Computational Statistics and Data Analysis 56 (2012) 143–150

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

A quantile estimation for massive data with generalized Pareto distribution[☆]

Jongwoo Song[a,*], Seongjoo Song[b]

[a] *Department of Statistics, Ewha Womans University, 11-1 Daehyun-dong, Seodaemun-gu, Seoul 120-750, Republic of Korea*
[b] *Department of Statistics, Korea University, 5-1 Anam-dong, Seongbuk-Gu, Seoul, 136-701, Republic of Korea*

ARTICLE INFO

Article history:
Received 6 June 2010
Received in revised form 4 May 2011
Accepted 23 June 2011
Available online 7 July 2011

Keywords:
Quantile estimation
Generalized Pareto distribution
Peak over threshold
Massive data
Parameter estimation
Nonlinear least squares

ABSTRACT

This paper proposes a new method of estimating extreme quantiles of heavy-tailed distributions for massive data. The method utilizes the Peak Over Threshold (POT) method with generalized Pareto distribution (GPD) that is commonly used to estimate extreme quantiles and the parameter estimation of GPD using the empirical distribution function (EDF) and nonlinear least squares (NLS). We first estimate the parameters of GPD using EDF and NLS and then, estimate multiple high quantiles for massive data based on observations over a certain threshold value using the conventional POT. The simulation results demonstrate that our parameter estimation method has a smaller Mean square error (MSE) than other common methods when the shape parameter of GPD is at least 0. The estimated quantiles also show the best performance in terms of root MSE (RMSE) and absolute relative bias (ARB) for heavy-tailed distributions.

1. Introduction

Massive datasets resulting from, for instance, internet traffic, sales transactions, cell phone call histories and query histories of internet search engines have become very common these days. These huge datasets contain so much information that they are sometimes hard to analyze, such as the difficulty in the computation of quantiles. Quantiles have essential information on the underlying distribution from which a sample is taken. When the sample size is small, we can compute exact sample quantiles easily and quickly. However, the computation time of quantiles increases very rapidly with increasing sample size. See Fig. 1 for the actual computation time with statistical package R for the mean and median of samples of different sizes. Computation of quantiles needs huge space as well because it needs to sort the observed values. Therefore, the actual computation time for quantiles usually takes more than the theoretical time because of the memory I/O time and other reasons. A lot of literature in the field of computer science has focused on this computational issue. For instance, Blum et al. (1973) showed that at most $5.43N$ comparisons are needed to find the kth largest observation of the dataset of size N. Later, the upper bound of the number of comparison became much tighter. Currently, the upper bound is shown to be less than $3N$ and the lower bound is $(2+\epsilon)N$ where ϵ is the order of 2^{-40}. One can refer to Paterson (1997) or Manku et al. (1998) for details. We also find many quantile (or simply median) computation algorithms for massive data in the computer science literature. Refer, for example, to Manku et al. (1998), Liechty et al. (2003) and Chen et al. (2000). In the statistics literature, we can find a few similar approaches such as the “Remedian” proposed by Rousseeuw and Bassett (1990). Most of these methods are designed to have two properties – single pass and low storage – because the computation should be done in a

[☆] This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2010-0004196 (J. Song) and No. 2010-0017185 (S. Song)).

[*] Corresponding author.
E-mail addresses: josong@ewha.ac.kr (J. Song), sjsong@korea.ac.kr (S. Song).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.06.030

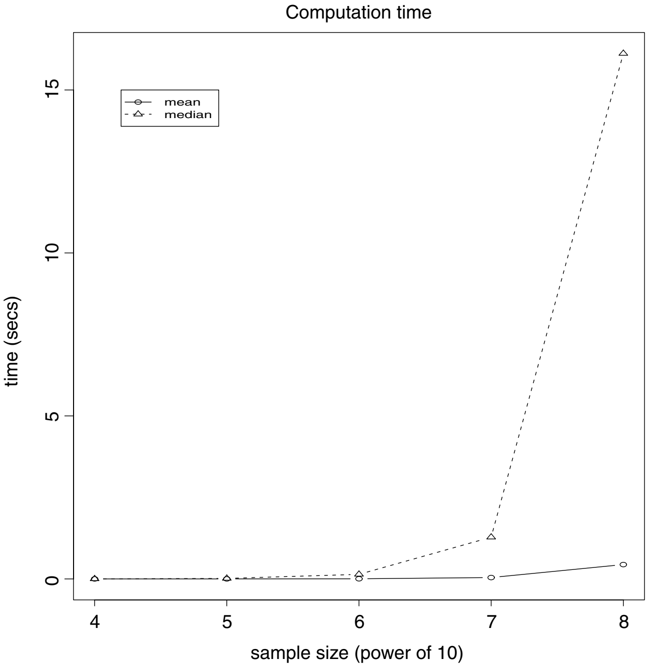

Fig. 1. Computation time for mean and median with different sample sizes.

limited memory environment and computation time should also be reduced. However, a different viewpoint that has also been presented in the statistics literature aims to provide a good estimation of high quantiles, especially for heavy-tailed distributions.

Many of the statisticians who have been interested in this problem have used extreme value theory as their major tool. Among the several methods considered, the peak over threshold (POT) method with generalized Pareto distribution (GPD) has been quite successful and very widely used. A common application of the POT method is to estimate extreme quantiles with datasets in the field of insurance, such as flood data, in which cases, extreme events are rarely observed. In other words, the POT method has been commonly used to estimate extreme quantiles when the datasets are not massive. In this paper, we would like to use this POT method to estimate high quantiles of massive data. When the dataset has a huge number of observations, we first take a sample, apply the POT method to the sample, and estimate high quantiles. To apply the POT method, the GPD parameter estimation is essential. Here, we propose a new method of parameter estimation and use this method in the quantile estimation by the conventional POT. Since we use a small sample from a massive dataset unlike many approaches in the computer science literature, our method can reduce the computation time and storage substantially from methods using all of the data.

Of the many GPD parameter estimation methods that have been used in the literature, Pickands proposed a method using order statistics of samples in Pickands (1975), Hosking and Wallis (1998) suggested probability weighted moments, and Dupuis and Tsao (1998) proposed hybrid probability weighted moments, Juarez and Schucany (2004) introduced a minimum density power divergence method, and Zhang (2007) proposed a likelihood moment estimator. Zhang and Stephens (2009) proposed a new efficient estimator and Zhang (2010) improved the previous version of the estimator. The maximum likelihood estimation (MLE) for GPD discussed by Davison (1984) or Smith (1985) is also one of the most popular methods. Rather than a complete survey of all existing methods, we mostly focus in the following sections on methods that are available in statistical programming environment R (R Development Core Team, 2010).

The goal of this paper is to propose a new method of estimating extreme quantiles of heavy-tailed distribution for massive data. The remainder of this paper is organized as follows. Section 2 introduces the POT method with GPD for estimating high quantiles. Section 3 proposes a new parameter estimation method for GPD using the empirical distribution function (EDF) and nonlinear least squares (NLS) method. In Section 4, we compare the performances of parameter estimation methods for GPD with several simulated datasets. Section 5 shows the results of quantile estimation using the POT method based on several different parameter estimation methods with simulated and real datasets. Section 6 concludes the paper.

2. Peak over threshold (POT)

Many applications need estimation of high quantiles for heavy-tailed distributions such as in the case of extreme events with low occurrence but great impact, especially in finance and insurance applications. Statisticians have been using extreme value theory to estimate high quantiles and the POT method is one of the most widely used methods. Pickands (1975) and Balkema and de Haan (2004) showed that the distribution of excesses can be approximated by the GPD if the distribution is in the maximum domain of attraction. In fact, most of the common continuous distributions are in the maximum domain of attraction as seen in Embrechts et al. (1997). Therefore, we may apply the POT method even if we do not know the underlying distribution in many applications.

The distribution function of GPD is defined as follows.

$$F_{\xi,\sigma}(x) = \begin{cases} 1-(1+\xi x/\sigma)^{-1/\xi} & \text{if } \xi \neq 0, \\ 1-\exp(-x/\sigma) & \text{if } \xi = 0, \end{cases}$$

where $\sigma > 0. x \geq 0$ when $\xi \geq 0$ and $0 \leq x \leq -\sigma/\xi$ when $\xi < 0$. This GPD family can be also extended to define a GPD with a location parameter, μ, as $F_{\xi,\mu,\sigma}(x) = F_{\xi,\sigma}(x-\mu)$. We define σ as the scale parameter and ξ as the shape parameter of the GPD. In general, we consider that the distribution is heavy-tailed when the shape parameter $\xi > 0$, medium-tailed when $\xi = 0$, and short-tailed when $\xi < 0$. According to this rule, distributions such as Pareto, Burr, log-gamma, and Cauchy are in the class of heavy-tailed distributions, distributions such as normal, exponential, gamma, and lognormal are in the class of medium-tailed distributions, and distributions such as uniform and beta are in the class of short-tailed distributions. For details of GPD, one can refer to McNeil and Saladin (1997). Our interest in this paper is in medium- to heavy-tailed distributions, so we focus on the cases when $\xi \geq 0$ in the following sections. Moreover, when we apply the POT method, the location parameter μ is not relevant because we consider observations that are larger than a certain threshold value. Therefore, we are only interested in estimating σ and ξ.

Now, let us briefly explain the POT method to estimate high quantiles in the following few steps.

1. For a given sample $X_1, \ldots, X_n$, pick a threshold value u.
2. Fit the GPD to data with observations above u only and estimate (σ, ξ).
3. Approximate the original distribution function as follows.

   $$\hat{F}(x) = (1-F_n(u))F_{\hat{\xi},u,\hat{\sigma}}(x) + F_n(u), \tag{1}$$

   where $F_n(x)$ is the EDF of the sample and $F_{\hat{\xi},u,\hat{\sigma}}(x)$ is the theoretical distribution function of GPD fitted with observations over u.
4. Compute the pth quantile by inverting (1) as follows.

   $$X_p = F^{-1}_{\hat{\xi},u,\hat{\sigma}}\left(1 - \frac{1-p}{1-F_n(u)}\right).$$

   For more theoretical details on POT, refer to Embrechts et al. (1997).

3. A new parameter estimation method for generalized Pareto distribution

The proposed method of estimating the GPD parameters minimizes the sum of squared deviations between the EDF of the data and the theoretical distribution function of GPD at observed values. It is similar to Hill's estimator (Hill, 1975) in the sense that Hill's estimator uses the quantile–quantile comparison to estimate the parameters while we use the probability–probability comparison. We use only observations that are greater than a threshold value u to estimate the parameters as in the usual POT method.

Suppose we have n observations $x_1, \ldots, x_n$ and n_u observations among them are greater than the threshold value u. Let $z_1, \ldots, z_{n_u}$ be those observations. We would like to find (σ, ξ) that minimizes the squared deviations between the empirical distribution and the theoretical GPD at these observations. Therefore, our proposed estimator for (σ, ξ) is

$$\arg\min_{(\sigma,\xi)} \sum_{i=1}^{n_u} (F_n(z_i) - F_{\xi,u,\sigma}(z_i))^2, \tag{2}$$

where $F_n(x)$ is the EDF and $F_{\xi,u,\sigma}(x)$ is the theoretical distribution function of GPD.

To find estimates of (σ, ξ) as in (2), we use a NLS method with the EDF as the response variable and the theoretical distribution function of GPD as the explanatory variable. However, a direct fitting does not work well because the distribution function of GPD is very sensitive to the shape parameter and is therefore not easy to fit. In other words, if the initial value of the shape parameter required for the NLS procedure is not close to the true value, then the theoretical GPD distribution function gives very extreme values. To check the sensitivity of the GPD distribution function to the shape parameter, we ran a simple simulation, generating 10^4 GPD observations with the shape parameter $\xi = 2$ and the scale parameter $\sigma = 1$. We used the 90th quantile as the threshold value u, so 1000 observations were selected. We computed the EDF with these observations and also computed the values of GPD distribution function with estimated parameters using various initial values of the shape parameter. Our range of EDF values was 0.9 to 1, but as we see from Table 1, most of the theoretical GPD function values are very close to 1 when the initial value of the shape parameter is not close to the true value. Thus, the NLS method cannot be used to fit this case because all the explanatory observations are almost identical.

So we propose a two-step fitting procedure. As we see from Table 1, if we take the log of the GPD distribution function, then it gets quite stable in the sense that the explanatory variable has long enough ranges even when the initial value of the

Table 1
Ranges of GPD distribution function values when the true shape parameter $\xi = 2$ with different initial values.

Initial ξ	Original GPD		Log GPD	
	Minimum	Maximum	Minimum	Maximum
0.1	1	1	−155.381	−17.714
0.5	0.999	1	−34.295	−6.469
1.0	0.980	1	−17.841	−3.908
1.5	0.943	1	−12.164	−2.871
2	0.900	1	−9.267	−2.295
3	0.811	0.998	−6.313	−1.664

shape parameter is not close to the true value. Therefore, we fit the log of EDF to the log of the theoretical GPD distribution function. From this fitting, we find the initial values to be used in the original fitting problem. The minimization of the residual sum of squares from the NLS model is computed by the "optim" function in R. The "optim" is a general-purpose optimization function with several optimization algorithms. We use the default optimization algorithm "Nelder–Mead" (Nelder and Mead, 1965) to obtain very stable results. The two-step fitting algorithm is summarized as follows.

1. Pick a threshold value u and compute the EDF using observations above u.
2. Take the log of (1-EDF) and the log of (1-GPD distribution function), which are set as the response and explanatory variables, respectively. Since the GPD distribution function is $F_{\xi,\sigma}(x) = 1 - (1 + \frac{\xi x}{\sigma})^{-\frac{1}{\xi}}$ for $\xi \neq 0$, $\log(1 - F_{\xi,\sigma}(x)) = -\frac{1}{\xi}\log(1 + \frac{\xi x}{\sigma})$. When $F_n(x)$ is the empirical distribution function, we fit the regression with a model
$$\log(1 - F_n(x)) = -\frac{1}{\xi}\log\left(1 + \frac{\xi x}{\sigma}\right) + \text{error}$$
as in Step 3 below. In other words, we find (ξ, σ) that minimizes
$$\sum\left(y_i + \frac{1}{\xi}\log\left(1 + \frac{\xi x_i}{\sigma}\right)\right)^2$$
where $y_i = \log(1 - F_n(x_i))$.
3. Fit the NLS regression of the log of (1-EDF) on the log of (1-GPD distribution function) as explained in Step 2 with initial values of $(\sigma, \xi) = (0.1, 0.01)$. Let us denote the estimated parameters as $(\hat{\sigma}^1, \hat{\xi}^1)$.
4. Fit the NLS regression of the original EDF on the original GPD function with the initial values of $(\hat{\sigma}^1, \hat{\xi}^1)$.

We selected the initial value for the shape parameter as 0.01 because we are interested in estimating quantiles for heavy-tailed distributions. The initial value for the scale parameter does not affect the NLS fitting procedure much. We compare the performance of our estimator with several different estimators in the next section.

4. Simulation results

Our interest is in the case when the shape parameter $\xi \geq 0$; *i.e.* when the distribution is medium- or heavy-tailed. Since our goal is to estimate the high quantiles using POT, we set up the following simulation studies.

1. Generate iid observations following GPD.
2. Pick a threshold value.
3. Estimate the parameters of (σ, ξ) with the observations above the threshold value using several different methods.
4. Repeat steps 1–3 1000 times.
5. Compute the mean square error (MSE) of each estimator.

We generated GPD random variables with the shape parameter, $\xi = 0, 0.5, 1, 2, 3, 4, 5$ and the scale parameter, $\sigma = 1, 10, 100$. We tried sample sizes of 5000, 10,000, and 20,000 and used the 90th quantile as the threshold value; i.e., with 500, 1000, and 2000 observations (sample size) used in the estimation, respectively. In this paper, we only present the result for the case of the sample size of 1000, but the results for the two other sample sizes were very similar. The number of simulations was 1000 in each case and we used $(\sigma, \xi) = (0, 1, 0, 01)$ as the initial values for our proposed estimator in all cases.

Tables 2 and 3 show the mean square errors (MSE) of several estimators that we were comparing. We use the following names in the tables: "mme", "pwmu", "pwmb", "mdpd", "med", "pickands", "lme", "Zhang", "nls-1" and "nls-2", which correspond respectively to the method of moments estimator, the unbiased probability weighted moment method by Hosking and Wallis (1998), the biased probability weighted moment method by Hosking and Wallis (1998), the minimum density power divergence method by Juarez and Schucany (2004), the method of median by He and Fung (1999), the Pickands' method (Pickands, 1975), the likelihood moment method by Zhang (2007), the latest method from Zhang (2010), and, in our own proposed methods, "nls-1" for the one-step NLS method with log empirical distribution and log GPD and "nls-2"

Table 2
Mean square errors (MSE) of estimators of (σ, ξ): fixed σ, varying ξ.

	mme	pwmu	pwmb	Med	Pickands	lme	Zhang	nls-1	nls-2
$\sigma = 1$	0.0019	0.0024	0.0024	0.0041	0.0063	0.0021	0.0121	0.0013	0.0010
$\xi = 0$	0.0010	0.0014	0.0014	0.0068	0.0134	0.0012	0.0010	0.0004	0.0004
$\sigma = 1$	0.2931	0.0297	0.0300	0.1207	0.1742	0.0217	0.0217	0.0044	0.0023
$\xi = 0.5$	0.0119	0.0031	0.0031	0.0117	0.0156	0.0021	0.0022	0.0011	0.0008
$\sigma = 1$	4157	0.5231	0.5393	0.1730	0.2100	0.0402	0.0398	0.0147	0.0061
$\xi = 1$	0.2625	0.0260	0.0262	0.0179	0.0189	0.0042	0.0042	0.0027	0.0017
$\sigma = 1$	$>10^{15}$	$>10^{4}$	$>10^{8}$	0.2825	0.4393	0.1780	0.0720	0.0666	0.0228
$\xi = 2$	2.2545	1.0126	1.0141	0.0277	0.0298	0.0768	0.0084	0.0081	0.0048
$\sigma = 1$	$>10^{31}$	$>10^{12}$	$>10^{25}$	0.7175	0.7953	1.0000	0.1524	0.2164	0.0773
$\xi = 3$	6.2547	4.0051	4.0079	0.0524	0.0467	24.7717	0.0164	0.0191	0.0118
$\sigma = 1$	$>10^{36}$	$>10^{27}$	$>10^{30}$	1.0446	1.5199	1.0000	0.2588	0.4292	0.1258
$\xi = 4$	12.2555	9.0038	9.0081	0.0697	0.0715	52.7989	0.0254	0.0318	0.0195
$\sigma = 1$	$>10^{40}$	$>10^{32}$	$>10^{33}$	1.12507	3.4657	1.0000	0.4193	0.7090	0.2330
$\xi = 5$	20.2567	16.0041	16.0097	0.0814	0.1164	89.8281	0.0337	0.0422	0.0279

Table 3
Mean square errors (MSE) of estimators of (σ, ξ): fixed ξ, varying σ.

	Med	Pickands	lme	Zhang	nls-1	nls-2
$\sigma = 1$	0.1730	0.2100	0.0402	0.0398	0.0147	0.0061
$\xi = 1$	0.0179	0.0189	0.0042	0.0041	0.0027	0.0017
$\sigma = 10$	18.4337	19.8548	99.9763	3.7032	1.4756	0.6285
$\xi = 1$	0.0189	0.0182	8.4259	0.0040	0.0027	0.0018
$\sigma = 100$	1745.2108	2371.0748	9999.9997	341.9315	128.1775	55.6954
$\xi = 1$	0.0190	0.0202	32.9368	0.0037	0.0024	0.0017
$\sigma = 1$	0.2825	0.4393	0.1780	0.0720	0.0666	0.0228
$\xi = 2$	0.0277	0.0298	0.0768	0.0084	0.0081	0.0048
$\sigma = 10$	31.3171	28.6893	100.	8.3644	5.4098	2.4173
$\xi = 2$	0.0340	0.0316	28.5830	0.0091	0.0072	0.0051
$\sigma = 100$	2591.4170	3623.2438	10000.00	721.0839	688.5632	289.0821
$\xi = 2$	0.0330	0.0312	60.7200	0.0089	0.0092	0.0065

for the two-step NLS method with the initial values from “nls-1”. We use the R package “POT” to generate random observations from GPD and to estimate the parameters. See Ribatet (2009) for R procedures.

Although MLE is very popular, we did not include its performance in the simulation results because the estimation could not be done in most of the cases when the shape parameter ξ is greater than 0.5 due to the observed information matrix becoming singular in many cases. The “mdpd” method also gives very unstable results when the shape parameter is greater than 1. Therefore, we did not include its results.

As shown in Table 2 with a fixed scale parameter and a varying shape parameter, the performance of the estimators is very close when the shape parameter ξ is zero or 0.5. However, “mme”,“pwmu”,“pwmb”, and “mdpd” methods start to break down when the shape parameter reaches 1. When $\xi > 1$, these methods under-estimate the shape parameter and heavily over-estimate the scale parameter, although the actual estimated values are not reported in Table 2. The “lme” method performs well until the shape parameter is 2 but starts to break down when the shape parameter is greater than 2. Only “med”, “pickands”, and “Zhang” methods give results comparable with our proposed methods. In fact, Zhang’s method gives results very close to ours. Nevertheless, our proposed methods clearly give the best results.

In Table 3, we compare the performances of some estimators when the shape parameter is fixed and the scale parameter is varying. Again, our proposed methods clearly perform best in this case. Note that the “lme” estimator performs well with small values of the scale parameter, but it gives very high MSE values for large values of the scale parameter. Especially, the MSE of $\hat{\sigma}$ is almost the same as the square of the true σ, which is explained as follows. “lme” overestimates the shape parameter, ξ, when σ is large. With very large values of $\hat{\xi}$, the estimated scale parameter in the original GPD distribution tends to zero. When the “lme” estimator (as well as other estimators) is computed, the scale parameter of GPD is obtained by the following formula

$$\hat{\sigma} = \hat{\sigma^*}(1 - F_n(u))^{\hat{\xi}},$$

where σ^* is the scale parameter of GPD when the observations over the threshold value are treated as if they are an original sample from GPD. (See McNeil and Saladin (1997) for more details.) So if values of $\hat{\xi}$ are large, $\hat{\sigma}$ tends to zero regardless of $\hat{\sigma^*}$ and, thus, the MSE is close to the square of the true value of σ.

Our proposed method should perform better with larger sample sizes since the empirical distribution function converges to the true distribution function as the sample size increases. When we computed means and standard deviations for our proposed estimators with different sample sizes in Table 4, our methods estimated the parameters very accurately with the large sample size, as expected.

Table 4
Means of NLS parameter estimators with increasing sample size (standard deviations are in the parenthesis).

Sample size	nls-1		nls-2	
	$\sigma = 1$	$\xi = 1$	$\sigma = 1$	$\xi = 1$
10^2	1.2151(0.3516)	0.9201(0.1407)	1.0398(0.2530)	0.9832(0.1320)
10^3	1.0475(0.1018)	0.9832(0.0454)	1.0076(0.0724)	0.9991(0.0406)
10^4	1.0065(0.0345)	0.9978(0.0160)	1.0032(0.0251)	0.9992(0.0137)
10^5	1.0023(0.0095)	0.9992(0.0046)	1.0005(0.0079)	0.9999(0.0043)

Table 5
Small sample performance when $n = 50$ and $n = 100$: MSE of (σ, ξ) with Zhang's method and nls-2.

	Zhang		nls-2	
	σ	ξ	σ	ξ
$n = 50, \sigma = 1, \xi = 0$	0.4160	0.0256	0.0185	0.0082
$n = 50, \sigma = 1, \xi = 0.5$	1.1551	0.0521	0.0600	0.0188
$n = 50, \sigma = 1, \xi = 1$	2.0792	0.0783	0.1467	0.0364
$n = 50, \sigma = 1, \xi = 2$	11.5413	0.1772	0.8244	0.1061
$n = 100, \sigma = 1, \xi = 0$	0.1746	0.0114	0.0102	0.0043
$n = 100, \sigma = 1, \xi = 0.5$	0.3617	0.0255	0.0260	0.0092
$n = 100, \sigma = 1, \xi = 1$	0.5922	0.0397	0.0667	0.0178
$n = 100, \sigma = 1, \xi = 2$	2.0056	0.0941	0.3069	0.0559

Table 6
Quantile estimation results for GPD(10, 1): RMSE and ARB (Absolute relative bias, in parenthesis).

Quantile	0.95	0.99	0.999	0.9999
Sample quantiles	8.542(0.036)	95.761(0.079)	3246.604(0.252)	143679.53(0.708)
Pickands	8.798(0.036)	148.121(0.114)	5440.186(0.370)	133278.4(0.736)
Med	8.835(0.036)	155.226(0.123)	4929.983(0.362)	106102.5(0.681)
Zhang	8.264(0.034)	84.707(0.067)	2183.664(0.170)	39193.15(0.293)
nls-2	7.827(0.032)	91.219(0.072)	1877.578(0.145)	29665.25(0.225)

We also checked the small sample performance of our proposed method. We generated sample sizes of 500 and 1000 from GPD and used the 90th percentile for the threshold value. Therefore, the actual sample size is $n = 50$ or $n = 100$. We computed MSE for our "nls-2" estimator and Zhang's estimator (Zhang, 2010) and the results are given in Table 5. As shown in Table 5, MSE values for two estimators are close but our estimator performs better than Zhang's method in all cases.

5. Quantile estimation

In this section, we compare the performances of several estimators of quantiles using POT.

5.1. Simulated data

We generate 10^7 random observations from some heavy-tailed distributions such as GPD, Cauchy, and log-gamma. Then we take a sample of size 10^4 from this generated massive data. We use the 90th quantile for a threshold value to select the 1000 largest observations and estimate the GPD parameters with them. The high quantiles using POT were estimated and the root MSE (RMSE) and absolute relative bias (ARB) were computed to compare the performances of several estimation methods. The ARB is defined as $|\hat{\theta} - \theta|/\theta$ where $\hat{\theta}$ and θ are estimated and true quantiles, respectively. The number of simulations was 1000 in each case. We include results from "med", "pickands", "Zhang", and "nls-2" methods only because other methods gave worse results, as seen in Section 4. We selected only heavy-tailed distributions because, firstly, our main interest is in the heavy-tailed data and secondly, a simple sample quantiles from a large enough sample can be quite accurate when distributions are short-tailed.

Example 1. GPD(10, 1).

GPD is a very widely used distribution in applications of insurance and finance. We use this distribution with the scale parameter 10 and the shape parameter 1. As shown in Table 6, our proposed method has the smallest RMSE and ARB except the 99th percentile. This result is very natural because "nls-2" had the smallest MSE among all estimators considered in Section 4. The estimated quantiles by "nls-2" based on GPD have the smallest RMSE when the underlying distribution is exactly GPD.

Example 2. Cauchy(0, 1).

Table 7
Quantile estimation results for Cauchy(0, 1): RMSE and ARB (in parenthesis).

Quantile	0.95	0.99	0.999	0.9999
Sample quantiles	0.284(0.036)	3.201(0.081)	103.710(0.250)	4627.962(0.742)
Pickands	0.288(0.036)	4.599(0.115)	157.745(0.354)	3633.042(0.661)
Med	0.289(0.036)	5.113(0.127)	157.885(0.362)	3402.722(0.646)
Zhang	0.274(0.035)	2.790(0.068)	69.656(0.170)	1227.404(0.287)
nls-2	0.262(0.033)	2.987(0.075)	61.843(0.162)	966.542(0.253)

Table 8
Quantile estimation results for log-gamma(2, 1): RMSE and ARB (in parenthesis).

Quantile	0.95	0.99	0.999	0.9999
Sample quantiles	35.022(0.070)	1970.930(0.152)	8.8×10^5(0.540)	1.93×10^9(3.869)
Pickands	34.96(0.069)	2584.50(0.203)	8.33×10^5(0.549)	2.47×10^8(1.045)
Med	34.64(0.069)	3007.64(0.233)	9.59×10^6(0.608)	2.56×10^8(1.174)
Zhang	32.03(0.064)	1719.65(0.136)	4.10×10^5(0.306)	7.51×10^7(0.495)
nls-2	31.28(0.063)	1730.51(0.139)	3.42×10^5(0.265)	5.47×10^7(0.393)

Table 9
Quantile estimation results : absolute relative biases (ARB) for SOA data.

Quantile	0.95	0.99	0.999	0.9999
Pickands	0.026	0.054	0.324	0.856
Med	0.024	0.073	0.360	0.909
mle	0.041	0.147	0.100	0.245
Zhang	0.023	0.039	0.108	0.203
nls-2	0.023	0.045	0.100	0.163

Although the Cauchy distribution is not used much in statistical applications, it is a very famous heavy-tailed distribution. We use Cauchy with the location parameter 0 and the scale parameter 1. It is the same as t-distribution with 1 degree of freedom. As we see from Table 7, our proposed method performs best again.

Example 3. Log-gamma(2, 1).

Log-gamma is a distribution that is very commonly used in insurance claims data. We generate log-gamma distribution with the scale parameter 2 and the shape parameter 1. As shown in Table 8, the "nls-2" method performs best among the five methods that were compared.

Tables 6–8 show that the "sample quantiles" with 10^4 observations are not bad for the 95th quantiles but their performance gets worse for the extreme quantiles. Their performance depends on the shape of the population distribution, and especially when the population distribution has a very long tail, the sample quantiles have a high bias for extreme quantiles.

Since different threshold values can give different results when the underlying distribution is not the GPD, we tried a few different threshold values (95th and 99th quantiles) and obtained very similar results. Our method gives the best results for all high quantiles. Although the choice of threshold values depends on data distribution, we recommend a threshold value between the 90th and 94th quantiles if the sample size is sufficiently large ($n \geq 1000$).

5.2. Real data applications

The dataset used in this section is the SOA group medical insurance large claims data in 1991. It is from Beirlant et al. (2004) and is available on the Internet at http://lstat.kuleuven.be/Wiley/. There are 75,789 observations (claims) with a range of values from 25,000 to 4518,000. Although not a very massive dataset, it can be a relevant example to compare the performances of the estimators because this dataset has a long tail. Unlike simulated datasets, it is important to pick the optimum threshold value. If the selected threshold value is too low then the distribution of the excess can be far from GPD, leading to high bias, but if the threshold value is too high then the variance of the estimator will be increased. Thus, there is a trade-off between low threshold and high threshold. We used the mean excess plot (see Davison and Smith, 1990, for details) and selected the 94th quantile for the threshold value, as a common practice in using the POT method. We took a sample of size 5000 and estimated the GPD parameters using observations over the 94th quantile value of this sample with different methods. Then we estimated the high quantiles using POT and checked the biases. The number of simulations is 1000. Table 9 shows the ARB for the SOA data. The 95th quantile is estimated well for most of the methods but for the 99th and higher quantiles, our proposed method and Zhang's method have the smallest biases. Our method has the smallest biases at $p = 0.95$, $p = 0.999$ and $p = 0.9999$ and Zhang's method has the smallest biases at $p = 0.95$ and $p = 0.99$. For

Table 10
Computation time in seconds.

Sample size	10^2	10^3	10^4	10^5
Zhang	0.0021	0.0030	0.0910	1.0940
MLE	0.0053	0.0281	0.1957	1.7795
NLS-2	0.0142	0.0433	0.3000	3.4210

these data, we can use MLE for the parameter estimation, and the results are included in Table 9. The quantile estimation with MLE gives a better result only for $p = 0.999$ and $p = 0.9999$ than "pickands" and "med" methods. Our proposed method gives better results than MLE for all quantiles. We also tried to estimate high quantiles using the 92nd and 96th quantiles as different threshold values. The results were very similar to those with the 94th quantile; that is, our method shows better results than other methods we compare.

6. Conclusions

We proposed a new, fast and stable parameter estimation method for extreme quantiles of heavy-tailed distributions with massive data. Our estimation method is fast to compute, very stable, and does not depend on initial values. We compared the computation times for Zhang's method, MLE and our own method. The results are presented in Table 10. Due to the very short computational time required to estimate the parameters with a small sample size, we measure the time for 100 iterations and compute the mean time, using a Dell PC with Intel E6600 (2.4 GHz) and 4 GB memory. As shown in Table 10, the computation time of our method remains quite short even for a large sample size, although it is longer than that of any other method.

When we compared our new method of estimating the GPD parameters with other existing methods, it gave the smallest MSE when the shape parameter is greater than or equal to zero. In both of the large and small sample cases, our method worked very well compared to the other methods. Although we did not include the simulation results when the shape parameter is less than zero, it still estimates the parameters well if the shape parameter is greater than −0.5. If the shape parameter is less than −0.5, there is a numerical problem in the NLS function in the statistical package R. This problem may be solvable by changing the initial values adaptively, but such a solution was not investigated further because of our focus on estimating quantiles for heavy-tailed distributions.

We also compared the performance of quantile estimation using POT for simulated massive data and real data with a long tail. Again, our method performed best for some common heavy-tailed distributions and for the real data that we presented. For real datasets, it is very important to pick the right threshold value because, in general, real datasets can be noisy and are sometimes sourced from a mixture of several distributions. The estimation function we propose is available at http://home.ewha.ac.kr/~josong/POT/index.html.

References

Balkema, A.A., de Haan, L., 2004. Residual life time at great age. Annals of Probability 2, 792–804.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., 2004. Statistics of Extremes: Theory and Applications. Wiley, West Sussex, England.

Blum, M., Floyd, W., Pratt, V.R., Rivest, R.L., Tarjan, R.E., 1973. Time bounds for selection. Journal of Computer and System Sciences 7, 448–461.

Chen, F., Lamber, D., Pinheiro, J.C., 2000. Incremental quantile estimation for massive tracking. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 516–522.

Davison, A.C., 1984. Modelling excesses over high thresholds, with an application. In: Tiago de Lliveira, J. (Ed.), Statistical Extremes and Applications. D. Reidel, Dordrecht, pp. 461–482.

Davison, A.C., Smith, R.L., 1990. Models for exceedances over high thresholds. Journal of the Royal Statistical Society, Series B, Methodological 52, 393–442.

Dupuis, D., Tsao, M., 1998. A hybrid estimator for generalized pareto and extreme-value distributions. Communications in Statistics-Theory and Methods 27, 925–941.

Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modelling Extremal Events for Insurance and Finance. Springer-Verlag, Berlin Heidelberg.

He, X., Fung, W., 1999. Method of medians for life time data with weibull models. Statistics in Medicine 1993–2009.

Hill, B.M., 1975. A simple general approach to inference about the tail of a distribution. Annals of Statistics 1163–1174.

Hosking, J., Wallis, J., 1998. Parameters and quantile estimation for the generalized pareto distribution. Technometrics 29, 339–349.

Juarez, S., Schucany, W., 2004. Robust and efficient estimation for the generalized pareto distribution. Extremes 7, 237–251.

Liechty, J.C., Lin, D.K.J., Mcdermott, J.P., 2003. Single-pass low-storage arbitrar quantile estimation for massive datasets. Statistics and Computing 13, 91–100.

Manku, G.S., Rajagopalan, S., Lindsay, B.G., 1998. Approximate medians and other quantiles in one pass with limited memory. ACM SIGMOD Record 27 (2), 426–435.

McNeil, A.J., Saladin, T., 1997. The peaks over thresholds method for estimating high quantiles of loss distributions. In: Proceedings of 28th International ASTIN Colloquium.

Nelder, J., Mead, R., 1965. A simplex algorithm for function minimization. Computer Journal 7, 308–313.

Paterson, M.R., Progress in Selection. University of Warwick, Coventry, UK, 1997.

Pickands, J., 1975. Statistical inference using extreme order statistics. Annals of Statistics 3, 119–131.

R Development Core Team. 2010. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. http://www.R-project.org.

Ribatet, M., 2009. R package POT. R Foundation for Statistical Computing. http://r-forge.r-project.org/projects/pot/.

Rousseeuw, P.J., Bassett, G.W.J., 1990. The remedian: a robust averaing method for large data sets. Journal of the American Statistical Association 85 (409), 97–104.

Smith, R.L., 1985. Maximum likelihood estimation in a class of nonregular cases. Biometrika 67–90.

Zhang, J., 2007. Likelihood moment estimation for the generalized pareto distribution. Australian and New Zealand Journal of Statistics 49 (1), 69–77.

Zhang, J., 2010. Improving on estimation for the generalized pareto distribution. Technometrics 52, 335–339.

Zhang, J., Stephens, M., 2009. A new and efficient estimation method for the generalized pareto distribution. Technometrics 51, 316–325.