Statistica Sinica **20** (2010), 1257-1272

A THRESHOLD APPROACH FOR PEAKS-OVER-THRESHOLD MODELING USING MAXIMUM PRODUCT OF SPACINGS

Tony Siu Tung Wong and Wai Keung Li

The University of Hong Kong

Abstract: We propose a threshold model extending the generalized Pareto distribution for exceedances over a threshold. The threshold is solely determined within the model and is shown to be super-consistent under the maximum product of spacings estimation method. We apply the model to some insurance data and demonstrate the merit of having a full parametric model for the entire data set.

Key words and phrases: Exceedances, generalized Pareto distribution, maximum product of spacings, super-consistency, threshold model.

1. Introduction

Pickands (1975) demonstrated that the conditional survival distribution of exceedances (or peaks, or excesses) $X-u$ over a sufficiently high threshold u, given $X \geq u$, is a generalized Pareto distribution (GPD)

$$1-G_u\left(x;\gamma,\sigma\right)=\begin{cases}\left\{1+\frac{\gamma(x-u)}{\sigma}\right\}^{-1/\gamma}, & x\in(u,\infty) \quad \text{if} \quad \gamma>0,\\ \exp\left\{-\frac{(x-u)}{\sigma}\right\}, & x\in(u,\infty) \quad \text{if} \quad \gamma=0,\\ \left\{1+\frac{\gamma(x-u)}{\sigma}\right\}^{-1/\gamma}, & x\in(u,u-\frac{\sigma}{\gamma}) \text{ if} \quad \gamma<0.\end{cases} \tag{1.1}$$

The parameter γ, termed the extreme value index (EVI), is a key quantity in the literature of extreme value analysis. Its sign is the dominant factor in describing the tail of the underlying distribution $F(x)$.

In order to work out the relevant estimators $\widehat{\gamma}$ and $\widehat{\sigma}$, an input of u is needed, and this choice is very much an open matter. In the literature, not much attention has been given to this aspect. It is possible to choose an optimal u by the quantification of a bias versus variance trade-off. As in the case of the Hill estimator (Hill (1975)), choosing an optimal threshold is similar to choosing the number of upper order statistics; a compromise between bias and variance has to be reached. Davison and Smith (1990) proposed the use of a mean excess plot based on the linearity of the mean excess function for the GPD. See also

Embrechts, Klüppelberg, and Mikosch (1997), Beirlant et al. (2004), Castillo et al. (2005), and de Haan and Ferreira (2006) for extensive discussions about extreme value analysis. In this paper, we propose a threshold method that allows for different probability models in different portions of the sample space. The threshold (or the change point) of the model becomes one of the unknown parameters. There have been remarkable successes of the threshold method in the application of time series analysis, and in other fields. See for example, Tong (1978) and Tong and Lim (1980).

Given a value of u, the estimation of the GPD parameters can be performed in a variety of ways. A popular method is maximum likelihood. Maximum likelihood estimators are consistent if $\gamma > -1/2$, but the log-likelihood function is unbounded. Maximizing the log-likelihood function with respect to the parameters involves the term $-\left(1+1/\gamma\right)\log\left[1+\gamma\left(x-u\right)/\sigma\right]$. As $x \downarrow u-\sigma/\gamma$, the log-likelihood function approaches positive infinity when $\gamma < -1$. An alternative to maximum likelihood is the maximum product of spacings (MPS) method introduced by Cheng and Amin (1983). The objective function of the MPS method is bounded from above by $-\left(k+1\right)\log\left(k+1\right)$, where k is the number of exceedances above u. Cheng and Stephens (1989) proved that, under regularity conditions, the MPS estimators have an asymptotic normal distribution and differ from the maximum likelihood estimators by $o_p\left(n^{-1/2}\right)$. Comparisons between the two methods on the inference of GPD parameters can be found in Fitzgerald (1996) and Wong and Li (2006).

The paper is organized as follows. Section 2 describes the threshold model and presents its asymptotic properties. It is shown that the threshold estimate is $n-$consistent and that the GPD parameter estimates are $\sqrt{k}-$consistent. Section 3 gives a summary of the methods of Guillou and Hall (2001) and Beirlant, Joossens, and Segers (2004). Simulation studies are reported in Section 4. Finally two examples are presented in Section 5. Section 6 gives a concluding remark.

2. The Threshold Model

The events $X \le u$ and $X > u$ on the real line partition the sample space according to the threshold u. The distribution function $P\left(X \le x\right)$ can be written as

$$P\left(X \le x \cap X \le u\right) + P\left(X > u\right) P\left(X \le x \left| X > u\right.\right).$$

As $u \to \infty$, the term involving a conditional probability can be approximated by (1.1). We model the left-hand side of the sample space defined by u by a truncated distribution function L with parameter $\theta \in \mathbb{R}^p$. This leads to the threshold model

$$F\left(x;\theta,\gamma,\sigma\right) = \begin{cases} L\left(x;\theta\right), & x \le u, \\ L\left(u;\theta\right) + \left(1-L\left(u;\theta\right)\right) G_u\left(x;\gamma,\sigma\right), & x > u. \end{cases} \tag{2.1}$$

The traditional approach concentrates on the k upper order statistics for any fixed k. Suppose that $n\left(1-F\left(u_n\right)\right)\rightarrow\tau$ holds for $0<\tau<\infty$. Here u_n typically becomes higher with n. If the GPD is valid for the k excesses over the threshold u_0, it should be equally valid for all thresholds $u>u_0$ subject to an appropriate change of σ. The major drawbacks of this approach are that much information is ignored as the sample size increases, and the whereabouts of the true threshold is always ambiguous. In contrast, (2.1) assumes a fixed large threshold such that k can tend to ∞ at a rate slower than n as $n\rightarrow\infty$. We can see that the GPD is also valid in this case. For a pair of sequences a_n and b_n with $a_n>0$ and a continuous distribution function $\Lambda\left(x\right)$, Pickands (1975) showed that if

$$\lim_{n\rightarrow\infty}\frac{\left[1-F\left(a_nx+b_n\right)\right]}{1-F\left(b_n\right)}=\frac{\log\Lambda\left(x\right)}{\log\Lambda\left(0\right)} \tag{2.2}$$

holds, the right-hand side is the GPD. Note that $\Lambda\left(x\right)$ is an extremal distribution function. Smith (1987) argued that the limiting results in the present context are usually conditional on both k and u, and that they can be interpreted as unconditional results when either k or u is treated as fixed and the other random, depending on n. As in Smith (1987), we adopt the view that u is fixed in such a way that as $n\rightarrow\infty$, $n\left(1-F\left(u\right)\right)\rightarrow\infty$ and $k^{-1}n\left(1-F\left(u\right)\right)\rightarrow_p 1$.

Our approach has several advantages. First, the full data set is used so that there is no loss of information. The model provides a global fit and also an appropriate tail fit. Second, the determination of the threshold is automatically data-driven. In particular, the estimate of u differs from the true parameter by an amount which is of order n^{-1}. Lastly, the model can provide better insight into the structure of the data. The value of a model is greatly determined by its ability to predict the future. Extrapolation beyond the data based on n observations is more persuasive than on k excesses in the traditional approach. The threshold value also has an interesting interpretation. In the insurance context, a high-excess loss layer with an attachment point u is of interest; a payout on the loss $X-u$ is related to an actuarial pricing problem. Thus, estimation of the threshold u is of both practical and methodological importance.

Denote by $(\theta_0,\gamma_0,\sigma_0,u_0)$ and $(\tilde{\theta},\tilde{\gamma},\tilde{\sigma},\tilde{u})$ the true parameter and the MPS estimator of (θ,γ,σ,u), respectively. Given an estimate of the threshold $\tilde{u}$, $(\tilde{\theta},\tilde{\gamma},\tilde{\sigma})$ is found by maximizing the objective function

$$M\left(\theta,\gamma,\sigma\right)=\sum\nolimits_{i=1}^{n+1}\log\left(F\left(x_{(i)}\right)-F\left(x_{(i-1)}\right)\right), \tag{2.3}$$

where $F\left(x_{(0)}\right)=0$, $F\left(x_{(n+1)}\right)=1$, and $x_{(1)}\leq\cdots\leq x_{(n)}$ are the ordered realizations of the sample. If $x_{(j)}=x_{(j-1)}$, $j=2,\ldots,n$, we replace the quantity $F\left(x_{(i)}\right)-F\left(x_{(i-1)}\right)$ by the density function $f\left(x_{(i)}\right)$, as in Cheng and Amin

(1983). The estimate of the threshold $\tilde{u}$ is obtained by choosing $x_{(i)}$ successively, as in Tong and Lim (1980), as possible candidates and picking the one for which the process (2.3) yields the maximum value. Pickands (1975) chose k from 1 to $[n/4]$ where the empirical upper tail is closest to the GPD. We adopt a similar approach. Note that when $u = x_{(n)}$, the entire sample is fitted with L.

We show the super-consistency of $\tilde{u}$ and the large sample distribution of other model parameter estimates. This result implies that statistical inference on the other parameters of model (2.1) can be conducted as if u_0 is known.

Theorem 2.1. *Under certain very general regularity conditions, $\widetilde{u}$ is super consistent, with $\widetilde{u} - u_0 = O_p\left(n^{-1}\right)$.*

The proof of Theorem 2.1 and the regularity conditions are given in the Appendix.

Theorem 2.2. *If $\gamma > -1/2$, the MPS estimator $(\widetilde{\gamma}, \widetilde{\sigma})$ is asymptotically normal with*

$$\left[\sqrt{k}\left(\tilde{\gamma} - \gamma_0\right), \sqrt{k}\left(\tilde{\sigma} - \sigma_0\right)\right] \xrightarrow{D} N\left(\mathbf{0}, \mathbf{V}\right),$$

$$\mathbf{V} = (1+\gamma)\begin{pmatrix} 1+\gamma & -\sigma \\ -\sigma & 2\sigma^2 \end{pmatrix}.$$

Proof of Theorem 2.2. The maximum product of spacings estimator has an asymptotic normal distribution with variance given by $k^{-1}\mathbf{V}$ where $\mathbf{V}$ is the inverse of the Fisher information matrix; see Theorem 1 in Cheng and Amin (1983). The log-likelihood function of the threshold model is

$$h\left(\theta, \gamma, \sigma\right) = \sum_{i=1}^{n-k} \log l\left(x_{(i)}; \theta\right) + \sum_{i=n-k+1}^{n} \log\left\{1 - L\left(u; \theta\right)\right\} + \sum_{i=n-k+1}^{n} \log g_u\left(x_{(i)}; \gamma, \sigma\right).$$

Since the first two terms on the right-hand side are independent of (γ, σ), the variance-covariance matrix is the same as that of the GPD. The closed form of $\mathbf{V}$ can be obtained from, for example, Beirlant et al. (2004).

3. Competing Methods

To study the performance of the threshold model, we consider two competing methods.

Guillou and Hall (2001) suggested an easily computed diagnostic for choosing the threshold when the Hill estimator is used to estimate the tail exponent. The procedure can be considered as an asymptotic test for the hypothesis of zero bias. The value of k is the least integer such that the mean of the bias significantly differs from zero. The authors found by numerical simulation that the optimal choice of the critical value c_{crit} for the test occurs at a value between 1.25 and 1.5. Unless otherwise specified, we follow the same procedure as in Section 3 of Guillou and Hall (2001) and use $c_{crit} = 1.25$.

Beirlant, Joossens, and Segers (2004) proposed an extension of the GPD with a single parameter by a second-order refinement of the extreme value theory. The model is

$$F\left(x\right) = 1 - \left[\frac{\gamma}{\sigma}x - \left(\frac{\gamma}{\sigma}u - 1\right)\left(\frac{x}{u}\right)^{\rho+1}\right]^{-1/\gamma},$$

with $\gamma \geq \sigma u^{-1} \max\left(0, 1 + \rho^{-1}\right)$. All the parameters are in parallel with our threshold model except that $\rho < 0$ is an additional parameter. Note that the special case $\rho = -1$ gives the GPD. The model fits the SOA Group Medical Insurance data of 1991 well, even for the lowest possible threshold. Hence, we consider it as a potential candidate for the entire data set.

4. Numerical Simulation

To examine the finite sample properties of our model, we undertook a simulation experiment. We used independent and identically distributed samples of sizes $n = 250, 500$ and replicated them 1,000 times independently. Samples were drawn from (2.1) with L being one of the following:

(a) a Weibull (a, b, c) distribution, $F\left(x\right) = 1 - \exp\left\{-\left[(x-b)/a\right]^c\right\}$;

(b) an exponential distribution with parameter λ, $F\left(x\right) = 1 - \exp\left(-\lambda x\right)$;

(c) a gamma (a, b, c) distribution, $F\left(x\right) = \int_b^x \left(t-b\right)^{c-1} \exp\left[-(t-b)/a\right]/(a^c\Gamma\left(c\right))$ dt;

(d) a Normal distribution with mean μ and variance β;

(e) a Student's $t-$distribution with degrees of freedom v;

(f) a Burr (a, b, c) distribution(type XII), $F\left(x\right) = 1 - \left(b/(b + x^c)\right)^a$;

(g) a Burr (a, b, c) distribution(type III), $F\left(x\right) = \left(b/(b + x^{-c})\right)^a$.

In each case, the distribution function is truncated at $u = \inf\left\{x : F\left(x\right) \geq p\right\}$ for p sufficiently large. For brevity, we present our results only in cases where p is 0.9 and the GPD parameters γ and σ are 0.4 and 5.0, respectively. We experimented with different values of n, γ, and p. Overall they were not significantly different.

Our results are summarized in Table 4.1. There, the mechanism generating the data was model (2.1), with L being one of the distributions above. The second and third columns give average values of $\widetilde{u}$ and $\widetilde{\gamma}$, respectively, given that L is known. The next column gives the average values of $\widetilde{u}_{GH}$ by the adaptive threshold selection method (Guillou and Hall (2001)). The fourth and fifth columns, respectively, give averages of the Hill estimator $\widetilde{\gamma}_{GH,Hill}$ and the

Table 4.1. Average values of the estimates for the seven models. $\widetilde{u}$ and $\widetilde{\gamma}$ are the MPS estimates of our models; $\widetilde{u}_{GH}$ is the threshold estimate by Guillou and Hall's method; $\widetilde{\gamma}_{GH,Hill}$ is the Hill estimator; $\widetilde{\gamma}_{GH,GPD}$ is the EVI estimator of the GPD given $\widetilde{u}_{GH}$; $\widetilde{\gamma}_{BJS}$ is the estimator by Beirlant, Joossens, and Segers's method. Standard errors are shown in brackets.

n	$\widetilde{u}$	$\widetilde{\gamma}$	$\widetilde{u}_{GH}$	$\widetilde{\gamma}_{GH,Hill}$	$\widetilde{\gamma}_{GH,GPD}$	$\widetilde{\gamma}_{BJS}$
(a) Weibull $(1.0, 0.0, 5.0)$ distribution with $u_0 = 1.18$						
250	1.18	0.53	7.69	0.76	0.83	0.44
	(0.01)	(0.37)	(6.23)	(0.33)	(8.92)	(0.08)
500	1.18	0.48	11.59	0.64	0.31	0.43
	(<0.01)	(0.24)	(8.60)	(0.26)	(2.40)	(0.06)
(b) Exponential distribution $(\lambda = 1.0)$ with $u_0 = 2.30$						
250	2.20	0.55	7.81	0.68	0.87	0.50
	(0.23)	(0.46)	(6.79)	(0.26)	(8.18)	(0.09)
500	2.27	0.47	12.00	0.60	0.31	0.50
	(0.11)	(0.25)	(8.99)	(0.24)	(2.32)	(0.07)
(c) Gamma $(1.0, 0.0, 5.0)$ distribution with $u_0 = 7.99$						
250	7.64	0.54	11.40	0.45	0.76	0.11
	(0.63)	(0.46)	(7.18)	(0.16)	(6.47)	(0.08)
500	7.83	0.47	14.98	0.45	0.36	0.10
	(0.43)	(0.26)	(10.01)	(0.16)	(2.14)	(0.06)
(d) Normal distribution $(\mu, \beta) = (10.0, 1.0)$ with $u_0 = 11.28$						
250	11.25	0.53	15.12	0.40	0.74	0.14
	(0.09)	(0.43)	(6.68)	(0.15)	(7.45)	(0.05)
500	11.27	0.48	18.31	0.40	0.37	0.13
	(0.04)	(0.24)	(9.73)	(0.14)	(2.21)	(0.04)
(e) Student's $t-$distribution $v = 5.0$ with $u_0 = 1.48$						
250	1.42	0.54	15.28	0.40	0.69	0.07
	(0.15)	(0.44)	(6.69)	(0.14)	(6.77)	(0.07)
500	1.46	0.48	18.47	0.40	0.35	0.05
	(0.07)	(0.25)	(9.74)	(0.14)	(2.16)	(0.08)
(f) Burr $(1.0, 1.0, 5.0)$ distribution(type XII) with $u_0 = 1.55$						
250	1.54	0.53	7.77	0.74	0.86	0.45
	(0.04)	(0.39)	(6.39)	(0.31)	(8.71)	(0.08)
500	1.55	0.48	11.74	0.63	0.30	0.43
	(0.01)	(0.24)	(8.72)	(0.26)	(2.38)	(0.06)
(g) Burr $(1.0, 1.0, 5.0)$ distribution(type III) with $u_0 = 1.55$						
250	1.53	0.63	7.77	0.74	0.86	0.45
	(0.07)	(0.56)	(6.39)	(0.31)	(8.71)	(0.08)
500	1.54	0.51	11.74	0.63	0.30	0.43
	(0.04)	(0.30)	(8.72)	(0.26)	(2.38)	(0.06)

EVI estimator $\widetilde{\gamma}_{GH,GPD}$ of the GPD given $\widetilde{u}_{GH}$. The last column gives average values of $\widetilde{\gamma}_{BJS}$ (Beirlant, Joossens, and Segers (2004)). It is clear from the information that $\widetilde{u}$ was very accurate. The ratio of the root mean squared error of $\widetilde{u}$ in the case of $n = 250$ to that of $n = 500$ took values between 1.45 and 2.76. These values have an average of 2.14 which is very close to the ratio of the sample size. This reflects the fact that the order of convergence is $O(n)$ instead of the usual $O(\sqrt{n})$. Though Guillou and Hall's method overestimated the threshold, good performance was obtained in some instances. In eight cases, $\widetilde{\gamma}_{GH,Hill}$ gave unfavourable results. This may be due to the fact that a strict Pareto distribution was assumed. In addition, the variation of $\widetilde{\gamma}_{GH,GPD}$ was rather unappealing. Even for the case $n = 500$, its standard error was nine times of that of $\widetilde{\gamma}$. This is likely due to the overestimation of the threshold and the large standard error of $\widetilde{u}_{GH}$. On the other hand, Beirlant, Joossens, and Segers's method on some occasions yielded an average value of $\widetilde{\gamma}_{BJS}$ that was much different from $\gamma_0 = 0.4$. Our approach compared favourably, producing average values of $\widetilde{\gamma}$ which were the closest to 0.4 among all other estimators in more than half of all cases under investigation.

To conduct a fair comparison, we also considered samples in favour of the two competing methods. We drew samples from one of the following null distributions:

(a) a Pareto distribution with parameter α given by $F(x) = 1 - x^{-\alpha}$, for which $\gamma = \alpha^{-1}$;

(b) a GPD.

In the former case, we gave explicit results for $\alpha = 5.0$. Our method gave average values of $\widetilde{\gamma}$ from 0.20 to 0.30. We encountered some difficulties in applying the threshold selection procedure by Guillou and Hall's method. Altogether 303 replications out of 1,000 failed to select a threshold. A change to $c_{crit} = 1.0$ gave 85 failures. A related note is that the samples may have a thin tail when the Hill estimator is not designed for the EVI close to zero. After removing the 303 failure cases, the average value of $\widetilde{\gamma}_{GH,Hill}$ was 0.19. The other competing method using $\widetilde{\gamma}_{BJS}$ tended to underestimate the EVI, giving an average value of 0.16. In the GPD samples with $\gamma = 0.4$, our method yielded average values of $\widetilde{\gamma}$ between 0.39 and 0.53. Guillou and Hall's method overestimated γ by yielding an average value of 0.49. In six of the seven models, our method outperformed $\widetilde{\gamma}_{GH,Hill}$. Beirlant, Joossens, and Segers's method performed well with an average value of 0.39, and this is because the special case $\rho = -1$ gives the GPD.

5. Data Examples

5.1. Secura Belgian Re data

The first data set under consideration is the Secura Belgian Re data. These are automobile claims in millions from 1988 to 2001 at several European insurance

Table 4.2. Average values of the estimates for the Pareto distribution and the GPD. $\widetilde{\gamma}$ is the MPS estimate of our models, with L being one of the seven distributions (a) to (g). $\widetilde{\gamma}_{GH,Hill}$ is the Hill estimator and $\widetilde{\gamma}_{GH,GPD}$ is the EVI estimator of the GPD by Guillou and Hall's method. $\widetilde{\gamma}_{BJS}$ is the estimator by Beirlant, Joossens, and Segers's method. Standard errors are shown in brackets.

	Pareto ($\alpha = 5.0$) distribution		GPD $(\gamma, \sigma) = (0.4, 1.0)$	
(a) $\widetilde{\gamma}$	0.22	(0.30)	0.39	(0.23)
(b) $\widetilde{\gamma}$	0.25	(0.12)	0.39	(0.15)
(c) $\widetilde{\gamma}$	0.21	(0.25)	0.39	(0.19)
(d) $\widetilde{\gamma}$	0.23	(0.11)	0.43	(0.14)
(e) $\widetilde{\gamma}$	0.24	(0.08)	0.44	(0.13)
(f) $\widetilde{\gamma}$	0.30	(0.32)	0.53	(0.36)
(g) $\widetilde{\gamma}$	0.20	(0.12)	0.47	(0.17)
$\widetilde{\gamma}_{GH,Hill}$	0.19	(0.08)	0.49	(0.18)
$\widetilde{\gamma}_{GH,GPD}$	0.28	(1.55)	0.34	(1.99)
$\widetilde{\gamma}_{BJS}$	0.16	(0.10)	0.39	(0.07)

Table 5.1. The MPS estimates of the threshold models, for different L, for the Secura Belgian Re data. The cases (a) to (g) refer to the corresponding distribution functions in Section 4.

Case	k	$\tilde{\gamma}$	$s.e.(\tilde{\gamma})$	$\tilde{\sigma}$	$s.e.(\tilde{\sigma})$	$\tilde{u}$	p-value
(a)	46	0.097	0.155	1.208	0.253	3.029	0.015
(b)	91	0.429	0.145	0.606	0.104	2.627	0.010
(c)	91	0.429	0.150	0.606	0.107	2.627	0.015
(d)	81	0.337	0.168	0.725	0.149	2.671	0.000
(e)	37	0.162	0.274	1.125	0.405	3.322	0.000
(f)	91	0.429	0.139	0.606	0.100	2.627	0.002
(g)	91	0.429	0.156	0.606	0.112	2.627	0.002

companies. There are 371 observations of at least 1.2 million euros. A study of the data set can be found in Beirlant et al. (2004).

We fitted model (2.1) to the data using various distribution functions for L. In the following, by cases (a) to (g) we mean the corresponding distribution functions in Section 4. The results are summarized in Table 5.1. We proposed two approaches in choosing a suitable L. The first method was to use Moran's statistic, which is a by-product of using the MPS method. The statistic M in (2.3) can be used for testing the goodness of fit of a random sample to a distribution function. Asymptotically, M suitably normalized has an approximate chi-squared distribution (Cheng and Stephens (1989)). Hence, the model with the largest p-value for the goodness-of-fit test is most favourable. The second approach is by means of a Quantile-Quantile plot (QQ-plot). For any class of distributions, the theoretical quantiles are linearly related to the corresponding quantiles of a

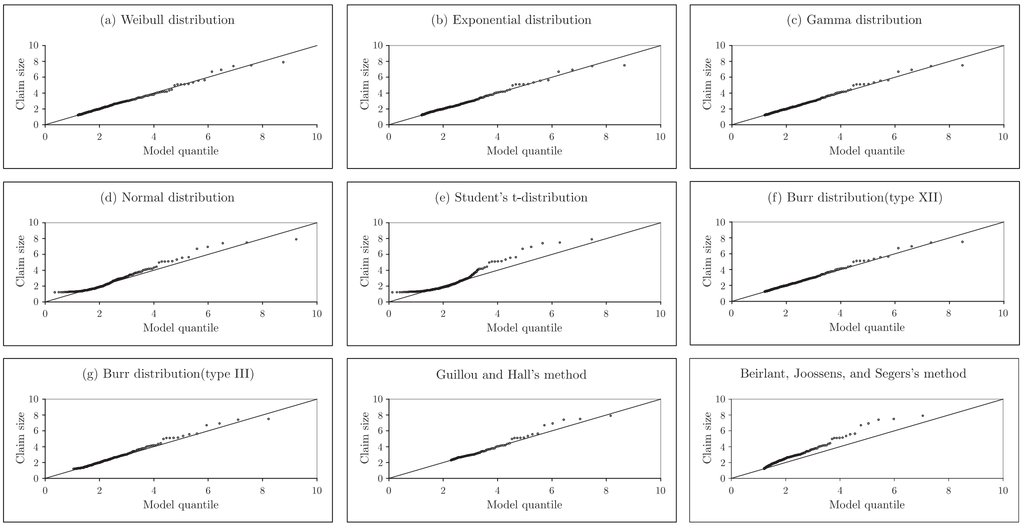


Figure 5.1. QQ plots for the Secura Belgian Re data.

random sample from that class. Hence, a straight line pattern is expected in a scatter plot if the model provides a good statistical fit. Figure 5.1 shows the QQ plots of the models. Based on the above two criteria, model (2.1) with L a Weibull distribution seems to provide the best fit to the data.

To compare our model with the other two approaches, we judged the overall goodness of fit by the average scaled absolute error (Castillo et al. (2005)),

$$ASAE = \frac{1}{k}\sum_{i=n-k+1}^{n} \frac{\left|x_{(i)} - \hat{x}_{(i)}\right|}{\left(x_{(n)} - x_{(n-k+1)}\right)},$$

where $\hat{x}_{(i)}$ are the expected quantiles. In applying Guillou and Hall's method, $c_{crit} = 1.25$ yielded $k = 4$. This was too small to be accepted. A change to $c_{crit} = 1.5$ gave $ASAE = 1.87$ based on 126 exceedances. Beirlant, Joossens, and Segers's method gave $ASAE = 20.77$. Significant improvement was obtained by our model which gave $ASAE = 1.83$ based on the entire data set, and $ASAE = 1.50$ based on 46 excesses over the estimated threshold. The goodness of fit of our model is also apparent from the QQ plots in Figure 5.1.

Our model can provide better insight into the structure of the data. We demonstrate this with the Secura Belgian Re data. The presence of the threshold indicates a heavy tailed claim size distribution and a loss in excess of the threshold $X > u$ can be severe. The model suggests that $u = 3.029$ is an appropriate reference point in pricing an automobile insurance contract. On the other hand,

Table 5.2. Estimates of $\Pi(R)$, in thousands, at different retention levels R, in millions.

R	3.00	4.00	5.00	7.50	10.00
$\widehat{\Pi(R)}$	183.37	89.15	45.65	10.30	2.85

in a reinsurance contract, the net premium $\Pi(R)$ is calculated on the basis of a retention level R,

$$\Pi(R) = E\left((X-R)_+\right) = \int_R^{x^*} (1-F(y))\,dy, \tag{5.1}$$

where x^* is the upper end-point (Beirlant et al. (2004)). To apply (5.1) and the proposed model with L a Weibull distribution, we have for $\gamma < 1$,

$$\Pi(R) = \begin{cases} \exp\left\{-\left[\frac{(u-b)}{a}\right]^c\right\}\left[1+\frac{\gamma(R-u)}{\sigma}\right]^{-1/\gamma+1}\frac{\sigma}{(1-\gamma)}, & R > u, \\ a[g(\frac{u-b}{a}) - g(\frac{R-b}{a})] + \exp\left\{-\left[\frac{(u-b)}{a}\right]^c\right\}\frac{\sigma}{(1-\gamma)}, & R \le u, \end{cases} \tag{5.2}$$

where $g(y) = \sum_{k=0}^{\infty}(-1)^k y^{kc+1}/[k!\,(kc+1)]$.

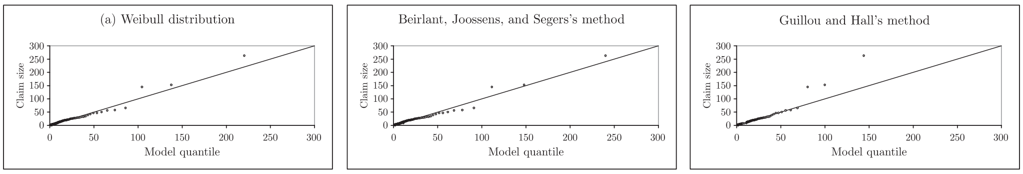
An estimate of $\Pi(R)$ can be obtained by substituting the MPS estimates into (5.2) at different retention levels R. Table 5.2 gives some numerical examples of $\widehat{\Pi(R)}$. Based on our estimates, the mean drops significantly with an increasing retention level R.

5.2. Danish fire claim data

This data set contains insurance losses over one million Danish kroner, from 1980 to 1990. Sample size is 2,157. Our model is based on a Weibull distribution for L. Judging from the overall fit, as measured by the $ASAE$ criterion, our method and Beirlant, Joossens, and Segers's method yielded values of 1.77 and 1.92, respectively, based on the entire data set. Guillou and Hall's method has the smallest $ASAE$ value of 1.16 based on 92 exceedances. However, its QQ plot, as shown in the right panel of Figure 5.2, shows a large departure for each of the three largest claims. The value of a model is determined by its ability to predict future observations. In particular, a model in extreme value analysis should describe the tail adequately. In this sense, our model seems more appealing.

6. Conclusion

There is a long history in the application of the peaks-over-threshold method in diverse fields. The selection of a threshold is an important and challenging problem. We find that there are difficulties in applying some of the existing



(a) Weibull distribution

Beirlant, Joossens, and Segers's method

Guillou and Hall's method

Figure 5.2. QQ plots for the Danish fire claim data.

methods. Guillou and Hall's approach requires a specification of some arbitrary parameters. In its application to the Secura Belgian Re data, the method yielded an inappropriate number of upper order statistics. Simulation experiments also revealed that the result may not be reliable when the extreme value index is close to zero. On the other hand, Beirlant, Joossens, and Segers's method does not always guarantee a good fit in application. In light of this, our approach seems appealing; in addition to providing an estimate of the threshold based on the entire sample, it provides a global fit to the data with an appropriate tail fit. The estimate of the threshold is shown to be super-consistent, and this leads to a much better estimation of the tail parameter. From extensive simulation experiments and two case studies, our method seems to be more reliable and flexible in modeling extreme value data. The sampling distribution of the threshold estimate is clearly an important open problem that deserves further investigation.

Acknowledgement

Partial support by the Area of Excellence Scheme under the University Grants Committee of the Hong Kong Special Administration Region, China (Project AoE/P-04/2004 and GRF grant HKU7036/06P) is gratefully acknowledged. We thank the referees and the Editors for comments that led to improvement of the paper.

Appendix. Proof of Theorem 2.1

We first describe the asymptotic framework for model (2.1), then regularity conditions are outlined. The proof of Theorem 2.1 is completed after two lemmas are presented and proved.

As in Smith (1987), we assume that $k^{-1}n\,(1 - F\,(u)) \to_p 1$ such that the GPD holds. Under very general conditions, the MPS estimators including $\widetilde{u}$ are consistent (Shao (2001)). In the following, let l and g_u be the density functions of L and the GPD, respectively. Let $\phi = (\gamma, \sigma)$ and denote the true parameter and the MPS estimator by (θ_0, ϕ_0, u_0) and $\left(\widetilde{\theta}, \widetilde{\phi}, \widetilde{u}\right)$, respectively. Let

$$\begin{aligned}
&D_L\left(x_{(i)}, \theta, \phi, u\right) = L\left(x_{(i)}; \theta\right) - L\left(x_{(i-1)}; \theta\right) \quad \text{and} \\
&D_U\left(x_{(i)}, \theta, \phi, u\right) = \left(1 - L\left(u; \theta\right)\right)\left(G_u\left(x_{(i)}; \phi\right) - G_u\left(x_{(i-1)}; \phi\right)\right).
\end{aligned}$$

Condition 1. For all x and for all (θ, ϕ), the partial derivatives $\partial M/\partial\theta$, $\partial M/\partial\phi$, $\partial^2 M/d\phi^2$ and $\partial^2 M/\partial\theta^2$ exist.

Condition 2. The first partial derivatives $|\partial M(x, \theta, \phi, u)/\partial\theta|$ and $|\partial M(x, \theta, \phi, u)/\partial\phi|$ are bounded by integrable functions.

Condition 3. For points in the interval $(u_0, u]$ or $[u, u_0)$, the spacings can be approximated by $F\left(x_{(i)}\right) - F\left(x_{(i-1)}\right) = \left[f\left(u_0\right) + o(1)\right]\left(x_{(i)} - x_{(i-1)}\right)$, where f is the density function of F. As $u \to u_0$, we have the limits

$$\begin{aligned}
&\lim_{x_{(i)} \le u, u \to u_0} D_L\left(x_{(i)}, \theta, \phi, u\right) = \left[a\left(\theta, \phi, u_0\right) + o\left(1\right)\right]\left(x_i - x_{i-1}\right), \\
&\lim_{x_{(i)} > u, u \to u_0} D_U\left(x_{(i)}, \theta, \phi, u\right) = \left[b\left(\theta, \phi, u_0\right) + o\left(1\right)\right]\left(x_i - x_{i-1}\right),
\end{aligned}$$

where $a\left(\theta, \phi, u_0\right) = l\left(u_0; \theta\right)$ and $b\left(\theta, \phi, u_0\right) = \left(1 - L\left(u_0; \theta\right)\right)\sigma^{-1}$.

Condition 4. For points in the interval (u, ∞) but not in (u_0, u) or (u_0, ∞) but not in (u, u_0), we have

(i) $D_U\left(x_{(i)}, \theta, \phi, u\right) = \left(1 - L\left(u; \theta\right)\right) g_u\left(\xi_{(i)}; \phi\right)\left(x_{(i)} - x_{(i-1)}\right)$ for some $\xi_{(i)}$ in $\left(x_{(i)}, x_{(i-1)}\right)$;

(ii) the first order derivative of $\log D_U\left(x_{(i)}, \theta, \phi, u\right)$ with respect to u,

$$\frac{\partial}{\partial u}\log D_U\left(x_{(i)}, \theta, \phi, u\right) = \frac{-l\left(u; \theta\right) g_u\left(\xi_{(i)}; \phi\right) + \left(1 - L\left(u; \theta\right)\right)\partial g_u\left(\xi_{(i)}; \phi\right)/\partial u}{\left(1 - L\left(u; \theta\right)\right) g_u\left(\xi_{(i)}; \phi\right)},$$

is bounded;

(iii) the first order derivative in (ii), evaluated at the true parameters, has an expected value with respect to the true distribution given by

$$\begin{aligned}
&= \int_{u_0}^{x^*} -l\left(u_0; \theta_0\right) g_{u_0}\left(x; \phi_0\right) dx + \int_{u_0}^{x^*} \left(1 - L\left(u_0; \theta_0\right)\right)\frac{\partial}{\partial u} g_{u_0}\left(x; \phi_0\right) dx \\
&= -l\left(u_0; \theta_0\right)\int_{u_0}^{x^*} g_{u_0}\left(x; \phi_0\right) dx - \int_{u_0}^{x^*} \left(1 - L\left(u_0; \theta_0\right)\right) d g_{u_0}\left(x; \phi_0\right) \\
&= -l\left(u_0; \theta_0\right) + \left(1 - L\left(u_0; \theta_0\right)\right)\sigma_0^{-1} \\
&= -a\left(\theta_0, \phi_0, u_0\right) + b\left(\theta_0, \phi_0, u_0\right),
\end{aligned}$$

where $x^* > u_0$ is the right end-point of the GPD.

Lemma 1. *Let $F_n(x)$ be the empirical distribution function and suppose Conditions* 3 *to* 4 *hold. Then,* $\max_u (1/n)M(\theta_0,\phi_0,u) - (1/n)M(\theta_0,\phi_0,u_0)$ *and* $\max_u \{F_n(u)-F_n(u_0)-c(\theta_0,\phi_0,u_0)(u-u_0)\}$ *are asymptotically equivalent, where*

$$c(\theta_0,\phi_0,u_0) = \frac{a(\theta_0,\phi_0,u_0) - b(\theta_0,\phi_0,u_0)}{\log a(\theta_0,\phi_0,u_0) - \log b(\theta_0,\phi_0,u_0)}.$$

Proof. The MPS method maximizes the function $M(\theta_0,\phi_0,u)$ with respect to u. Note that

$$\begin{aligned} &M(\theta_0,\phi_0,u)\\ &=\sum_{i=1}^{n+1} I\left(x_{(i)} \le u\right) \log D_L\left(x_{(i)},\theta_0,\phi_0,u\right) + \sum_{i=1}^{n+1} I\left(x_{(i)} > u\right) \log D_U\left(x_{(i)},\theta_0,\phi_0,u\right). \end{aligned}$$

Consider the difference

$$\begin{aligned} &M(\theta_0,\phi_0,u) - M(\theta_0,\phi_0,u_0)\\ &=\sum_{i=1}^{n+1} I\left(u_0 < x_{(i)} \le u\right)\left(\log D_L\left(x_{(i)},\theta_0,\phi_0,u\right) - \log D_U\left(x_{(i)},\theta_0,\phi_0,u_0\right)\right)\\ &\quad+\sum_{i=1}^{n+1} I\left(x_{(i)} > u, u > u_0\right)\left(\log D_U\left(x_{(i)},\theta_0,\phi_0,u\right) - \log D_U\left(x_{(i)},\theta_0,\phi_0,u_0\right)\right)\\ &\quad-\sum_{i=1}^{n+1} I\left(u < x_{(i)} < u_0\right)\left(\log D_L\left(x_{(i)},\theta_0,\phi_0,u_0\right) - \log D_U\left(x_{(i)},\theta_0,\phi_0,u\right)\right)\\ &\quad+\sum_{i=1}^{n+1} I\left(x_{(i)} > u_0, u < u_0\right)\left(\log D_U\left(x_{(i)},\theta_0,\phi_0,u\right) - \log D_U\left(x_{(i)},\theta_0,\phi_0,u_0\right)\right). \end{aligned}$$

For points in the interval $(u_0, u]$ or (u, u_0), apply Condition 3 to the first and third lines above to get

$$\begin{aligned} &\log D_L\left(x_{(i)},\theta_0,\phi_0,u\right) - \log D_U\left(x_{(i)},\theta_0,\phi_0,u\right)\\ &\quad= \log a(\theta_0,\phi_0,u_0) - \log b(\theta_0,\phi_0,u_0) + o(1). \end{aligned}$$

For points in the interval (u,∞) but not in (u_0,u) or (u_0,∞) but not in (u,u_0), consider the Taylor series expansion of $\log D_U$ around $u = u_0$ and apply Conditions 4(i) and (ii) to the second and fourth lines above to get

$$\begin{aligned} &\log D_U\left(x_{(i)},\theta_0,\phi_0,u\right) - \log D_U\left(x_{(i)},\theta_0,\phi_0,u_0\right)\\ &\quad= (u-u_0)\frac{\partial}{\partial u}\log D_U\left(x_{(i)},\theta_0,\phi_0,u_0\right) + o_p(1). \end{aligned}$$

By the Strong Law of Large Numbers and Condition 4(iii),

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n+1} I\left(x_{(i)} > u, u > u_0\right)(u-u_0)\frac{\partial}{\partial u}\log D_U\left(x_{(i)},\theta_0,\phi_0,u_0\right)\\ &\qquad+\frac{1}{n}\sum_{i=1}^{n+1} I\left(x_{(i)} > u_0, u < u_0\right)(u-u_0)\frac{\partial}{\partial u}\log D_U\left(x_{(i)},\theta_0,\phi_0,u_0\right)\\ &\quad= -\left(u-u_0\right)\left[a\left(\theta_0,\phi_0,u_0\right) - b\left(\theta_0,\phi_0,u_0\right)\right] + o_p(1).\end{aligned}$$

Consider the difference per observation and replace the indicator function by the empirical distribution. Then,

$$\begin{aligned}&\frac{1}{n}M\left(\theta_0,\phi_0,u\right) - \frac{1}{n}M\left(\theta_0,\phi_0,u_0\right)\\ &\quad= \left[F_n(u) - F_n(u_0)\right]\left[\log a\left(\theta_0,\phi_0,u_0\right) - \log b\left(\theta_0,\phi_0,u_0\right) + o(1)\right]\\ &\qquad -\left(u-u_0\right)\left[a\left(\theta_0,\phi_0,u_0\right) - b\left(\theta_0,\phi_0,u_0\right)\right] + o_p(1).\end{aligned}$$

Hence, the problem is translated into maximizing $[F_n(u) - F_n(u_0)] - c(\theta_0,\phi_0,u_0) \times (u-u_0)$ with respect to u, where $c(\theta_0,\phi_0,u_0)$ is defined in the statement of the Lemma.

Lemma 2 below is a modified version of Lemma 2 in Chernoff and Rubin (1956). We choose u such that $F(u) - F(u_0)$ is arbitrarily close to $F_n(u) - F_n(u_0)$ with large probability, provided u and u_0 are large. As in Chernoff and Rubin (1956) it suffices to consider the uniform distribution in a small range.

Lemma 2. *For the uniform distribution, for each $\varepsilon_1 > 0$ and $\eta_1 > 0$, there are $0 < K_1 < K_2$ such that*

$$P\left(\max_{K_1/n \le y \le K_2/n}\left|\frac{F_n(y)}{y} - 1\right| < \eta_1\right) > 1 - \varepsilon_1.$$

Proof. Let $Y_1, \ldots, Y_n$ be i.i.d. random variables from the uniform distribution on $[0,1]$. It is easy to check that the indicator function $I(Y_1 \le y)$ has mean y and variance $y(1-y)$ for $0 < y < 1$. Hence,

$$\frac{F_n(y)}{y} = \frac{1}{yn}\sum_{i=1}^{n} I\left(Y_{(i)} \le y\right)$$

has mean one and variance $(1-y)/(ny)$. By Chebyshev's inequality,

$$P\left(\left|\frac{F_n(y)}{y} - 1\right| > \eta_2\right) < \frac{1-y}{\eta_2^2 ny} < \frac{1}{\eta_2^2 ny},$$

for $\eta_2 > 0$ and for $a > 1$, we have

$$P\left(\max_{i=0,1,\dots,r}\left|\frac{F_n\left(a^iK_1/n\right)}{\left(a^iK_1/n\right)}-1\right|>\eta_2\right)<\frac{1}{\eta_2^2 n}\sum_{i=0}^{r}\left(\frac{n}{a^iK_1}\right)=\frac{a^{r+1}-1}{\eta_2^2K_1a^r\left(a-1\right)}.$$

If

$$\left|\frac{F_n\left(y\right)}{y}-1\right|<\eta_2 \quad\text{and}\quad \left|\frac{F_n\left(ay\right)}{ay}-1\right|<\eta_2,$$

then for $y \le z \le ay$,

$$-\frac{1}{a}\eta_2+\frac{1}{a}-1<\frac{F_n\left(y\right)}{ay}-1<\frac{F_n\left(z\right)}{z}-1<\frac{F_n\left(ay\right)}{y}-1<a\eta_2+a-1.$$

We may select η_2 such that $a\eta_2+a-1<\eta_1$ and $-\eta_2/a+1/a-1>-\eta_1$. Then, select K_1 and K_2 such that $\eta_2^2K_1a^r\left(a-1\right)/\left(a^{r+1}-1\right)>1/\varepsilon_1$ and $K_2 \ge a^rK_1$.

Proof of Theorem 2.1. Let $\psi = (\theta, \phi)$. Consider the Taylor series expansion of M around $\widetilde{\psi} = \psi_0$:

$$\frac{1}{n}M\left(\widetilde{\psi},\widetilde{u}\right)=\frac{1}{n}M\left(\psi_0,\widetilde{u}\right)+\left(\widetilde{\psi}-\psi_0\right)\frac{1}{n}\frac{\partial}{\partial\psi}M\left(\overline{\psi},\widetilde{u}\right)+o_p\left(1\right),$$

where $\overline{\psi}$ is between $\widetilde{\psi}$ and ψ_0. By the consistency of $\widetilde{\psi}$ and Condition 2 that $n^{-1}\partial M\left(\overline{\psi},\widetilde{u}\right)/\partial\psi = O_p\left(1\right)$, we can focus on the first term on the right-hand side. By Lemma 1 and Chernoff and Rubin (1956, Lemma 4), we can treat $\widetilde{u}$ as the maximizer of the following

$$\frac{1}{n}M\left(\widetilde{\theta},\widetilde{\phi},\widetilde{u}\right)=\frac{1}{n}M\left(\theta_0,\phi_0,u_0\right)+F_n\left(\widetilde{u}\right)-F_n\left(u_0\right)-c\left(\theta_0,\phi_0,u_0\right)\left(\widetilde{u}-u_0\right)+o_p\left(1\right).$$

Now the first term on the right-hand side is a constant. The rate of convergence of $\widetilde{u}$ can be determined by $H\left(u\right) = F_n\left(u\right) - F_n\left(u_0\right) - c\left(\theta_0,\phi_0,u_0\right)\left(u-u_0\right)$. Since $H\left(u_0\right) = 0$, it will suffice to show that, for u outside a neighborhood of u_0, $H\left(u\right) < 0$. By Condition 3, we have

$$F\left(u\right)-F\left(u_0\right)=\begin{cases}\left[a\left(\theta_0,\phi_0,u_0\right)+o\left(1\right)\right]\left(u-u_0\right), & u<u_0,\\ \left[b\left(\theta_0,\phi_0,u_0\right)+o\left(1\right)\right]\left(u-u_0\right), & u>u_0.\end{cases}$$

We have shown in Lemma 2 that, with large probability, $F\left(u\right) - F\left(u_0\right)$ is arbitrarily close to $F_n\left(u\right) - F_n\left(u_0\right)$ for $n(u-u_0)$ large enough. Using the fact that $w < (v-w)/(\log v - \log w) < v$ for any positive constants $v > w$, we have $a\left(\theta_0,\phi_0,u_0\right) - c\left(\theta_0,\phi_0,u_0\right) > 0$ and $b\left(\theta_0,\phi_0,u_0\right) - c\left(\theta_0,\phi_0,u_0\right) < 0$. Hence, for each ε there is a K such that

$$P\left(\max_{K/n<\left|u-u_0\right|}H\left(u\right)<0\right)>1-\varepsilon.$$

References

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley, Chichester.

Beirlant, J., Joossens, E. and Segers, J. (2004). Discussion of "Generalized Pareto fit to the society of actuaries' large claims database" by A. Cebrian, M. Denuit and P. Lambert. *North Amer. Actuarial J.* **8**, 108-111.

Castillo, E., Hadi, A. S., Balakrishnan, N. and Sarabia, J. M. (2005). *Extreme Value and Related Models with Applications in Engineering and Science.* John Wiley & Sons, New Jersey.

Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society B* **45**, 394-403.

Cheng, R. C. H. and Stephens, M A. (1989). A goodness-of-fit test using Moran's statistic with estimated parameters. *Biometrika* **76**, 385-392.

Chernoff, H. and Rubin, H. (1956). The estimation of the location of a discontinuity in density. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**, 19-37. University of California Press, Berkeley.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society B* **52**, 393-442.

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

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events.* Springer-Verlag, Berlin.

Fitzgerald, D. L. (1996). Maximum product of spacings estimators for the generalized Pareto and log-logistic distributions. *Stochastic Hydrology and Hydraulics* **10**, 1-15.

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.

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

Pickands, J. III (1975). Statistical inference using extreme order statistics. *The Annals of Statistics* **3**, 119-131.

Shao, Y. (2001). Consistency of the maximum product of spacings method and estimation of a unimodal distribution. *Statistica Sinica* **11**, 1125-1140.

Smith, R. L. (1987). Estimating tails of probability distributions. *The Annals of Statistics* **15**, 1174-1207.

Tong, H. (1978). On a threshold model. In *Pattern Recognition and Signal Processing* (Edited by C. H. Chen). Sijthoff and Noordhoff, Amsterdam.

Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data (with Discussion). *Journal of the Royal Statistical Society B* **42**, 245-292.

Wong, T. S. T. and Li, W. K. (2006). A note on the estimation of extreme value distributions using maximum product of spacings. *IMS Lecture Notes Monograph Series* **52**, 272-283.

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong.

E-mail: h0127272@hkusua.hku.hk

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong.

E-mail: hrntlwk@hku.hk

(Received January 2008; accepted March 2009)