrate the methodology with a practical example.

2. A KERNEL GOODNESS-OF-FIT STATISTIC FOR PARETO-TYPE BEHAVIOR

Consider $X_1, ..., X_n$ i.i.d. $Pa(1/\gamma)$ random variables, where $Pa(1/\gamma)$ denotes the strict Pareto distribution with Pareto index $1/\gamma$, i.e. $F(x) = 1-x^{-1/\gamma}$, $x > 1$, and the corresponding ascending order statistics $X_{1,n} \leq ... \leq X_{n,n}$. Then the ratios $Y_{j,k} = X_{n-k+j,n}/X_{n-k,n}$, $j = 1, ..., k$, are jointly distributed as the order statistics of a random sample of size k from the $Pa(1/\gamma)$ distribution. Consequently, $Y^*_{j,k} = \log Y_{j,k}$ behave as $Exp(1/\gamma)$ order statistics, where $Exp(1/\gamma)$ denotes the exponential distribution with mean γ. In case the data originate from a Pareto-type distribution these properties hold approximately above a sufficiently high threshold. This close link between the Pareto-type and the exponential model will be exploited in the derivation of goodness-of-fit tests for the former. The literature on testing whether a sample is consistent with an exponential distribution is quite extensive, see for instance Stephens (1986) and Henze and Meintanis (2005), and the references therein. These exponential goodness-of-fit test statistics are quite often a ratio of two estimators for the exponential scale parameter (e.g. Lewis, 1965, Jackson, 1967, de Wet and Venter, 1973). Inspired by this and based on the above properties of $Pa(1/\gamma)$ order statistics, we apply a similar ratio to the k largest order statistics, leading to the following test statistic

$$\frac{\frac{1}{k}\sum_{j=1}^{k} K\Big(\frac{j}{k+1}\Big)\, Z_j}{H_{k,n}}\,, \tag{2.1}$$

with K denoting a kernel function satisfying $\int_0^1 K(u)\,du = 0$, $Z_j = j(\log X_{n-j+1,n} - \log X_{n-j,n})$, and $H_{k,n} = \frac{1}{k}\sum_{j=1}^{k} Z_j$, the Hill estimator for γ (Hill, 1975).

In Goegebeur *et al.* (2007), generalizing Beirlant *et al.* (2006), the statistic in (2.1) was proposed and its limiting distribution derived under the hypothesis stated in (1.6), some mild regularity conditions on K, and an intermediate

k sequence, i.e. $k = k_n \to \infty$, $k_n = o(n)$ as $n \to \infty$. We use $\log^+ u$ to denote $\max\{\log u, 1\}$.

Theorem 2.1. *Consider $X_1, ..., X_n$ i.i.d. random variables according to distribution function F, where $F \in \mathcal{D}(G_\gamma)$ for some $\gamma > 0$. Assume ℓ_U satisfies $\mathcal{R}_\ell$ and let $K(t) = \frac{1}{t} \int_0^t u(v)\, dv$ for some function u satisfying $\left|k \int_{(j-1)/k}^{j/k} u(t)\, dt\right| \le f\left(\frac{j}{k+1}\right)$ for some positive continuous function f defined on $(0,1)$ such that $\int_0^1 \log^+(1/w) f(w)\, dw < \infty$ in case $\rho < 0$ and $\int_0^1 w^{-\xi} f(w)\, dw < \infty$ for some small $\xi > 0$ in case $\rho = 0$, $\int_0^1 |K(w)|^{2+\delta} dw < \infty$ for some $\delta > 0$ and $\frac{1}{\sqrt{k}} \sum_{j=1}^k K\left(\frac{j}{k+1}\right) \to 0$ as $k \to \infty$. Then as $k, n \to \infty$, $k/n \to 0$ and $\sqrt{k}\, b(n/k) \to c$,*

$$\frac{\sqrt{k}}{H_{k,n}} \frac{1}{k} \sum_{j=1}^k K\left(\frac{j}{k+1}\right) Z_j \xrightarrow{\mathcal{L}} N\left(\frac{c}{\gamma} \int_0^1 K(u)\, u^{-\rho}\, du,\ \int_0^1 K^2(u)\, du\right).$$

Using this theorem, the decision rule for testing the hypothesis (1.6) at the significance level α is to reject H_0 if

$$\sqrt{k} \left| \frac{1}{k\, H_{k,n}} \sum_{j=1}^k K\left(\frac{j}{k+1}\right) Z_j - \frac{b(n/k)}{\gamma} \int_0^1 K(u)\, u^{-\rho} du \right| >$$
$$> \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\int_0^1 K^2(u)\, du}\ ,$$

where Φ^{-1} denotes the standard normal quantile function. However, for practical application this rule is not very helpful as it depends on the unknown function b as well as on the parameters γ and ρ. A way out of this is to choose k relatively small, i.e. small enough to guarantee that $\sqrt{k}\, b(n/k) \approx 0$, which then leads to the rule to reject H_0 if

$$\frac{\sqrt{k}}{H_{k,n}} \left| \frac{1}{k} \sum_{j=1}^k K\left(\frac{j}{k+1}\right) Z_j \right| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\int_0^1 K^2(u)\, du}\ .$$

For a detailed description of the fundamental properties of the goodness-of-fit statistic and for an evaluation of its small sample performance through a simulation study, we refer to Goegebeur *et al.* (2007). We will now describe two important special cases of this kernel-type goodness-of-fit statistic, the Jackson (Jackson, 1967) and the Lewis (Lewis, 1965) statistics, in more detail.

2.1. Jackson kernel function

We modify the Jackson statistic (Jackson, 1967), originally proposed as a goodness-of-fit statistic for testing exponentiality, in such a way that it measures the linearity of the k largest observations on the Pareto quantile plot. Consider $X_1, ..., X_n$ i.i.d. *Exp*(δ) random variables. The Jackson statistic is given by

$$T_{\mathrm{J}} = \frac{\sum_{j=1}^{n} t_{j,n} X_{j,n}}{\sum_{j=1}^{n} X_j} \tag{2.2}$$

where $t_{j,n} = \delta E(X_{j,n}) = \sum_{i=1}^{j} (n-i+1)^{-1}$. The numerator is clearly a sum of cross products of order statistics and their expected values. The denominator is introduced to eliminate the dependence on the nuisance parameter δ. The Jackson statistic can hence be considered as a 'correlation like' statistic based on the exponential quantile plot. The limiting distribution of the appropriately normalized Jackson statistic was derived by Jackson (1976), in particular $\sqrt{n}(T_{\mathrm{J}} - 2) \xrightarrow{\mathcal{D}} N(0,1)$, as $n \to \infty$. For our purposes it is more convenient to express (2.2) in terms of the standardized spacings $V_j = (n-j+1)(X_{j,n} - X_{j-1,n})$, $j = 1, ..., n$. From the Rényi representation these are known to be i.i.d. *Exp*(δ) random variables. Rearranging terms of (2.2), it can be shown that

$$T_{\mathrm{J}} = \frac{\sum_{j=1}^{n} C_{j,n} V_j}{\sum_{j=1}^{n} V_j}$$

where $C_{1,n} = 1$ and $C_{j,n} = 1 + t_{j-1,n}$, $j = 2, ..., n$.

We will now adjust the Jackson statistic in such a way that it measures the linearity of the k upper order statistics on the Pareto quantile plot. Consider a random sample $X_1, ..., X_n$ of Pareto-type distributed random variables. Application of the Jackson statistic to $Y^*_{j,k}$, $j = 1, ..., k$, yields, after suitable normalization and rearranging terms,

$$T^{\mathrm{J}}_{k,n} = \sqrt{k}\, \frac{\frac{1}{k} \sum_{j=1}^{k} K_{\mathrm{J}}\Big(\frac{j}{k+1}\Big) Z_j}{H_{k,n}}$$

where $K_{\mathrm{J}}(u) = -1 - \log u$, see also Beirlant *et al.* (2006). The kernel function K_{J} satisfies the conditions of Theorem 2.1 with $u(s) = -2 - \log s$, and hence we can state the following proposition.

Proposition 2.1. *Assume $X_1, ..., X_n$ i.i.d. random variables according to distribution function F, where $F \in \mathcal{D}(G_\gamma)$ for some $\gamma > 0$ and ℓ_U satisfying $\mathcal{R}_\ell$. Then as $k, n \to \infty$, $k/n \to 0$ and $\sqrt{k}\, b(n/k) \to c$,*

$$\frac{\sqrt{k}}{H_{k,n}} \frac{1}{k} \sum_{j=1}^{k} K_{\mathrm{J}}\bigg(\frac{j}{k+1}\bigg) Z_j \xrightarrow{\mathcal{L}} N\bigg(\frac{c\,\rho}{\gamma(1-\rho)^2}, 1\bigg) .$$

Note that the normal limit is not necessarily centered at zero, i.e. the statistic may exhibit some bias. The centering depends, besides γ, on the slowly varying function ℓ_U through the parameters ρ and c.

2.2. Lewis kernel function

As a second example we study the Lewis goodness-of-fit statistic. Consider a sample $X_1, ..., X_n$ of i.i.d. $Exp(\delta)$ random variables. The Lewis statistic is given by

$$T_{\mathrm{L}} = \frac{\sum_{j=1}^{n} \frac{j}{n+1}\, V_{n-j+1}}{\sum_{j=1}^{n} X_j} \ ,$$

and $\sqrt{n}(T_{\mathrm{L}} - 1/2) \xrightarrow{\mathcal{L}} N(0, 1/12)$, as $n \to \infty$ (Lewis, 1965). In case of a random sample $X_1, ..., X_n$ of Pareto-type random variables, we can apply the Lewis statistic to $Y^*_{j,k}$, $j = 1, ..., k$, yielding, after appropriate normalization and rearranging terms,

$$T^{\mathrm{L}}_{k,n} = \sqrt{k}\ \frac{\frac{1}{k} \sum_{j=1}^{k} K_{\mathrm{L}}\Big(\frac{j}{k+1}\Big)\, Z_j}{H_{k,n}} \ ,$$

with $K_{\mathrm{L}}(u) = u - 0.5$. The function K_{L} satisfies the conditions of Theorem 2.1 with $u(s) = 2s - 0.5$, leading to the following proposition:

Proposition 2.2. *Assume $X_1, ..., X_n$ i.i.d. random variables according to distribution function F, where $F \in \mathcal{D}(G_\gamma)$ for some $\gamma > 0$ and ℓ_U satisfying $\mathcal{R}_\ell$. Then as $k, n \to \infty$, $k/n \to 0$ and $\sqrt{k}\, b(n/k) \to c$,*

$$\frac{\sqrt{k}}{H_{k,n}}\, \frac{1}{k} \sum_{j=1}^{k} K_{\mathrm{L}}\Big(\frac{j}{k+1}\Big) Z_j \xrightarrow{\mathcal{L}} N\Big(-\frac{c\,\rho}{2\,\gamma\,(1-\rho)(2-\rho)}\,,\, \frac{1}{12}\Big) \ .$$

Note that for the same value of c, the absolute value of the asymptotic bias of the Lewis statistic is smaller than the absolute bias of the Jackson statistic.

2.3. Bias-correction

As mentioned above, the bias of the kernel statistics may make it difficult to evaluate the nature of the tail behavior. It is, however, possible to derive, for a given kernel function K, a bias-corrected kernel function, denoted $K_{\mathrm{BC}}(\cdot;\rho)$, i.e. a kernel satisfying $\int_0^1 K_{\mathrm{BC}}(u;\rho)\, u^{-\rho}\, du = 0$. To obtain such a bias-corrected kernel, note that both the numerator and the denominator of the general kernel statistic (2.1) are weighted averages of the Z_j, $j = 1, ..., k$. Within the framework of Pareto-type tails and assuming condition $\mathcal{R}_\ell$ on ℓ_U holds, with $\rho < 0$, Beirlant *et al.* (1999) derived the following approximate representation for log-spacings of successive order statistics

$$Z_j \sim \gamma + b_{n,k} \Big(\frac{j}{k+1}\Big)^{-\rho} + \varepsilon_j \ , \qquad j = 1, ..., k \ , \tag{2.3}$$

where $b_{n,k} = b(n/k)$ and ε_j, $j=1,...,k$, are zero centered error terms, or, equivalently

$$Z_j - b_{n,k}\left(\frac{j}{k+1}\right)^{-\rho} \sim \gamma + \varepsilon_j \ , \qquad j=1,...,k \ .$$

This then motivates the following bias-corrected statistic

$$\text{(2.4)} \qquad \sqrt{k}\ \frac{\frac{1}{k}\sum_{j=1}^{k} K\left(\frac{j}{k+1}\right)\left(Z_j - \hat{b}_{\mathrm{LS},k}(\rho)\left(\frac{j}{k+1}\right)^{-\rho}\right)}{\hat{\gamma}_{\mathrm{LS},k}(\rho)} \ ,$$

with $\hat{\gamma}_{\mathrm{LS},k}(\rho)$ and $\hat{b}_{\mathrm{LS},k}(\rho)$ the least squares estimators for respectively γ and $b_{n,k}$ obtained from (2.3), taking ρ as fixed:

$$\text{(2.5)} \qquad \hat{\gamma}_{\mathrm{LS},k}(\rho) \ = \ \frac{1}{k}\sum_{j=1}^{k} Z_j - \frac{\hat{b}_{\mathrm{LS},k}(\rho)}{1-\rho} \ ,$$

$$\text{(2.6)} \qquad \hat{b}_{\mathrm{LS},k}(\rho) \ = \ \frac{(1-\rho)^2\,(1-2\rho)}{\rho^2}\ \frac{1}{k}\sum_{j=1}^{k}\left(\left(\frac{j}{k+1}\right)^{-\rho} - \frac{1}{1-\rho}\right) Z_j \ .$$

After some additional straightforward manipulations on (2.4), we obtain the bias-corrected kernel function:

$$\text{(2.7)} \qquad K_{\mathrm{BC}}(u;\rho) \ = \ K(u) - \frac{(1-\rho)^2\,(1-2\rho)}{\rho^2}\left(u^{-\rho} - \frac{1}{1-\rho}\right)\int_0^1 K(v)\,v^{-\rho}\,dv \ .$$

It is easy to verify that for kernel functions K satisfying the conditions of Theorem 2.1, K_{BC} will also satisfy these conditions with $\int_0^1 K_{\mathrm{BC}}(u;\rho)\,u^{-\rho}\,du = 0$, hence leading to an asymptotic normal distribution with null mean value, stated in the next theorem (Goegebeur *et al.*, 2007).

Theorem 2.2. *Consider $X_1,...,X_n$ i.i.d. random variables according to distribution function F, where $F \in \mathcal{D}(G_\gamma)$ for some $\gamma > 0$, and with ℓ_U satisfying $\mathcal{R}_\ell$, fixed $\rho < 0$. If K satisfies the conditions of Theorem 2.1, then if $k,n \to \infty$, $k/n \to 0$ and $\sqrt{k}\,b(n/k) \to c$,*

$$\frac{\sqrt{k}}{\hat{\gamma}_{\mathrm{LS},k}(\rho)}\ \frac{1}{k}\sum_{j=1}^{k} K_{\mathrm{BC}}\left(\frac{j}{k+1};\rho\right) Z_j \ \xrightarrow{\mathcal{L}}\ N\left(0, \int_0^1 K_{\mathrm{BC}}^2(u;\rho)\,du\right) .$$

The bias-correcting effect of the above described operation can be readily seen from the limiting distribution: whatever c the normal limit is centered at zero. In case of the bias-corrected Lewis kernel function, denoted K_{BCL}, obtained by plugging K_{L} into (2.7), $\int_0^1 K_{\mathrm{BCL}}^2(u;\rho)\,du = 0$ if $\rho = -1$, leading to a degenerate distribution at zero. When dealing with this kernel function we exclude the value $\rho = -1$.

3. SELECTION OF THE NUMBER OF UPPER ORDER STATISTICS FOR TAIL INDEX ESTIMATION

In this section we discuss the use of the kernel goodness-of-fit statistic for selecting the optimal threshold in tail index estimation. The discussion will be focused on the Hill estimator, but the idea can of course be equally well applied to other estimators for $\gamma > 0$. The basic idea is to exploit the relationship between the bias component of the asymptotic mean squared error of the Hill estimator, denoted $AMSE(H_{k,n})$, and the kernel goodness-of-fit statistics introduced above. It is well known that for the Hill estimator

$$\begin{aligned} AMSE(H_{k,n}) &= \frac{\gamma^2}{k} + \left(\frac{b_{n,k}}{1-\rho}\right)^2 \\ &= \gamma^2 \left[\frac{1}{k} + \left(\frac{b_{n,k}}{\gamma(1-\rho)}\right)^2\right] . \end{aligned}$$

From Theorem 2.1, we have for the general kernel goodness-of-fit statistic, for k, n large and k/n small,

$$\frac{\frac{1}{k}\sum_{j=1}^{k} K\Big(\frac{j}{k+1}\Big) Z_j}{H_{k,n}} \;\sim\; \frac{b_{n,k}}{\gamma} \int_0^1 K(u)\, u^{-\rho}\, du ,$$

and hence, provided $\int_0^1 K(u)\, u^{-\rho}\, du \neq 0$,

$$\frac{b_{n,k}}{\gamma(1-\rho)} \;\sim\; \frac{\frac{1}{k}\sum_{j=1}^{k} K\Big(\frac{j}{k+1}\Big) Z_j}{(1-\rho)\, H_{k,n} \int_0^1 K(u)\, u^{-\rho}\, du} , \tag{3.1}$$

leading to the following approximation to $AMSE(H_{k,n})$

$$\widehat{AMSE}(H_{k,n}) = \gamma^2 \left\{\frac{1}{k} + \left[\frac{\frac{1}{k}\sum_{j=1}^{k} K\Big(\frac{j}{k+1}\Big) Z_j}{(1-\rho)\, H_{k,n} \int_0^1 K(u)\, u^{-\rho}\, du}\right]^2\right\} .$$

The optimal choice of k is then approximated by

$$\hat{k}_{\text{opt}} = \arg\min \left\{\frac{1}{k} + \left[\frac{\frac{1}{k}\sum_{j=1}^{k} K\Big(\frac{j}{k+1}\Big) Z_j}{(1-\rho)\, H_{k,n} \int_0^1 K(u)\, u^{-\rho}\, du}\right]^2\right\} . \tag{3.2}$$

Note that the squared goodness-of-fit statistic is to be complemented by a penalty $1/k$ in order to prevent choosing too small values of k. Also the role of ρ is important: typically, the smaller $|\rho|$ the heavier the ρ-factor with the test statistic leading to small values of k.

In the remainder of this section we will concentrate on the Lewis goodness-of-fit statistic, but of course similar results can be easily obtained for other kernel functions. For the Lewis statistic, $K_{\text{L}}(u) = u - 0.5$, and hence, $\int_0^1 K(u)\, u^{-\rho}\, du =$

$|\rho|/\big[2\,(1-\rho)\,(2-\rho)\big]$, which leads to minimizing

$$\frac{1}{k} + \left[\frac{2\,(2-\rho)}{|\rho|\,\sqrt{k}}\;T^{\mathrm{L}}_{k,n}\right]^2 \tag{3.3}$$

with respect to k.

Practical implementations based on (3.2) or (3.3) require of course an estimate for the unknown parameter ρ. Gomes *et al.* (2002) proposed ratios involving different powers of statistics $M^{(r)}_{k,n}$, with

$$M^{(r)}_{k,n} \;=\; \frac{1}{k}\sum_{j=1}^{k}\Big(\log X_{n-j+1,n} - \log X_{n-k,n}\Big)^r$$

to derive estimators for ρ. In a similar fashion, we propose an estimator for ρ using a ratio of two kernel goodness-of-fit statistics. Define

$$T_i \;=\; \frac{1}{k}\sum_{j=1}^{k} K_i\bigg(\frac{j}{k+1}\bigg)\,Z_j\;, \qquad i=1,2\;,$$

where the indices 1 and 2 refer to the Jackson and Lewis goodness-of-fit statistics, for instance. From Proposition 2.1 and Proposition 2.2, we have, in probability,

$$\begin{aligned} T_1 \;&\sim\; b_{n,k}\,\frac{\rho}{(1-\rho)^2}\;, \\ T_2 \;&\sim\; -b_{n,k}\,\frac{\rho}{2\,(1-\rho)\,(2-\rho)}\;, \end{aligned}$$

and hence

$$\frac{T_1}{T_2} \;\sim\; -\frac{2\,(2-\rho)}{1-\rho}\;,$$

which can be solved for ρ, yielding

$$\hat{\rho}_k \;=\; \frac{4T_2+T_1}{2T_2+T_1}\;. \tag{3.4}$$

The asymptotic properties of this estimator will be discussed elsewhere.

As an alternative goodness-of-fit based procedure, the optimal k could be derived from comparing observed and fitted values on the Pareto quantile plot, for instance minimizing a weighted Cramér–von Mises statistic

$$\frac{1}{H^2_{k,n}}\,\frac{1}{k}\sum_{j=1}^{k}\frac{j}{k-j+1}\left(\log\frac{X_{n-j+1,n}}{X_{n-k,n}} + H_{k,n}\log\frac{j}{k+1}\right)^2\;. \tag{3.5}$$

Criteria of this type were considered in, for instance, Beirlant *et al.* (1996), and Dupuis and Victoria-Feser (2003). Unlike the goodness-of-fit based threshold selection procedure described above, this prediction error criterion does not require the estimation of the nuisance parameter ρ, but even asymptotically it will not minimize the AMSE.

The Lewis based AMSE criterion and the prediction error criterion will now be compared on the basis of a small sample simulation study. For the Lewis based AMSE criterion we consider three cases: ρ fixed at -1, correct specification of ρ and the case where ρ is replaced by (3.4). We simulated 500 samples of size $n = 500$ from the Burr(η, τ, λ) distribution, with distribution function given by

$$F(x) \,=\, 1 - \left(\frac{\eta}{\eta + x^\tau}\right)^\lambda, \qquad x > 0, \quad \eta, \tau, \lambda > 0 \;,$$

for which $\gamma = 1/(\lambda\tau)$ and $\rho = -1/\lambda$. In Table 1, we summarize the results of the simulation study by the empirical mean squared errors (MSE) of $H_{\hat{k}_{\text{opt}},n}$. For both procedures considered, the γ estimates deteriorate with increasing values of ρ. Clearly, the Lewis based AMSE approximation outperforms the prediction error criterion. Moreover, the gains in MSE tend to increase in ρ. Note that, although the Lewis based approximation of the AMSE requires an estimate for ρ, the results are quite insensitive with respect to the specification of ρ.

Table 1: Empirical MSE of $H_{\hat{k}_{\text{opt}},n}$.

Distribution	γ	ρ	Lewis based AMSE criterion			Prediction error criterion
			$\rho = -1$	correct ρ	$\hat{\rho}$	
Burr$(1, 2, 0.5)$	1	-2	0.0103	0.0100	0.0109	0.0109
Burr$(1, 1, 1)$	1	-1	0.0275	0.0275	0.0288	0.0359
Burr$(1, 0.5, 2)$	1	-0.5	0.1178	0.1018	0.1199	0.1996
Burr$(1, 0.25, 4)$	1	-0.25	0.6869	0.5156	0.7195	1.1239
Burr$(1, 4, 0.5)$	0.5	-2	0.0029	0.0025	0.0030	0.0027
Burr$(1, 2, 1)$	0.5	-1	0.0069	0.0069	0.0072	0.0089
Burr$(1, 1, 2)$	0.5	-0.5	0.0299	0.0271	0.0308	0.0464
Burr$(1, 0.5, 4)$	0.5	-0.25	0.1771	0.1316	0.1741	0.2756

Besides this prediction error criterion we will also compare our goodness-of-fit based approach with some other criteria recently proposed. The computational complexity of some of these is such that they are not easy to implement for comparison purposes. Beirlant *et al.* (2002) performed an extensive simulation study and we will refer to some of their results, along with those from Beirlant *et al.* (1996) and Matthys and Beirlant (2003).

To summarize, the procedures that will be compared are:

- Method 1: the Lewis based criterion given by (3.3),
- Method 2: the prediction error criterion given by (3.5),
- Method 3: Beirlant *et al.* (2002),

- Method 4: Danielsson *et al.* (2001),
- Method 5: Drees and Kaufmann (1998),
- Method 6: Guillou and Hall (2001).

The performance of these procedures is evaluated on the basis of a small sample simulation study. In this simulation we use, next to the Burr(η, τ, λ) distribution introduced above, the following distributions:

1. The Fréchet(α) distribution,

$$F(x) = \exp(-x^{-\alpha}) , \qquad x > 0, \quad \alpha > 0 ,$$

with $\gamma = 1/\alpha$ and $\rho = -1$. We set $\alpha = 2$.

2. The $|T_\nu|$ distribution,

$$F(x) = \int_0^x \frac{2\,\Gamma\big(\frac{\nu+1}{2}\big)}{\sqrt{\nu\pi}\,\Gamma\big(\frac{\nu}{2}\big)} \left(1 + \frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}} dy , \qquad x > 0, \quad \nu > 0 ,$$

with $\gamma = 1/\nu$ and $\rho = -2/\nu$. We took $\nu = 6$.

3. The loggamma(λ, α) distribution,

$$F(x) = \int_1^x \frac{\lambda^\alpha}{\Gamma(\alpha)} \, (\log y)^{\alpha-1}\, y^{-\lambda-1}\, dy , \qquad x > 1, \quad \lambda, \alpha > 0 ,$$

with $\gamma = 1/\lambda$ and $\rho = 0$. We set $\lambda = 1$ and $\alpha = 2$.

For each of the above models, 500 datasets of size $n = 500$ are simulated. The results of the simulation are summarized in Table 2 where we show the empirical mean squared error of $H_{\hat{k}_{\text{opt}},n}$ for the different methods and distributions considered. As is clear from Table 2 no single criterion performs uniformly best. The Lewis based approximation is clearly competitive and maintains itself in the first half of the methods considered.

Table 2: Empirical MSE of $H_{\hat{k}_{\text{opt}},n}$.

Method	Fréchet(2)	Burr$(1, 0.5, 2)$	$\lvert T_6\rvert$	loggamma$(1, 2)$
1	0.0047	0.1178	0.0148	0.0873
2	0.0054	0.1996	0.0242	0.1105
3	0.0052	0.0930	0.0110	0.0602
4	0.0109	0.1459	0.0176	0.0904
5	0.0041	0.1239	0.0129	0.0784
6	0.0049	0.1452	0.0190	0.0689

4. Case study: diamond data

Our case study can be situated in a geostatistical context and concerns the valuation of diamonds. The profitability of a mining exploration largely depends on the occurrence of precious stones, and consequently, accurate modeling of the tail of the diamond value distribution is of crucial importance. The data set considered here contains the value (in USD) of a sample of 1914 diamonds obtained from a kimberlite deposit. These data are publicly available at `http://ucs.kuleuven.be/Wiley/Data/diamond.txt`. Figure 1 (a) shows the exponential quantile plot for the variable value; Figure 1 (b) is the corresponding mean excess plot. The convex shape of the exponential quantile plot and the means excess function that is decreasing when considered as a function of k indicate sub-exponential tail behavior. To assess the hypothesis of Pareto-type behavior we also construct the Pareto quantile plot, see Figure 1 (c). The Pareto quantile plot is clearly approximately linear in the largest observations indicating a good fit of the value distribution by a Pareto-type model. The mean excess function of the log-transformed data, which is in fact the Hill estimator, given in Figure 1 (d), confirms this in the sense that it clearly shows a constant slope at the smaller $\log k$ values.

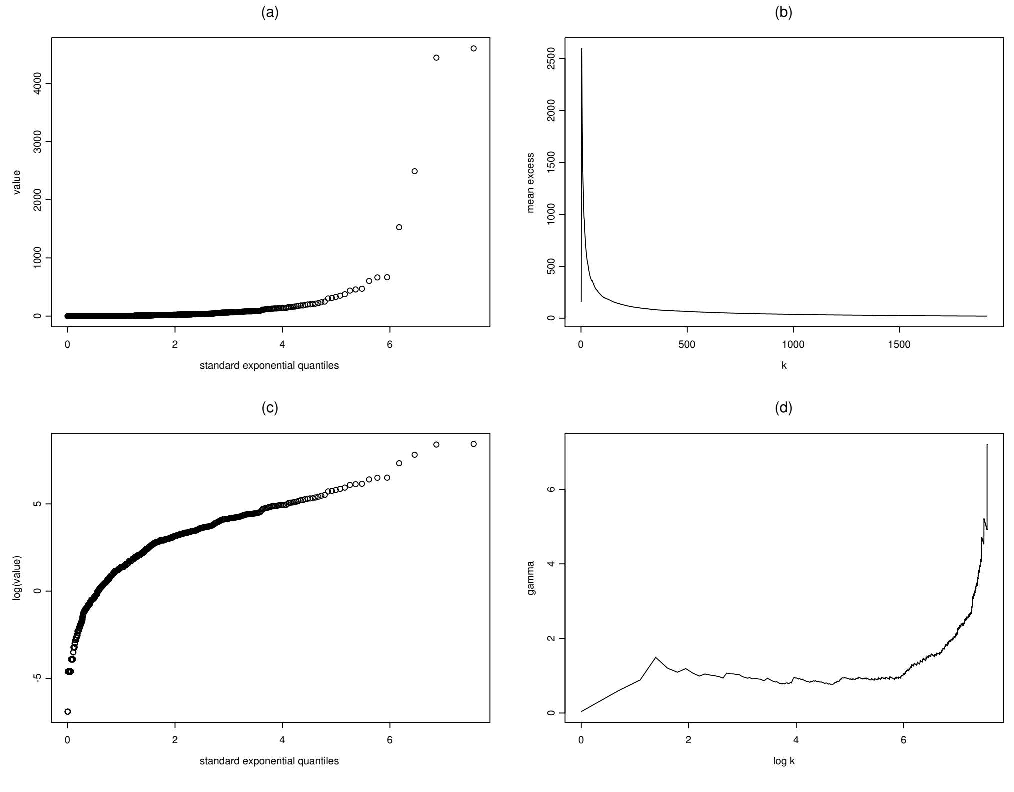

Figure 1: Diamond data: **(a)** exponential quantile plot, **(b)** mean excess plot, **(c)** Pareto quantile plot and **(d)** $H_{k,n}$ as a function of $\log k$.

In Figure 2 we show the four goodness-of-fit statistics together with the critical values of pointwise 5% hypothesis tests, as derived from the limiting distributions of the test statistics. For the ease of comparison we show the statistics in standardized format, i.e. we show $T^{\mathrm{J}}_{k,n}$, $\sqrt{12}\;T^{\mathrm{L}}_{k,n}$, $(1-\hat\rho)/|\hat\rho|\,T^{\mathrm{BCJ}}_{k,n}$ and $(2-\hat\rho)/|1+\hat\rho|\,\sqrt{12}\;T^{\mathrm{BCL}}_{k,n}$, where the scaling factors follow from the asymptotic variance expression $\int_0^1 K^2(u)\,du$, and where $T^{\mathrm{BCJ}}_{k,n}$ and $T^{\mathrm{BCL}}_{k,n}$ denote the bias-corrected Jackson and Lewis statistic, respectively, obtained by plugging K_{J} and K_{L} in (2.7). Globally, up to approximately $k = 380$, all statistics fail to reject H_0 of Pareto-type behavior with $\mathcal{R}_\ell$ on ℓ_U. The bias-corrected Lewis statistic shows two exceptions to this overall pattern, namely at the positions $k = 53$ and $k = 128$. These positions are indicated on the Pareto quantile plot given in Figure 3.

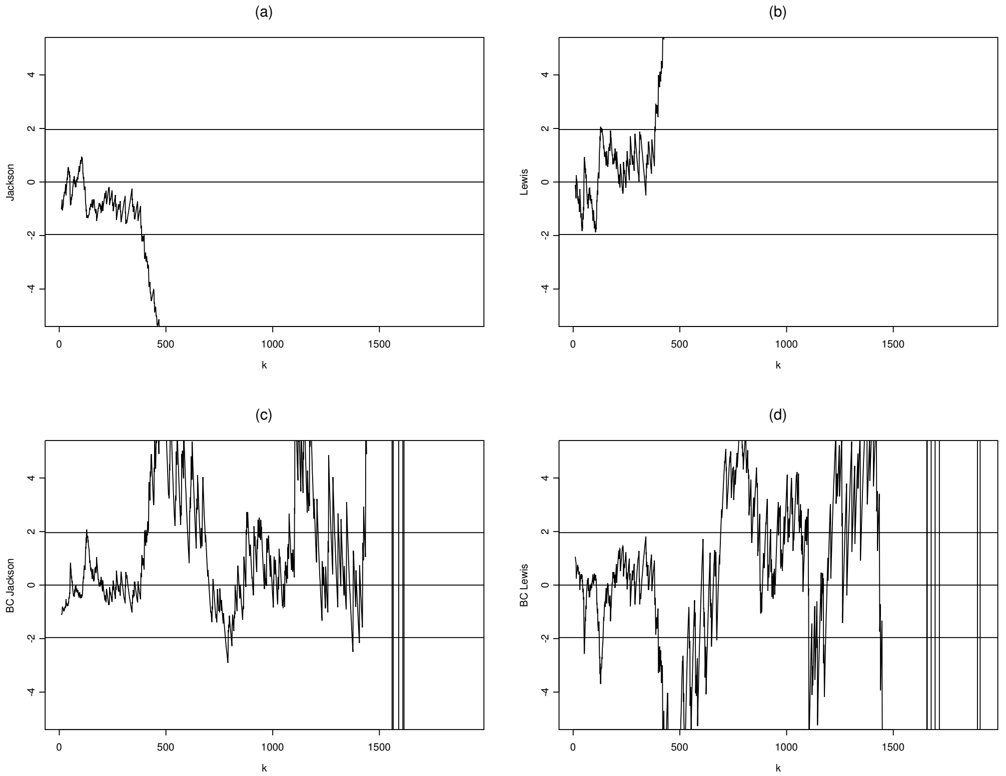

Figure 2: Diamond data: **(a)** $T^{\mathrm{J}}_{k,n}$, **(b)** $\sqrt{12}\,T^{\mathrm{L}}_{k,n}$, **(c)** $(1-\hat\rho)/|\hat\rho|\;T^{\mathrm{BCJ}}_{k,n}$, **(d)** $(2-\hat\rho)/|1+\hat\rho|\,\sqrt{12}\,T^{\mathrm{BCL}}_{k,n}$ as a function of k.

Clearly, at these positions the Pareto quantile plot makes vertical jumps. Beyond $k = 380$ the uncorrected statistics diverge and move outside the acceptance region, while the bias-corrected statistics fluctuate heavily and show portions of reasonable length both inside and outside the acceptance region, and hence give a more nuanced picture of the distributional behavior. A plausible explanation for this pattern can be found in the Pareto quantile plot. The Pareto quantile plot shows more or less linear segments on both the left- and right-hand side of the observation $k = 380$, although with different slopes. The uncorrected statistics can only handle the ultimate linear part of this plot and hence beyond this point they diverge. The bias-corrected statistics, through the inclusion of the second order tail condition, are also able to deal with the curved part. However, the portions inside and outside the acceptance region indicate special features of these data. Looking back at the Pareto quantile plot we find that also deeper in the data, i.e. at larger k-values, other linear portions with different slopes can be distinguished. This may indicate that the diamond value distribution is a mixture of several Pareto-type models with different Pareto indices. In fact, the tail of the diamond value distribution is known to be influenced by factors such as, among others, size and color (Beirlant and Goegebeur, 2003). In this analysis we ignored this information. It is a nice feature of the bias-corrected statistics that they indicate this change in distributional regime and give, compared to the uncorrected statistics, a more subtle view on the tail behavior.

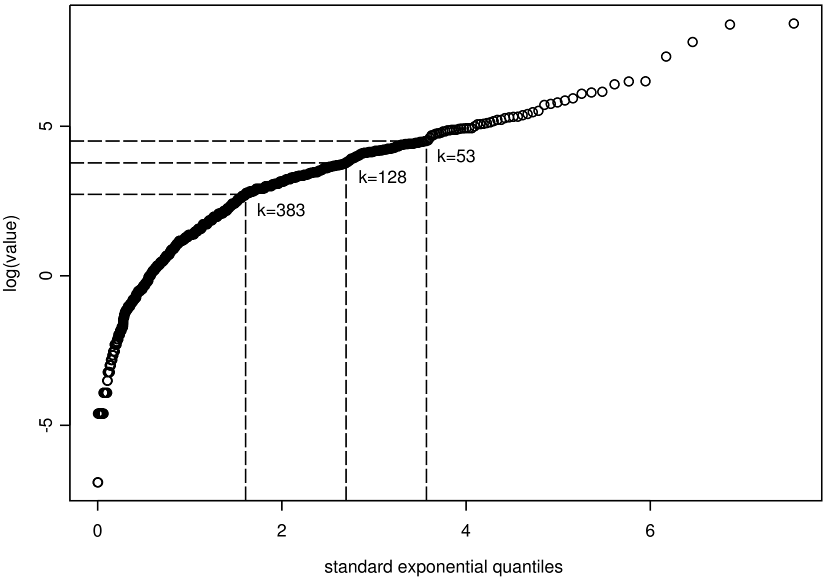

Figure 3: Diamond data: Pareto quantile plot with the positions where H_0 of Pareto-type behavior is rejected.

Given that the Pareto-type model provides a plausible explanation of these data, the analysis can be carried one step further, focusing on the estimation of the tail index γ. In this respect, Figure 4 (a) shows the Lewis based approximation to the asymptotic mean squared error of the Hill estimator, $\widehat{AMSE}(H_{k,n})$, obtained with $\hat{\rho} = -2.724$, as a function of k. This ρ-value is obtained from (3.4) together with the rule of thumb proposed by Gomes *et al.* (2002) that the k for the estimation of ρ can be taken as $k = \lfloor n^{0.995} \rfloor$, see also Figure 4 (e). The minimum value of $\widehat{AMSE}(H_{k,n})$ is reached at $\hat{k}_{\text{opt}} = 343$ and $H_{343,1914} = 0.917$. Note that

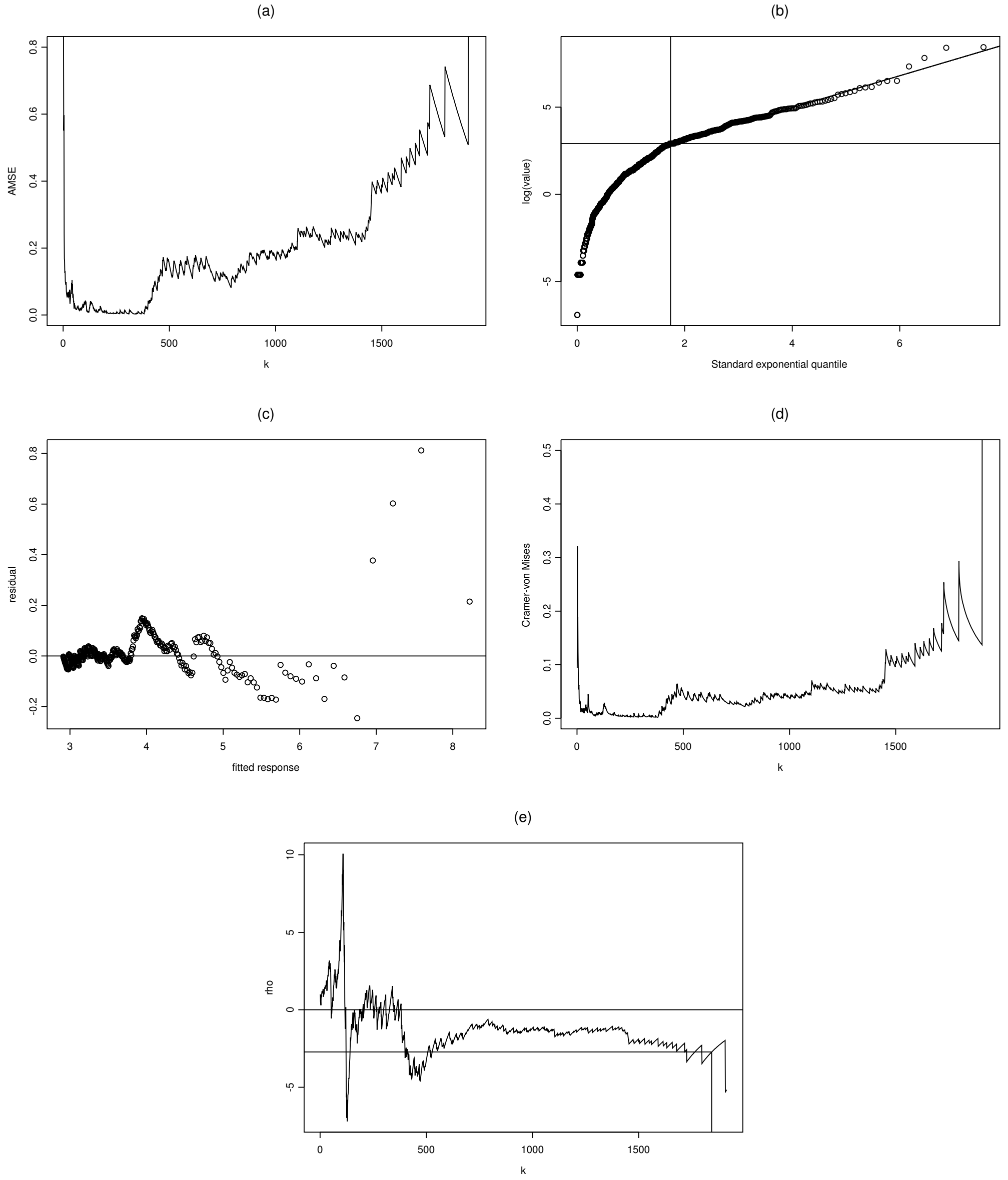

Figure 4: Diamond data: **(a)** $\widehat{AMSE}(H_{k,n})$ as a function of k, **(b)** Pareto quantile plot, **(c)** residuals versus fitted responses, **(d)** prediction error criterion, **(e)** $\hat{\rho}_k$ as a function of k.

the k-value minimizing the asymptotic mean squared error is smaller than the k-value beyond which goodness-of-fit tests consistently reject the null hypothesis given in (1.6). Alternatively, based on Figure 4 (e), we could also have taken $\hat{\rho} = -1$, which would result in $\hat{k}_{\text{opt}} = 336$ and $H_{336,1914} = 0.912$, results that are in line with those obtained with the former ρ-value. In Figure 4 (b) we indicate the 343th largest observation on the Pareto quantile plot of the variable value together with a straight line through this point and with slope $H_{\hat{k}_{\text{opt}},n}$. Clearly, the straight line summarizes the upper right portion of the Pareto quantile plot quite well, see also the Figure 4 (c) showing the residuals resulting from this line fit. Finally, in Figure 4 (d), we show the prediction error criterion (3.5) as a function of k. The prediction error reaches its minimum deeper in the data, at $k = 379$ and $H_{379,1914} = 0.940$, results that are comparable with the minimization of the asymptotic mean squared error.

5. CONCLUSION

In this paper we examined the relationship between Pareto-type goodness-of-fit testing and the selection of the upper sample fraction when estimating the tail index, for instance using Hill's estimator. To this end we considered the class of kernel statistics introduced in Goegebeur *et al.* (2007). Typically, goodness-of-fit tests are too conservative with respect to the null hypothesis, entailing too high k-values (or too small thresholds) relative to the minimum AMSE criterion, which led us to follow another route, exploiting the relationship between the kernel statistic and the bias component of the AMSE of the Hill estimator. The procedure was evaluated on a small sample simulation study and showed to be competitive with some of the better performing currently available algorithms. As a nice side result, we obtained a new estimator for the second order parameter ρ, of which the in-depth investigation is a topic of current research.

ACKNOWLEDGMENTS

Yuri Goegebeur's research was supported by a grant from the Danish Natural Science Research Council. Tertius de Wet's research was supported by National Research Foundation grant 2053332 and a research grant from the University of Stellenbosch.

REFERENCES

[1] Beirlant, J.; de Wet, T. and Goegebeur, Y. (2006). A goodness-of-fit statistic for Pareto-type behaviour, *Journal of Computational and Applied Mathematics*, **186**, 99–116.

[2] Beirlant, J.; Dierckx, G.; Goegebeur, Y. and Matthys, G. (1999). Tail index estimation and an exponential regression model, *Extremes*, **2**, 177–200.

[3] Beirlant, J.; Dierckx, G.; Guillou, A. and Stărică, C. (2002). On exponential representations of log-spacings of extreme order statistics, *Extremes*, **5**, 157–180.

[4] Beirlant, J. and Goegebeur, Y. (2003). Regression with response distributions of Pareto-type, *Computational Statistics and Data Analysis*, **42**, 595–619.

[5] Beirlant, J.; Goegebeur, Y.; Segers, J. and Teugels, J. (2004). *Statistics of Extremes – Theory and Applications*, Wiley Series in Probability and Statistics.

[6] Beirlant, J.; Vynckier, P. and Teugels, J.L. (1996). Tail index estimation, Pareto quantile plots, and regression diagnostics, *Journal of the American Statistical Association*, **91**, 1659–1667.

[7] Bingham, N.H.; Goldie, C.M. and Teugels, J.L. (1987). *Regular Variation*, Cambridge University Press, Cambridge.

[8] Danielsson, J.; de Haan, L.; Peng, L. and de Vries, C. (2001). Using a bootstrap method to choose sample fraction in tail index estimation, *Journal of Multivariate analysis*, **76**, 226–248.

[9] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*, Springer.

[10] de Wet, T. and Venter, J.H. (1973). A goodness-of-fit test for a scale parameter family of distributions, *South African Statistical Journal*, **7**, 35–46.

[11] Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation, *Stochastic Processes and their Applications*, **75**, 149–172.

[12] Dupuis, D. and Victoria-Feser, M.-P. (2003). *A prediction error criterion for choosing the lower quantile in Pareto-index estimation*, Cahiers de Recherche HEC no. 2003.19, University of Geneva.

[13] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire, *Annals of Mathematics*, **44**, 423–453.

[14] Goegebeur, Y.; Beirlant, J. and de Wet, T. (2007). *Kernel goodness-of-fit statistics for Pareto-type behavior*, submitted.

[15] Gomes, M.I.; de Haan, L. and Peng, L. (2002). Semi-parametric estimators of the second order parameter in statistics of extremes, *Extremes*, **5**, 387–414.

[16] Guillou, A. and Hall, P. (2001). A diagnostic for selecting the threshold in extreme value analysis, *Journal of the Royal Statistical Society B*, **63**, 293–305.

[17] Henze, N. and Meintanis, S.G. (2005). Recent and classical tests for exponentiality: a partial review with comparisons, *Metrika*, **61**, 29–45.

[18] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1163–1174.

[19] Jackson, O.A.Y. (1967). An analysis of departures from the exponential distribution, *Journal of the Royal Statistical Society B*, **29**, 540–549.

[20] Lewis, P.A.W. (1965). Some results on tests for Poisson processes, *Biometrika*, **52**, 67–77.

[21] Matthys, G. and Beirlant, J. (2003). Estimating the extreme value index and high quantiles with exponential regression models, *Statistica Sinica*, **13**, 853–880.

[22] Stephens, M.A. (1986). *Tests for the exponential distribution.* In "Goodness-Of-Fit Techniques" (R.B. D'Agostino and M.A. Stephens, Eds.), Marcel Dekker Inc., 421–459.