




Minimum Distance Estimation for the Generalized Pareto Distribution

Piao Chen, Zhi-Sheng Ye & Xingqiu Zhao


To cite this article: Piao Chen, Zhi-Sheng Ye & Xingqiu Zhao (2017) Minimum Distance Estimation for the Generalized Pareto Distribution, *Technometrics*, 59:4, 528-541, DOI: [10.1080/00401706.2016.1270857](https://doi.org/10.1080/00401706.2016.1270857)

To link to this article: <https://doi.org/10.1080/00401706.2016.1270857>

 View supplementary material 

 Accepted author version posted online: 20 Dec 2016.
Published online: 10 May 2017.

 Submit your article to this journal 

 Article views: 236

 View Crossmark data 

Minimum Distance Estimation for the Generalized Pareto Distribution

Piao Chen^a, Zhi-Sheng Ye^a, and Xingqiu Zhao^b

^aIndustrial and Systems Engineering, National University of Singapore, Singapore; ^bDepartment of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong

ABSTRACT

The generalized Pareto distribution (GPD) is widely used for extreme values over a threshold. Most existing methods for parameter estimation either perform unsatisfactorily when the shape parameter k is larger than 0.5, or they suffer from heavy computation as the sample size increases. In view of the fact that $k > 0.5$ is occasionally seen in numerous applications, including two illustrative examples used in this study, we remedy the deficiencies of existing methods by proposing two new estimators for the GPD parameters. The new estimators are inspired by the minimum distance estimation and the M -estimation in the linear regression. Through comprehensive simulation, the estimators are shown to perform well for all values of k under small and moderate sample sizes. They are comparable to the existing methods for $k < 0.5$ while perform much better for $k > 0.5$.

ARTICLE HISTORY

Received July 2016
Revised October 2016

KEYWORDS

Consistency; Extreme value;
 M -estimation; Peak over
threshold; Regression

1. Introduction

1.1 Background and Motivation

Extreme values are of interest in many statistical applications. For example, high-tide water levels are often collected to help design dikes, and extreme values of return series are useful for managing financial risks. Other examples where the extremes are of primary interest can be found in Finkenstadt and Rootzén (2003) and Castillo et al. (2005). A classical approach for analyzing the extreme values is based on the generalized extreme value distributions (Castillo and Hadi 1997). This family of distributions is found appropriate for the maximum of a sequence of independent and identically distributed (iid) random variables. Because only the maximum values are used in this approach, the information contained in other large sample values may be lost (Castillo and Hadi 1997). To make full use of all large observations, this classical approach is often extended to the peak over threshold (POT) method (Davis and Smith 1990). In the POT method, several of the largest order statistics are used instead of the maxima only. Usually, a threshold is set and peak values exceeding the threshold such as high-tide water levels are recorded. The differences between these peak values and the threshold are called exceedances over the threshold. Among all possible distributions for the exceedances, it is well recognized that the generalized Pareto distribution (GPD) is one of the most commonly used distributions (Davis and Smith 1990).

The cumulative distribution function (cdf) of a GPD with parameters $\theta = (k, \sigma)$ is

$$F_{\theta}(x) = \begin{cases} 1 - (1 - kx/\sigma)^{1/k}, & k \neq 0, \\ 1 - \exp(-x/\sigma), & k = 0, \end{cases} \quad (1)$$

where k is the shape parameter and $\sigma > 0$ is the scale parameter. The range of x is $x > 0$ for $k \leq 0$ and $0 < x < \sigma/k$ for $k > 0$. The GPD was first explicitly introduced by Pickands III (1975) as a distribution for the exceedances. Later, it is found that many distributions used for long-tailed data can be well approximated by a GPD (Choulakian and Stephens 2001). An interesting property of the GPD is that if $X \sim \text{GPD}(k, \sigma)$, then $(X - t) \sim \text{GPD}(k, \sigma - kt)$ given $X > t$ (Zhang and Stephens 2009). This property makes GPD a natural model for exceedances. In fact, the GPD is extensively used in various applications involving extreme values. Areas of applications include engineering (Zagorski and Wnek 2007), finance (Krehbiel and Adkins 2008), medical science (Davis and Feldstein 1979), environments (Moisello 2007), and behavioral science (Rootzén and Zholud 2016), to name a few.

One premise of successful applications of the GPD is an efficient inference procedure. Estimating the GPD parameters has a long history dating back to the early work of Gumbel (1958). Since then, numerous efforts have been found on this problem. However, there is still a lack of uniformly good methods for estimating the GPD parameters for the time being. Most existing methods either perform well only for certain values of k , or they become computationally burdensome as the sample size increases. A brief review of some popular inference methods for the GPD is provided in the next subsection.

1.2 Review of Existing Methods

Let x_1, \dots, x_n be n iid realizations of X , where $X \sim \text{GPD}(k, \sigma)$. The ML estimation for a GPD is first explicitly studied by Smith (1984). In the ML estimation, it is more convenient to deal with

the parameters (k, θ) , where $\theta = k/\sigma$. The ML estimator of θ is the solution of

$$1 - \frac{n}{\sum_i (1 - \theta x_i)^{-1}} + \frac{\sum_i \log(1 - \theta x_i)}{n} = 0, \quad \theta < 1/x_{(n)}, \quad (2)$$

where $x_{(n)}$ is the maximum among all the x_i 's. After $\hat{\theta}_{ML}$ is obtained, the ML estimates for k and σ can be obtained as

$$\hat{k}_{ML} = -1/n \sum_i \log(1 - \hat{\theta}_{ML} x_i), \quad \hat{\sigma}_{ML} = \hat{k}_{ML}/\hat{\theta}_{ML}. \quad (3)$$

When $k > 1$, the log-likelihood function can be made as large as possible by taking θ arbitrary close to $1/x_{(n)}$, and thus the ML estimators do not exist. In addition, when $0.5 < k < 1$, the ML estimators do not perform well (Grimshaw 1993; Castellanos and Cabras 2007). As a consequence, the ML method may be only applicable when $k < 0.5$, under which the asymptotic properties are well established in Smith (1984).

Since the ML estimation is not adequate in many situations, Hosking and Wallis (1987) proposed the method of moments (MOM) and the method of probability-weighted moments (PWM) for GPD parameters. The MOM uses the first two moments of a GPD. The MOM estimators of $\theta = (k, \sigma)$ based on an observed sample can be expressed as

$$\hat{k}_{MOM} = (\bar{x}^2/s^2 - 1)/2, \quad \hat{\sigma}_{MOM} = \bar{x}(\bar{x}^2/s^2 + 1)/2,$$

where \bar{x} and s^2 are the sample mean and sample variance, respectively. Alternatively, the PWM only make use of the first moment. The PWM estimators can be expressed as

$$\hat{k}_{PWM} = \bar{x}/(\bar{x} - 2u) - 2, \quad \hat{\sigma}_{PWM} = 2\bar{x}u/(\bar{x} - 2u),$$

where one choice of u is $u = \sum_i \frac{n-i}{n-1} x_{(i)}$, and $x_{(i)}$ is the i th-order statistic in a sample of size n (Zhang and Stephens 2009).

The first two moments of a GPD exist only when $k > -1$ and $k > -0.5$, respectively. Therefore, the application of MOM and PWM is restricted to these values of k . Although one can always obtain estimators for GPD parameters by the above formulas, it is not uncommon that the obtained estimates based on the MOM and PWM are inconsistent with the observed sample. For example, it is possible that $\hat{\sigma}/\hat{k} < x_{(n)}$ when $\hat{k} > 0$, which is inconsistent with the range of x when $k > 0$ (Zhang and Stephens 2009).

Since the performance of the conventional methods above is often not satisfactory, Zhang (2010) proposed to estimate the parameter $\theta = k/\sigma$ from a Bayesian perspective. The resulting estimators for θ is defined as $\hat{\theta}_{ZJ} = \sum_{j=1}^m w_j \theta_j$, where

$$\theta_j = \frac{n-1}{n+1} x_{(n)}^{-1} - \frac{\sigma^*}{k^*} \left[1 - \left(\frac{j-0.5}{m} \right)^{k^*} \right].$$

Here, $m = 20 + \lfloor \sqrt{n} \rfloor$ with $\lfloor \sqrt{n} \rfloor$ being the integer part of \sqrt{n} , and (k^*, σ^*) are prefixed by the sample. In addition, $w_j = 1/\sum_{k=1}^m \exp[l(\theta_k) - l(\theta_j)]$ is treated as weights, where $l(\theta)$ is the log-likelihood function based on θ . After having the value of $\hat{\theta}_{ZJ}$, one can obtain \hat{k}_{ZJ} and $\hat{\sigma}_{ZJ}$ by substituting $\hat{\theta}_{ZJ}$ into (3). This method (ZJ method hereafter) works well when $k < 0.5$, which

is the most commonly used region for k . However, its performance deteriorates sharply when k is outside this range. However, it is not uncommon to find that $k > 0.5$ in many practical examples (e.g., Castillo et al. 2005). In Section 4, two real applications are provided to further support this argument. On the other hand, when $k = 1$, the GPD degenerates to a uniform distribution on the support $[0, \sigma]$, which plays an important role in many statistical applications. In addition, Castillo and Hadi (1997) argued that at least from a theoretical viewpoint, estimating GPD for all possible values of its parameters is of interest. Another disadvantage of the ZJ method is that asymptotic properties of the estimators are unclear. According to our simulation in Section 3, these estimators might not be consistent for $k > 0.5$.

Castillo and Hadi (1997) proposed an inference procedure, called the elemental percentile method (EPM), for the GPD parameters. The EPM suffices for all ranges of k values. In the first step of EPM, two distinct order statistics $x_{(i)}$ and $x_{(j)}$ are extracted from the observed sample. By solving the equations

$$F_{\theta}(x_{(i)}) = i/(n+1) \quad \text{and} \quad F_{\theta}(x_{(j)}) = j/(n+1), \quad (4)$$

one can obtain an estimator of (k, σ) , denoted as $(\hat{k}(i, j), \hat{\sigma}(i, j))$. Since at most $n(n-1)/2$ pairs of such order statistics can be extracted from a sample of size n , the ensemble estimators are defined as the median of all the $n(n-1)/2$ estimators as

$$\hat{k}_{EPM} = \text{median}\{\hat{k}(1, 2), \dots, \hat{k}(n-1, n)\}, \\ \hat{\sigma}_{EPM} = \text{median}\{\hat{\sigma}(1, 2), \dots, \hat{\sigma}(n-1, n)\}.$$

Because the EPM needs to estimate the parameters $n(n-1)/2$ times, the computational cost is $O(n^2)$. This may be too high when a large sample is concerned. Though some random sampling schemes may help reduce the number of calculations, it is better to make full use of the data. In addition, asymptotic results for the ensemble estimator are quite difficult to obtain. The simulation results in Castillo and Hadi (1997) as well as our results in Section 3 indicate that there is still room for improvement in terms of estimating the shape parameter k .

1.3 Aims and Outline

In view of the importance of estimating the GPD parameters and the deficiencies in the existing methods, we propose two new estimators for the GPD parameters in this study. The key idea is to use the minimum distance estimation together with the M -estimation in a linear regression. The remainder of this article is organized as follows. In Section 2, we develop the new estimators of the GPD parameters. Asymptotic properties for the proposed estimators are also established. In addition, the bootstrap- t is used for interval estimation of the parameters. Simulation studies are conducted to compare different estimation methods in Section 3. Section 4 provides two illustrative examples. Section 5 concludes the article.

2. Proposed Methods

In this section, two new estimators for the GPD parameters are proposed based on the M -estimation. We first briefly review the

M -estimation method in a linear regression context. Next, we show that a simple regression model can be established for estimating the GPD and then the M -estimation procedure can be applied. Consistency of the proposed estimators is also established. The bootstrap- t method is then used for interval estimation of the GPD parameters are constructed. At last, we provide a simple optimization routine to obtain the proposed estimators.

2.1 M -Estimation in Linear Regression

Consider a linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i, \quad 1 \leq i \leq n,$$

where $y_i \in \mathbb{R}$ is the response variable, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^p$ are the explanatory variables, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T \in \mathbb{R}^p$ are the unknown parameters and e_i is the zero-mean error term. Let $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^T \boldsymbol{\beta}$ be the residuals. Then the least-square (LS) estimator of $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}_n = \arg \min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^n [r_i(\boldsymbol{\beta})]^2.$$

If the error terms e_i 's are iid normal and e_i is independent of \mathbf{x}_i , then the LS estimator is efficient. However, the LS estimator is sensitive to outliers in the sample. If the distribution of the error terms is not exactly normal, the estimator may be far from optimal. Therefore, many robust estimators have been proposed to achieve a balance between efficiency and stability in the presence of outliers. See, for example, the L -estimator (Jaeckel 1972), the R -estimator (Rousseeuw 1984), and the M -estimator (Huber 1973). For a book-length account on the robust estimators, readers are referred to Huber (2011).

Due to its generality and nice properties, for example, high efficiency, the M -estimator appears to be a popular robust method (Huber 2011, chap. 3). The basic idea of the M -estimation is to replace the squared residuals $[r_i(\boldsymbol{\beta})]^2$ by a less rapidly increasing function ρ of the residuals, leading to the M -estimator as

$$\tilde{\boldsymbol{\beta}}_n = \arg \min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^n \rho[r_i(\boldsymbol{\beta})].$$

In the literature, many ρ functions have been proposed (Rey 2012). One popular choice of the function ρ is the Tukey biweight function (Yohai and Zamar 1988) defined as

$$\rho_c(u) = \begin{cases} \frac{u^2}{2} \left(1 - \frac{u^2}{c^2} + \frac{u^4}{3c^4} \right) & |u| \leq c, \\ \frac{c^2}{6}, & |u| > c, \end{cases} \quad (5)$$

where c is called the tuning parameter. By setting $c = 4.6851$, the M -estimator with the Tukey biweight function has an efficiency of 95% under independent Gaussian errors. The consistency and asymptotic normality of an M -estimator for the linear regression with iid errors are well established in Huber (2011, chap. 11).

When e_i 's are heteroscedastic and suppose the variance of e_i is proportional to w_i^2 , it is natural to extend the M -estimator to

the weighted M -estimator defined as

$$\tilde{\boldsymbol{\beta}}_n^* = \arg \min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^n \rho[r_i^*(\boldsymbol{\beta})],$$

where $r_i^*(\boldsymbol{\beta}) = r_i(\boldsymbol{\beta})/w_i$.

2.2 New Estimators for the GPD

Consider an iid sample x_1, \dots, x_n from $\text{GPD}(k_0, \sigma_0)$ and let $\boldsymbol{\theta}_0 = (k_0, \sigma_0)$. Let $x_{(1)}, \dots, x_{(n)}$ be the corresponding order statistics of x_1, \dots, x_n . Let $F_n(x)$ be the corresponding empirical distribution function. At the discontinuity points $x_{(1)}, \dots, x_{(n)}$, we define $F_n(x_{(i)}) = (i - 0.5)/n$, $i = 1, \dots, n$, as a ‘‘continuity correction’’ to the empirical distribution (Zhang 2002). Although $F_n(x)$ is modified at the discontinuity points, in the Appendix we show that the Glivenko–Cantelli theorem still ensures $F_n(x) \rightarrow F_{\boldsymbol{\theta}_0}(x)$ almost surely, and the convergence is uniform in x . Intuitively, if we define residuals similar to those in a linear regression, that is,

$$r_i(\boldsymbol{\theta}) = F_n(x_i) - F_{\boldsymbol{\theta}}(x_i), \quad i = 1, \dots, n, \quad (6)$$

we may be able to obtain an estimator of $\boldsymbol{\theta}$ by minimizing a function $\rho(\cdot)$ (also called a distance measure) of $r_i(\boldsymbol{\theta})$.

This idea is known as the minimum distance estimation in the literature, which was explicitly developed by Wolfowitz (1957). Recent applications of the minimum distance estimation can be found in Ferger (2010), Hart and Cañette (2011), and Politis (2013). Obviously, an important task in the minimum distance estimation is to find an appropriate function that measures the distance between the empirical distribution function and the true one. Conventional distance measures are closely related to the goodness-of-fit statistics, for example, the Kolmogorov statistic and the Cramer–von Mises statistic. The estimators obtained from these distance measures are often satisfactory if the theoretical distribution function only involves the scale and location parameters, that is, the location-scale models (Parr and Schucany 1980). However, as stated by Huber (1972) and Parr and Schucany (1980), problems occur when the theoretical distribution involves a shape parameter. For the GPD, it is found that the estimators obtained by minimizing $\frac{1}{n} \sum_{i=1}^n [r_i(\boldsymbol{\theta})]^2$ are sensitive to the shape parameter k (Song and Song 2012). This is because at some values of k , some observations in the sample may have decisive impact in minimizing $\frac{1}{n} \sum_{i=1}^n [r_i(\boldsymbol{\theta})]^2$, especially when the sample size is small. Borrowing the idea from M -estimation, we may reduce the influence of these values in $r_i(\boldsymbol{\theta})$ by using an appropriate distance function $\rho(\cdot)$.

Following the M -estimation procedure in the last subsection, the M -estimator for $\boldsymbol{\theta}$ is defined as

$$\tilde{\boldsymbol{\theta}}_n = \arg \min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^n \rho[r_i(\boldsymbol{\theta})]. \quad (7)$$

Observe that $r_i(\boldsymbol{\theta}_0)$'s are asymptotically normal and have different asymptotic variances (Van der Vaart 1998, chap. 19), that is,

$$\sqrt{nr_i(\boldsymbol{\theta}_0)} = \sqrt{n}[F_n(x_i) - F_{\boldsymbol{\theta}_0}(x_i)] \xrightarrow{d} N[0, F_{\boldsymbol{\theta}_0}(x_i)(1 - F_{\boldsymbol{\theta}_0}(x_i))],$$

where $N[a, b]$ denotes a normal distribution with mean a and variance b . Therefore, we can construct a weighted M -estimator

as

$$\tilde{\theta}_n^* = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \rho[r_i^*(\theta)], \tag{8}$$

where $r_i^*(\theta) = r_i(\theta)/w_i$ and w_i 's are weights. In view of the asymptotic variance of $r_i(\theta)$, an appropriate weight can be $w_i(\theta) = \sqrt{F_{\theta}(x_i)(1 - F_{\theta}(x_i))}$. However, optimizing (8) with this weight is sometimes difficult, as $\rho(\cdot)$ becomes a very complicated function of the parameters θ . Alternatively, we may use the weight $w_i(\tilde{\theta}_n) = \sqrt{F_{\tilde{\theta}_n}(x_i)(1 - F_{\tilde{\theta}_n}(x_i))}$, where $\tilde{\theta}_n$ is the unweighted M -estimator obtained from (7). To use this weight requires estimating $\tilde{\theta}_n$ first, and thus $\tilde{\theta}_n^*$ can be regarded as a two-stage estimator.

2.3 Asymptotic Properties

This subsection shows that the two proposed estimators are consistent under some mild regularity conditions. The following assumption is made on the distance function $\rho(\cdot)$.

Assumption 1. Let ρ be a real function satisfying the following properties:

- (a) $\rho(0) = 0$ and $\rho(x) = \rho(-x)$;
- (b) if $0 < x_1 < x_2$, then $\rho(x_1) \leq \rho(x_2)$;
- (c) ρ is continuously differentiable.

Clearly, the Tukey biweight function defined in (5) satisfies Assumption 1. In addition, other popular ρ functions such as the Huber's function and the Cauchy's function also satisfy Assumption 1. For definitions of these ρ functions, we refer readers to Rey (2012). The following theorem establishes the consistency of $\tilde{\theta}_n$ and the proof is given in the Appendix.

Theorem 2.1. Assume Assumptions 1 holds. Consider an iid sample x_1, \dots, x_n from F_{θ_0} with θ_0 being the true parameters. Let $F_n(x)$ be the empirical distribution function and $r_i(\theta)$ the residual defined in (6). Consider the M -estimator $\tilde{\theta}_n$ given in (7). Then $\tilde{\theta}_n \rightarrow \theta_0$ in probability.

By assuming that the weight sequence $\{w_i\}$ satisfies Assumption 2 in the Appendix, the consistency of the weighted M -estimator $\tilde{\theta}_n^*$ is shown in the following theorem.

Theorem 2.2. Assume Assumptions 1–2 hold. Consider an iid sample x_1, \dots, x_n from F_{θ_0} with θ_0 being the true parameters. Let $F_n(x)$ be the empirical distribution function and $r_i(\theta)$ the residual defined in (6). Consider the weighted M -estimator given in (8). Then $\tilde{\theta}_n^* \rightarrow \theta_0$ in probability.

2.4 Confidence Interval

To capture uncertainties of the point estimation, another important task is to construct confidence intervals (e.g., Chen and Ye 2017). Unfortunately, interval estimation of the GPD parameters has not been well received in the literature. Because the ML estimators do not exist when $k > 1$ and the ML estimation performs poorly when $k > 0.5$, the application of the large-sample approximation to the ML estimators is restricted to $k < 0.5$. Alternatively, resampling-based methods such as the bootstrap were often used for interval estimation (Castillo and Hadi 1997).

In this section, the bootstrap- t method (Efron and Tibshirani 1994, chap. 12) is used with respect to different point estimation methods. The detailed procedures are summarized below.

1. For a given dataset of size n , estimate the parameter of k and σ by (7) (or (8)), denoted as \hat{k} and $\hat{\sigma}$; let $\widehat{se}_{\hat{k}}$ be the estimated standard error of \hat{k} .
2. Generate B bootstrap samples each with size n from $GPD(\hat{k}, \hat{\sigma})$.
3. For each bootstrap sample, estimate k by (7) (or (8)), denoted as \hat{k}_b and compute $t_b = (\hat{k}_b - \hat{k})/\widehat{se}_{\hat{k}_b}$, where $\widehat{se}_{\hat{k}_b}$ is the estimated standard error of \hat{k}_b .
4. The equal-tailed $100(1 - \alpha)\%$ confidence interval for k is $(\hat{k} - t_{1-\alpha/2}\widehat{se}_{\hat{k}}, \hat{k} + t_{\alpha/2}\widehat{se}_{\hat{k}})$, where t_{α} is the α percentile point of t_b 's.

Remark: Because the standard errors of \hat{k} and \hat{k}_b do not have closed forms, the bootstrap could again be invoked (Efron and Tibshirani 1994, chap. 6). For example, to estimate the standard error $\widehat{se}_{\hat{k}}$ of \hat{k} , we first generate 200 bootstrap samples each with size n from $GPD(\hat{k}, \hat{\sigma})$ and for each bootstrap sample, we estimate k by (7) (or (8)). Afterward, we will have 200 estimates of k , and then $\widehat{se}_{\hat{k}}$ is approximated by the sample standard deviation of these 200 estimates.

2.5 Optimization Method

Due to the complicated form of the distance function $\rho(\cdot)$, sometimes it may not be easy to obtain $\tilde{\theta}_n$ and $\tilde{\theta}_n^*$ by directly solving (7) and (8), respectively. In such cases, the method of iterative reweighted least squares (IRLS) (e.g., Green 1984) can be used to obtain these two estimators. Without loss of generality, we consider the IRLS algorithm for $\tilde{\theta}_n$ here, and $\tilde{\theta}_n^*$ can be obtained in a similar vein.

Let $\psi(u) = \partial\rho(u)/\partial u$, which is called the influence function in the context of robust statistics. If $\rho(u)$ satisfies Assumption 1 in the last subsection, then solving (7) is equivalent to solving the following two equations

$$\sum_{i=1}^n \psi[r_i(\theta)] \frac{\partial r_i(\theta)}{\partial \theta_j} = 0, \quad j = 1, 2, \tag{9}$$

where $\theta_1 = k$ and $\theta_2 = \sigma$. If we define $\phi(u) = \psi(u)/u$, then Equation (9) becomes

$$\sum_{i=1}^n \phi[r_i(\theta)] r_i(\theta) \frac{\partial r_i(\theta)}{\partial \theta_j} = 0, \quad j = 1, 2, \tag{10}$$

which is a typical problem that can be solved by the IRLS algorithm (e.g., Chaudhury 2013). In each iteration of the IRLS algorithm, we solve

$$\tilde{\theta}_n^{(p)} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \phi[r_i(\tilde{\theta}_n^{(p-1)})][r_i(\theta)]^2, \tag{11}$$

where p is the iteration number.

As can be seen from Equation (11), $\phi[r_i(\tilde{\theta}_n^{(p-1)})]$ is treated as known in each iteration. Therefore, the famous Gauss–Newton algorithm can be used to solve (11), which is much simpler than solving (7) directly. Our simulation experiences suggest

that the IRLS algorithm is more stable than direct maximization of (7). Estimators by the EPM and the ZJ method can be used as $\theta_n^{(0)}$, as they exist for all values of k . From a practical point of view, we would suggest the ZJ method, which is much less time-consuming. According to our simulation, if $\rho(u)$ is the Tukey biweight function defined in (5), then the sequence of $\tilde{\theta}_n^{(p)}$ always converges within 10 iterations in our simulation trials.

3. Simulation

3.1 Point Estimation

In this subsection, a simulation study is conducted to compare different estimation methods for the GPD parameters. Specifically, the proposed methods in the last section are compared with the ML method, the EPM and the ZJ method. The implementation of these methods can be found in Section 1.2. In

the proposed methods, we use the Tukey biweight function defined in (5) with the tuning parameter $c = 4.6851$. According to our preliminary simulation results (not reported), other popular ρ functions such as the Huber's function have a similar performance with the Tukey biweight function. The method of moments and the method of probability-weighted moments are not considered here because they are not better than the EPM in almost all the cases considered in Castillo and Hadi (1997). Because the simulation results are invariant of the scale parameter σ , we set $\sigma = 1$ throughout this section. A wide range of k values are considered, that is, $k = -4, -3.9, -3.8, \dots, 3.9, 4$. Because the ML estimators exit for $k \leq 1$, such k values are used for the ML method. When the estimated k value is greater than one, the ML estimates are set as $(1, x_{(n)})$, as suggested by Zhang and Stephens (2009). Sample sizes $n = 20, 50, 100$ are considered. For each combination of k, σ , and n , biases and mean squared errors (MSEs) of the estimators are obtained based

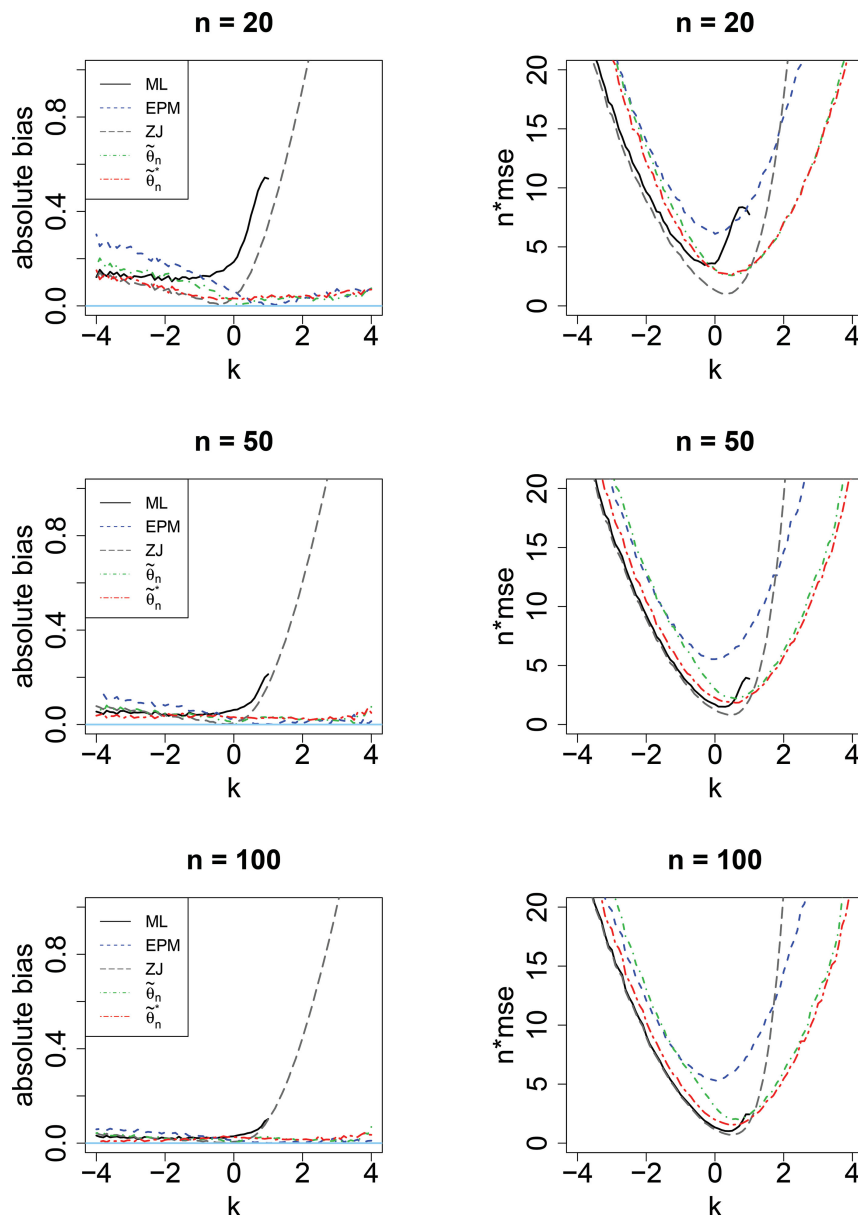


Figure 1. Absolute value of bias and $n \times \text{MSE}$ for the shape parameter k based on the ML method, the elemental percentile method (EPM), the method in Zhang (2010) (ZJ), and the two proposed estimators ($\tilde{\theta}_n$ and $\hat{\theta}_n$).

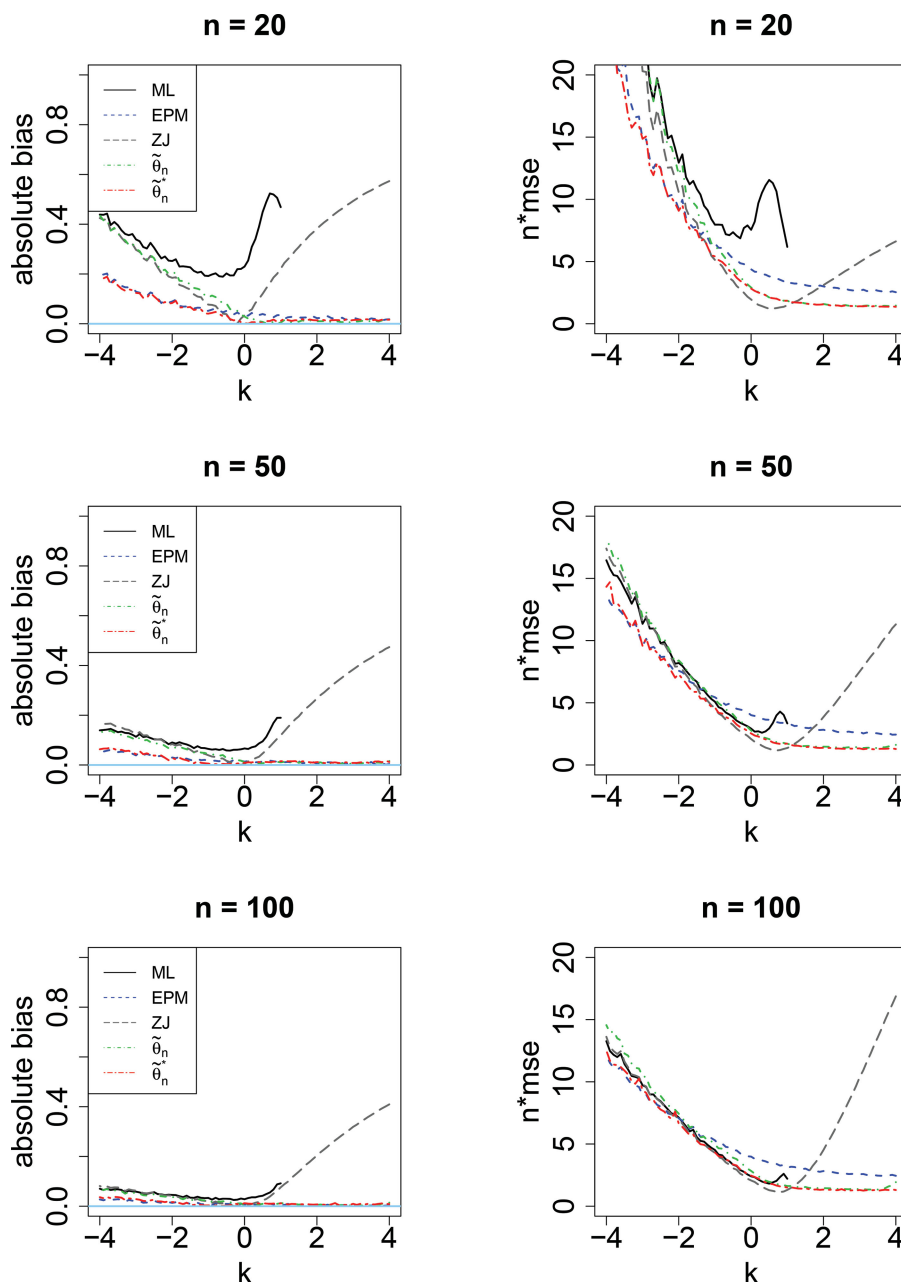


Figure 2. Absolute value of bias and $n \times \text{MSE}$ for the scale parameter σ based on the ML method, the elemental percentile method (EPM), the method in Zhang (2010) (ZJ), and the two proposed estimators ($\hat{\theta}_n^*$ and $\hat{\theta}_n^w$).

on 10,000 Monte Carlo replications. The results for the shape parameter k and the scale parameter σ are shown in Figures 1 and 2, respectively.

As can be seen from these two figures, the two proposed estimators seem to work well overall, as small biases and MSEs are achieved in all the scenarios. Between these two estimators, the weighted M -estimator performs slightly better. As far as the shape parameter k is concerned, the proposed methods are comparable with the ZJ method in the commonly used range $k < 0.5$, while they perform much better when $k > 0.5$. In addition, the proposed methods have a much better performance for estimating the scale parameter σ than the ZJ method. The EPM works well in terms of the biases but the MSEs seem a bit large for both parameters. Moreover, the application of EPM generally requires more computing time than the other methods (not

reported). As for the ML method, it does not work well when $0.5 < k < 1$, which is consistent with the findings in Castellanos and Cabras (2007).

3.2 Interval Estimation

This subsection compares the performance of the bootstrap- t method based on different point estimation procedures, that is, the ML method, the EPM, the ZJ method, and the two proposed estimators. The comparison is based on the coverage probabilities. Because the implementation of the bootstrap requires the existence of the point estimator and the ML estimator only exists for $k \leq 1$, we consider $k = -1, -0.5, 0, 0.5, 1$ here. In fact, the range $-1 < k < 0.5$ is most commonly observed in practical applications (Zhang and Stephens 2009). The scale parameter σ

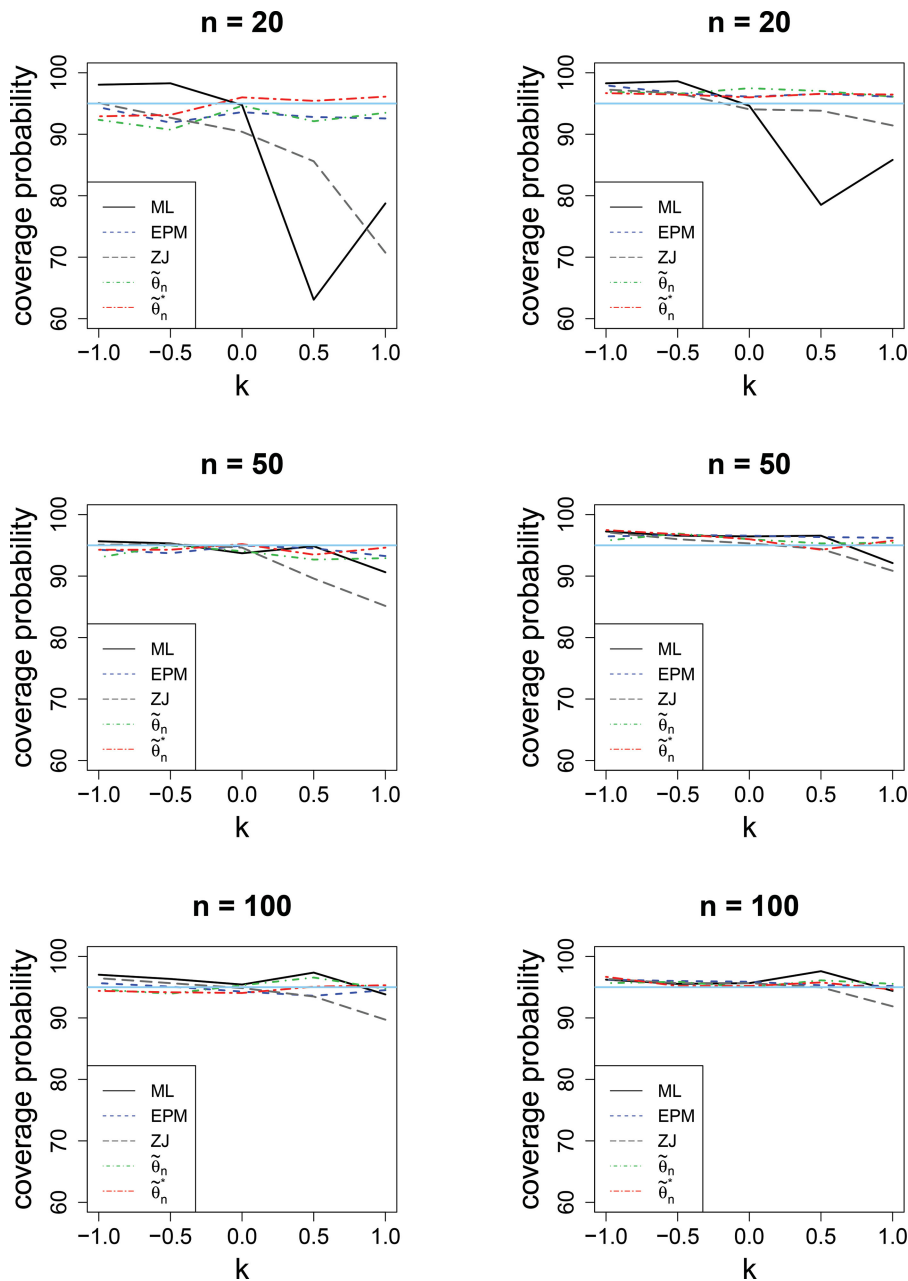


Figure 3. Coverage probabilities of shape parameter k (left) and σ (right) based on ML method, elemental percentile method (EPM), method in Zhang (2010) (ZJ), and the proposed two methods ($\hat{\theta}_n$ and $\hat{\theta}_n^*$).

is set as 1. In addition, sample sizes $n = 20, 50, 100$ are used. We consider the 95% confidence intervals for k and σ . The numbers of both bootstrap samples and Monte Carlo replications are set as 10,000. The coverage probabilities by different methods are shown in Figure 3. As can be seen, the bootstrap- t based on the two proposed estimators and EPM seem to work uniformly well, and their performance improves with the sample size n . As far as the ML method and the ZJ method are concerned, both methods do not work well when $k > 0$ or $n < 100$.

3.3 Censored Data

The GPD is a popular model for extreme values over a threshold. Since extreme values are often seen in environmental sciences, financial time series, insurance industry, and hydraulics

engineering, the GPD is extensively applied in these areas, where most of the recorded data are complete. For example, see the various datasets of extreme values in the books by Coles et al. (2001); Finkenstadt and Rootzén (2003); Castillo et al. (2005); De Haan and Ferreira (2007). As a result, most GPD-related studies, including estimation and applications, are based on complete samples. The simulations in the last subsections showed that the GPD parameters based on a complete sample can be accurately estimated by our proposed methods. On the other hand, sometimes, the GPD is also found useful in the reliability area, where censoring is commonly seen. Because the empirical distribution function in (6) can be computed based on a censored sample, our proposed methods can also be used in the presence of censoring. For demonstration, consider Type II censoring where only the first m failures out of the n units

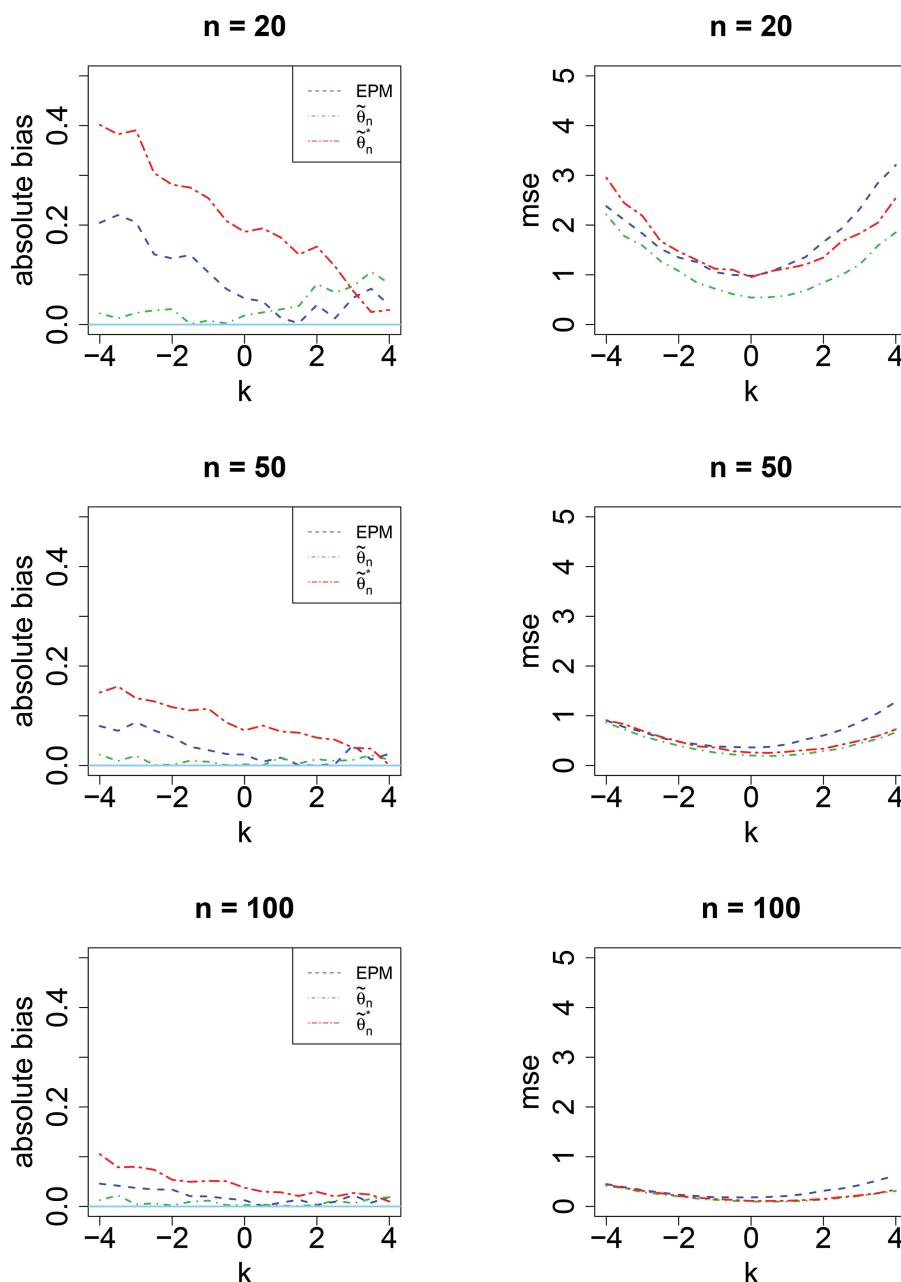


Figure 4. Absolute value of bias and MSE for the shape parameter k based on the elemental percentile method (EPM) and the two proposed estimators ($\tilde{\theta}_n$ and $\tilde{\theta}_n^*$) when the censoring rate is 20%.

are observed. Let $x_1 < x_2 < \dots < x_m$ be the ordered failure times. The empirical distribution at time x_i can be computed as $F_n(x_i) = (i - 0.5)/n$, $i = 1, \dots, m$. Then the M -estimator can be obtained as

$$\tilde{\theta}_n = \arg \min_{\theta} \frac{1}{m} \sum_{i=1}^m \rho[r_i(\theta)],$$

and the weighted M -estimator can be obtained as

$$\tilde{\theta}_n^* = \arg \min_{\theta} \frac{1}{m} \sum_{i=1}^m \rho[r_i^*(\theta)],$$

where $r_i^*(\theta) = r_i(\theta) / \sqrt{F_{\tilde{\theta}_n}(x_i)(1 - F_{\tilde{\theta}_n}(x_i))}$, $i = 1, \dots, m$. By similar procedure, the proposed methods can also be used to

deal with other types of censoring such as left censoring and interval censoring (Meeker and Escobar 1998).

In this subsection, a simulation is conducted to compare the proposed methods with the EPM based on the Type II censored data. The EPM is applicable because Equation (4) can be solved based on the censored sample, while other methods such as the MOM and the ZJ method cannot be easily extended to deal with censoring. Similar to the previous settings, we consider $n = 20, 50, 100$, $k = -4, -3.9, -3.8, \dots, 3.9, 4$, and $\sigma = 1$. Under each sample size n , four censoring rates $\gamma = 10\%, 20\%, 30\%$, and 40% are considered by carefully choosing m . Based on 10,000 Monte Carlo replications, the absolute values of biases and MSEs for the shape parameter k and the scale parameter σ when $\gamma = 20\%$ are shown in Figures 4 and 5, respectively. Other simulation results when $\gamma = 10\%, 30\%$, and 40% are given in the supplementary material. As we can

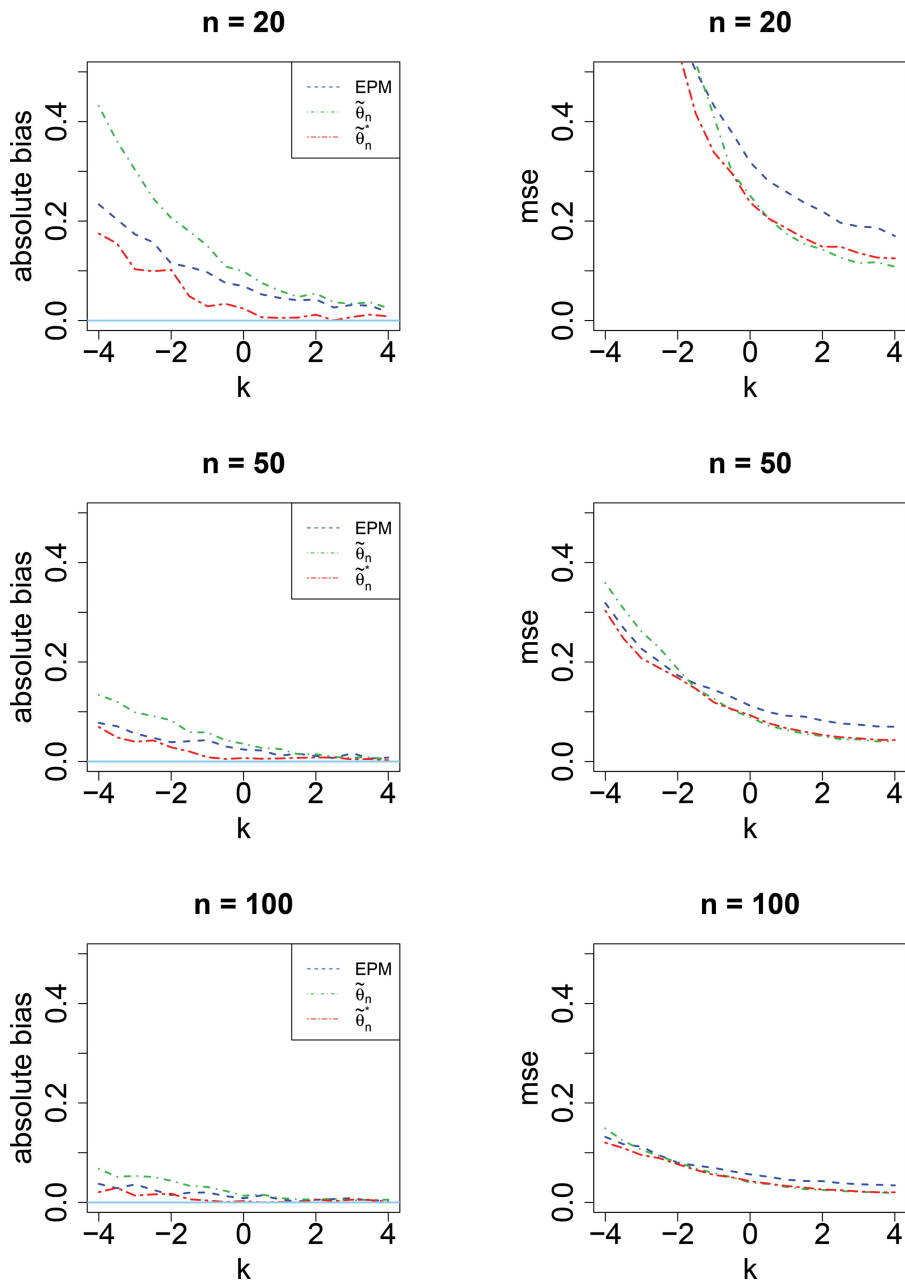


Figure 5. Absolute value of bias and MSE for the scale parameter σ based on the elemental percentile method (EPM) and the two proposed estimators ($\hat{\theta}_n$ and $\hat{\theta}_n^*$) when the censoring rate is 20%.

see, our proposed methods generally compete well with the EPM when $k < 0.5$ and they have a better performance than the EPM when $k > 0.5$. As with the complete sample cases, the EPM is more time-consuming compared with the proposed methods.

4. Example

In this section, two real examples in Castillo and Hadi (1997) are revisited. In each example, the generalized Pareto distribution is found to provide a good fit to the data. For both examples, all the estimated values of k are larger than 0.5 and sometimes even larger than 1. This finding affirms our previous arguments that $k > 0.5$ is not uncommon in practical applications.

4.1 Bilbao Waves Data

The first example consists of the zero-crossing hourly mean periods (in seconds) of the sea waves measured in Bilbao buoy, Spain. The data are used for studying the influence of periods on beach morphodynamics and other properties related to the right tail (Castillo and Hadi 1997). Only data above 7 sec are listed in Table 1. Fitting GPD to this dataset has been well studied in the literature (e.g., Castillo and Hadi 1997; Luceño 2006; Zhang and Stephens 2009; Del Castillo and Serra 2015). It is found that when the threshold $t \geq 7.5$, the GPD fits the exceedances very well (Zhang and Stephens 2009). Table 2 shows the estimates of the two GPD parameters based on the ML estimation, the EPM, the ZJ method, and the proposed methods, when the threshold time t is set as 7.5, 8.0, 8.5, 9.0, and 9.5, respectively. Note that the ML estimators do not exist when

Table 1. The Bilbao waves data: the zero-crossing hourly mean periods (in seconds), above 7 sec, of the sea waves measured in Bilbao buoy.

7.05	7.26	7.46	7.59	7.69	7.82	7.90	7.97	8.11	8.21	8.40	8.51	8.69	8.85
9.06	9.23	9.46	9.75	9.12	9.24	9.47	9.78	9.16	9.27	9.59	9.79	9.43	9.74
7.12	7.27	7.46	7.59	7.72	7.83	7.91	7.99	8.12	8.23	8.41	8.52	8.71	8.86
7.15	7.28	7.47	7.61	7.72	7.83	7.93	8.00	8.15	8.23	8.42	8.53	8.72	8.88
7.18	7.30	7.48	7.63	7.72	7.83	7.93	8.03	8.15	8.30	8.43	8.54	8.74	8.88
9.17	9.29	9.59	9.79	9.17	9.30	9.60	9.80	9.18	9.32	9.61	9.84	9.22	9.90
7.19	7.31	7.48	7.65	7.72	7.84	7.93	8.03	8.15	8.30	8.43	8.56	8.74	8.94
7.20	7.31	7.52	7.66	7.72	7.85	7.94	8.05	8.18	8.31	8.45	8.58	8.74	8.98
7.20	7.32	7.54	7.66	7.77	7.85	7.95	8.06	8.18	8.31	8.48	8.59	8.74	8.98
7.20	7.33	7.55	7.67	7.77	7.88	7.95	8.06	8.18	8.32	8.49	8.59	8.79	8.99
7.20	7.37	7.55	7.67	7.79	7.88	7.97	8.07	8.19	8.32	8.50	8.60	8.81	9.01
7.25	7.40	7.58	7.68	7.79	7.90	7.97	8.10	8.20	8.33	8.50	8.65	8.84	9.03
9.18	9.33	9.62	9.85	9.18	9.36	9.63	9.89	9.21	9.38	9.66			

Table 2. Estimates of (k, σ) for the Bilbao waves data based on the ML method (ML), the elemental percentile method (EPM), the method in Zhang (2010) (ZJ), and the proposed two estimators $(\tilde{\theta}_n$ and $\tilde{\theta}_n^*$): t is the threshold time and m is the number of exceedances; NA means nonexistence.

t	m	ML	EPM	ZJ	$\tilde{\theta}_n$	$\tilde{\theta}_n^*$
7.5	154	(0.768, 1.860)	(0.789, 1.995)	(0.686, 1.722)	(0.567, 1.583)	(0.602, 1.621)
8.0	106	(0.864, 1.648)	(0.787, 1.611)	(0.731, 1.462)	(0.638, 1.384)	(0.668, 1.406)
8.5	69	NA	(0.617, 1.104)	(0.767, 1.146)	(0.763, 1.163)	(0.771, 1.165)
9.0	41	NA	(0.833, 0.819)	(0.760, 0.756)	(0.806, 0.802)	(0.877, 0.836)
9.5	17	NA	(1.585, 0.626)	(0.736, 0.361)	(1.291, 0.518)	(1.274, 0.515)

Table 3. 95% confidence intervals of k and σ for the Bilbao waves data. t is the threshold time and m is the number of exceedances.

t	m	\tilde{k}	$\tilde{\sigma}$	\tilde{k}^*	$\tilde{\sigma}^*$
7.5	154	(0.316, 0.813)	(1.204, 1.952)	(0.403, 0.813)	(1.262, 1.982)
8.0	106	(0.308, 0.964)	(0.966, 1.789)	(0.400, 0.947)	(1.014, 1.797)
8.5	69	(0.355, 1.162)	(0.746, 1.565)	(0.420, 1.137)	(0.762, 1.565)
9.0	41	(0.240, 1.350)	(0.411, 1.171)	(0.387, 1.364)	(0.463, 1.196)
9.5	17	(0.618, 1.901)	(0.252, 0.758)	(0.601, 1.873)	(0.253, 0.761)

Table 4. Average scaled absolute errors (ASAEs) for the Bilbao waves data based on the ML method (ML), the elemental percentile method (EPM), the method in Zhang (2010) (ZJ), and the proposed two estimators $(\tilde{\theta}_n$ and $\tilde{\theta}_n^*$): t is the threshold time and m is the number of exceedances; NA means nonexistence.

t	m	ML	EPM	ZJ	$\tilde{\theta}_n$	$\tilde{\theta}_n^*$
7.5	154	0.0262	0.0444	0.0169	0.0121	0.0126
8.0	106	0.0307	0.0324	0.0152	0.0133	0.0134
8.5	69	NA	0.0283	0.0188	0.0179	0.0180
9.0	41	NA	0.0325	0.0333	0.0320	0.0327
9.5	17	NA	0.0800	0.0974	0.0631	0.0629

$t = 8.5, 9.0,$ and 9.5 . As can be seen, the estimates of k by all the methods under every threshold is larger than 0.5. In some cases, the estimated k is even larger than 1. Table 3 shows the 95% confidence intervals of k and σ based on the bootstrap- t method.

Table 5. Fatigue data: Kevlar/epoxy strand-life data in 10^4 hr tested at 70% stress level.

0.1051	0.1337	0.1389	0.1921	0.1942	0.2322	0.3629	0.4006	0.4012	0.4063	0.4921
0.5445	0.5620	0.5817	0.7886	0.8108	0.8546	0.8666	0.8831	0.9106	0.9711	0.9806
1.0205	1.0396	1.0861	1.1026	1.1214	1.1362	1.1604	1.1608	1.1745	1.1762	1.1895
1.4110	1.4496	1.5395	1.6179	1.7092	1.7568	1.7568	0.5905	0.5956	0.6068	0.6121
1.2044	1.3520	1.3670	0.6473	0.7501						

To measure the overall goodness of fit of these methods, we use the average scaled absolute error (ASAE) defined in Castillo and Hadi (1997) as

$$ASAE = \frac{1}{n} \sum_{i=1}^n \frac{|x_{(i)} - \hat{x}_{(i)}|}{x_{(n)} - x_{(1)}}, \tag{12}$$

where $\hat{x}_{(i)} = \hat{\sigma} [1 - (1 - i/(n + 1))^{\hat{k}}] / \hat{k}$ with \hat{k} and $\hat{\sigma}$ the estimates by different estimation methods. The ASAE values for the ML method, the EPM, the ZJ method, and the two proposed methods under different threshold are shown in Table 4. As can be seen, the ASAE values by the two proposed estimators are the smallest under every threshold. The results indicate better performance of the proposed estimators.

4.2 Fatigue Data

The second example concerns the fatigue data for the Kevlar/epoxy strand lifetime, which are reported by Andrews and Herzberg (1985) and reproduced in Table 5. The purpose of this dataset is to estimate the small quantiles of the lifetime and hence the interest is in the left tail. To apply the GPD, we shall transform the lower tail to the upper tail by considering negative values of these data, as suggested by Castillo and Hadi (1997). Based on this transformation, we are able to analyze exceedances over a prefixed threshold by the GPD.

We first need to determine the threshold over which the GPD provides a good fit to the exceedances. Two statistics proposed by Choulakian and Stephens (2001) for a GPD, that is, the Cramer-von Mises statistic W^2 and the Anderson-Darling statistic A^2 , are used here. Following Choulakian and Stephens (2001), these two statistics can be calculated as

$$W^2 = \sum_{i=1}^n [z_i - (2i - 1)/(2n)]^2 + 1/(12n)$$

and

$$A^2 = -n - (1/n) \sum_{i=1}^n (2i - 1) [\log(z_i) + \log(1 - z_{n+1-i})],$$

where $z_i = F_{\hat{\theta}}(x_{(i)})$ and $\hat{\theta}$ is the estimate of θ based on different estimation methods. We consider thresholds $t = -1.8, -1.6, -1.4, -1.2, -1.0,$ and -0.8 as in Castillo and Hadi (1997). Because the ML estimator does not exist in all these values of thresholds, we only consider the EPM, the ZJ method, and the proposed methods here. For each estimation method, the parametric bootstrap is used to obtain the p -values of the statistics W^2 and A^2 . As an example, consider the Cramer-von Mises statistic W^2 and the EPM with threshold t . Also denote m as the number of exceedances over t . We first use the EPM

Table 6. p -values (p_w, p_a) of the Cramer–von Mises statistic (p_w) and the Anderson–Darling statistic (p_a) for the elemental percentile method (EPM), the method in Zhang (2010) (ZJ), and the two proposed methods ($\hat{\theta}_n$ and $\hat{\theta}_n^*$) based on the fatigue data: t is the threshold time and m is the number of exceedances.

t	m	EPM	ZJ	$\hat{\theta}_n$	$\hat{\theta}_n^*$
-1.8	49	(0.266, 0.465)	(0.066, 0.085)	(0.109, 0.145)	(0.114, 0.224)
-1.6	45	(0.239, 0.466)	(0.122, 0.083)	(0.105, 0.131)	(0.136, 0.157)
-1.4	42	(0.578, 0.382)	(0.505, 0.408)	(0.495, 0.619)	(0.561, 0.536)
-1.2	39	(0.093, 0.214)	(0.531, 0.390)	(0.346, 0.437)	(0.293, 0.376)
-1.0	28	(0.664, 0.857)	(0.669, 0.756)	(0.628, 0.855)	(0.722, 0.855)
-0.8	21	(0.398, 0.623)	(0.360, 0.376)	(0.287, 0.554)	(0.318, 0.458)

Table 7. Estimates of (k, σ) based on the elemental percentile method (EPM), the method in Zhang (2010) (ZJ), and the two proposed estimators ($\hat{\theta}_n$ and $\hat{\theta}_n^*$) for the fatigue data: t is the threshold time and m is the number of exceedances.

t	m	EPM	ZJ	$\hat{\theta}_n$	$\hat{\theta}_n^*$
-1.4	42	(0.948, 1.219)	(0.748, 1.070)	(0.910, 1.210)	(0.908, 1.217)
-1.2	39	(0.020, 0.546)	(0.599, 0.767)	(0.523, 0.756)	(0.441, 0.705)
-1.0	28	(0.749, 0.760)	(0.675, 0.709)	(0.844, 0.815)	(0.849, 0.820)
-0.8	21	(0.798, 0.625)	(0.640, 0.550)	(0.861, 0.657)	(0.894, 0.673)

to obtain the estimate $\hat{\theta}$. Then, 1000 samples of size m are generated from the GPD with parameter $\hat{\theta}$. From each sample, we can obtain the corresponding value of W^2 . The p -value is then defined as the percentage of these W^2 's no less than the W^2 based on the original sample. The p -values are listed in Table 6. As we can see, when $t \geq -1.4$, the GPD seems to fit the exceedances very well. For these values of threshold t , the estimates of θ

Table 8. 95% confidence intervals of k and σ for the fatigue data. t is the threshold time and m is the number of exceedances.

t	m	\tilde{k}	$\tilde{\sigma}$	\tilde{k}^*	$\tilde{\sigma}^*$
-1.4	42	(0.490, 1.316)	(0.753, 1.641)	(0.591, 1.213)	(0.741, 1.652)
-1.2	39	(-0.202, 1.231)	(0.282, 1.210)	(-0.120, 1.115)	(0.288, 1.152)
-1.0	28	(0.281, 1.384)	(0.407, 1.198)	(0.312, 1.393)	(0.398, 1.224)
-0.8	21	(0.241, 1.446)	(0.284, 1.001)	(0.251, 1.553)	(0.274, 1.060)

Table 9. Average scaled absolute errors (ASAEs) for the fatigue data based on the elemental percentile method (EPM), the method in Zhang (2010) (ZJ), and the two proposed estimators ($\hat{\theta}_n$ and $\hat{\theta}_n^*$): t is the threshold time and m is the number of exceedances.

t	m	EPM	ZJ	$\hat{\theta}_n$	$\hat{\theta}_n^*$
-1.4	42	0.0271	0.0302	0.0249	0.0255
-1.2	39	0.0860	0.0355	0.0334	0.0366
-1.0	28	0.0311	0.0359	0.0310	0.0314
-0.8	21	0.0465	0.0507	0.0471	0.0485

based on the EPM, the ZJ method, and two proposed estimators are shown in Table 7. We again find that $\hat{k} > 0.5$ occurs quite often. The 95% confidence intervals of k and σ by the bootstrap- t method are shown in Table 8.

Similar to the foregoing Bilbao waves example, we use the average scaled absolute errors defined in (12) to measure the performance of these methods. The values of ASAE are shown in Table 9. As expected, overall our proposed methods have the smallest ASAE. On the other hand, we can display the empirical distribution function and the estimated distribution function on

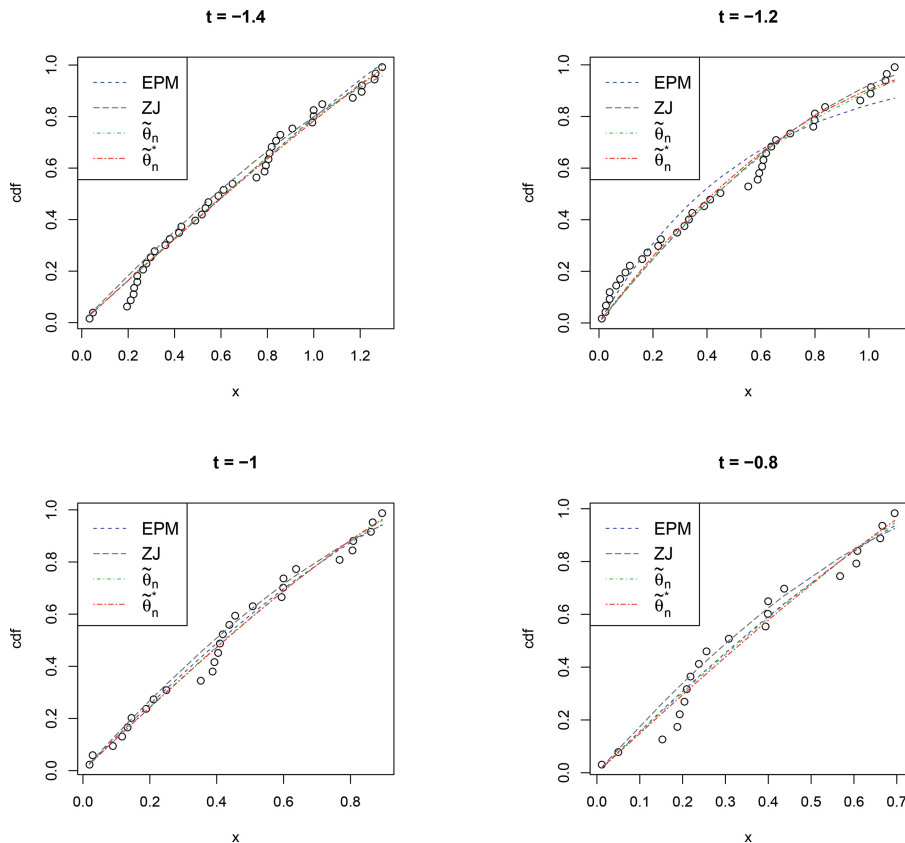


Figure 6. Plots of the empirical distribution function versus the estimated distribution functions based on the elemental percentile method (EPM), the method in Zhang (2010) (ZJ), and the two proposed estimators ($\hat{\theta}_n$ and $\hat{\theta}_n^*$).

the same plot to graphically check the fit (Zhang and Stephens 2009), as shown in Figure 6. We can see from this figure that the proposed methods fit the fatigue data better in the upper tail, which is a desirable property in real applications (Zhang and Stephens 2009).

5. Conclusion

The GPD is one of the most important distributions in applications involving extreme values. Because the ML estimation is applicable only when the shape parameter $k > 1$ and it performs poorly for small sample sizes, many other methods have been proposed for estimating the GPD parameters. However, most of these methods only work well for $k < 0.5$. In this article, we proposed two new estimators for the GPD parameters using the minimum distance estimation, where the Turkey biweight function is used as the distance measure. The first estimator minimizes the distance between the empirical distribution and the family of GPDs, while the second one is a weighted version of the first one, where the weight is computed based on the first estimator. The two estimators were shown to be consistent. Confidence intervals of the GPD parameters were successfully constructed based on the bootstrap- t method. Our simulations and illustrative examples show that the proposed methods not only compete well with the existing methods for $k < 0.5$ but also perform much better when $k > 0.5$.

Because the inference procedure is quite general, the proposed methods should also work well for other parametric models such as the Gompertz distribution where the ML estimators may not exist. Following our proofs in the Appendix, the asymptotic properties hold as long as if the cdfs of these parametric models satisfy some mild regularity conditions. On the other hand, because the distance measure $\rho(\cdot)$ in this study is borrowed from robust estimation, the proposed methods are robust to outlier contamination and the breakdown point is as high as 50%. Nevertheless, the robustness to outliers was not highlighted in this study, as our purpose is to develop a good estimation procedure for the GPD that is valid for all range of k . For applications involving extreme values, it is often seen that values of exceedances are small (De Zea Bermudez and Kotz 2010). In such cases, a single abnormal large value may damage the estimation procedure. With the M -estimation in our study, however, the effect of outliers can be greatly reduced. In this sense, our proposed estimation methods are more robust than the existing methods.

Appendix: Proofs

Here, we only present the proof of Theorem 2.1 since Theorem 2.2 can be proved similarly. Define $r_n(x, \theta) = F_n(x) - F(x; \theta)$ and $r(x, \theta) = F(x; \theta_0) - F(x; \theta)$. We first show that $r(x, \theta)$ is Donsker.

Lemma A.1. The class of functions $r(x, \theta)$ indexed by θ is Donsker.

Proof. First note that the total variation of functions $F(x; \theta)$ is bounded by 1, and hence $F(x; \theta)$ is Donsker (Van der Vaart and Wellner 1996, pp. 191). We can then treat $r(x, \theta)$ as the class of functions $x \mapsto r(F(x; \theta))$ with θ ranging over the parameter space Θ . It is easy to see that $|r(x, \theta_1) - r(x, \theta_2)|^2 = (F(x; \theta_1) - F(x; \theta_2))^2$ for every $\theta_1, \theta_2 \in \Theta$, and x . According to Theorem 2.10.6 in Van der Vaart and Wellner (1996), it suffices to show

that there exists $\theta \in \Theta$ such that $\int r^2(x, \theta) dx < \infty$. In fact, if we let $\theta = \theta_0$, then $\int r^2(x, \theta) dx = 0 < \infty$. This completes the proof. \square

Proof of Theorem 2.1. Let

$$M_n^*(\theta) = -\frac{1}{n} \sum_{i=1}^n \rho(r_n(x_i, \theta)) \quad \text{and}$$

$$M_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \rho(r(x_i, \theta)).$$

A first-order Taylor expansion shows that

$$\begin{aligned} & \sup_{\theta} |M_n^*(\theta) - M_n(\theta)| \\ & \leq \frac{1}{n} \sum_{i=1}^n \sup_{\theta} |\rho(r_n(x_i, \theta)) - \rho(r(x_i, \theta))| \\ & = \frac{1}{n} \sum_{i=1}^n \sup_{\theta} |[r_n(x_i, \theta) - r(x_i, \theta)]\rho'(\xi_i(\theta))|, \end{aligned}$$

where $\xi_i(\theta)$ is between $r(x_i, \theta)$ and $r_n(x_i, \theta)$. The first term inside the supremum is exactly $F_n(x) - F(x; \theta_0)$. Although $F_n(x)$ is modified at the discontinuity points x_1, \dots, x_n , if we denote $\tilde{F}_n(x)$ as the conventional empirical distribution function, it is clear that $\sup_x |F_n(x) - \tilde{F}_n(x)| = 0.5/n$, which converges to 0 when $n \rightarrow \infty$. Therefore, the Glivenko–Cantelli theorem still holds to the modified $F_n(x)$, that is, $\sup_x |F_n(x) - F(x; \theta_0)| \rightarrow 0$ as $n \rightarrow \infty$. This means the first term is uniformly $o_p(1)$. The second term is bounded by $\max\{\rho'(\xi) : -1 \leq \xi \leq 1\}$. Therefore,

$$\sup_{\theta} |M_n^*(\theta) - M_n(\theta)| = o_p(1). \quad (\text{A.1})$$

Let $M(\theta) = -E\{\rho(r(x, \theta))\}$. Next, we will use Corollary 3.2.3 of Van der Vaart and Wellner (1996) to establish the theorem. To use this corollary, we need to verify the following two conditions:

- (i) $\sup_{\theta} |M_n^*(\theta) - M(\theta)| = o_p(1)$;
- (ii) $M(\theta_0) > \sup_{\theta \notin G} M(\theta)$ for every open set G that contains θ_0 .

Assume X_1, \dots, X_n are iid with measure P and values in \mathcal{X} , and P_n is the empirical measure of X_i 's. Given a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, let $P_n f = \frac{1}{n} \sum_{i=1}^n f(X_i)$ and $P f = \int f dP$. Theorem 2.10.6 in Van der Vaart and Wellner (1996) ensures that $\rho(r(x, \theta))$ is Donsker as $r(x, \theta)$ is Donsker (Lemma A.1) and $\rho(\cdot)$ is a Lipschitz function (since $\rho(\cdot)$ is assumed continuously differentiable in Assumption 1). Since $M(\theta) - M_n(\theta) = (P_n - P)\rho(r(x, \theta))$, the Clivenko–Cantelli theorem ensures

$$\sup_{\theta} |M_n(\theta) - M(\theta)| = o_p(1). \quad (\text{A.2})$$

Condition (i) then follows from (A.1) and (A.2) as

$$\begin{aligned} \sup_{\theta} |M_n^*(\theta) - M(\theta)| & \leq \sup_{\theta} |M_n^*(\theta) - M_n(\theta)| \\ & \quad + \sup_{\theta} |M_n(\theta) - M(\theta)|. \end{aligned}$$

Condition (ii) holds since $M(\theta_0) = 0$ and $\sup_{\theta \notin G} M(\theta) < 0$. Thus, we conclude $\tilde{\theta}_n \rightarrow \theta_0$ in probability. \square

Note that the consistency of the weighted M -estimator $\tilde{\theta}_n^*$ can be similarly established as long as the weight sequence $\{w_i\}$ satisfies the following Assumption.

Assumption 2. There exists a constant $\alpha > 0$ such that

$$\limsup_{n \rightarrow \infty} \sup_{1 \leq i \leq n} w_i < \alpha \text{ a.s.}$$

and for $r_i(\theta)$ defined in (6)

$$\lim_{n \rightarrow \infty} \sup_{1 \leq i \leq n} |r_i(\theta_0)/w_i| = 0 \text{ a.s.}$$

Given the consistency of the M -estimator $\tilde{\theta}_n$, Einmahl (1989) showed that our constructed weight sequence $w_i(\tilde{\theta}_n) = \sqrt{F(x_i; \tilde{\theta}_n)(1 - F(x_i; \tilde{\theta}_n))}$ satisfies Assumption 2. Therefore, the constructed weighted M -estimator $\tilde{\theta}_n^*$ is consistent.

Supplementary Materials

Technical details: The PDF file provides additional simulation results based on the Type II censored data: the proposed methods are compared with the EPM when the censoring rate is set as 10%, 30%, and 40% (PDF file).

Source code: The R file contains the R code to implement the proposed M -estimation and weighted M -estimation (R file).

Acknowledgments

The authors thank the editor, associate editor, and two anonymous referees for their constructive comments and suggestions that have considerably improved the article. This work is supported by Natural Science Foundation of China (71601138) and Singapore AcRF Tier 1 funding (#R-266-000-081-133 and #R-266-000-095-112).

References

- Andrews, D., and Herzberg, A. (1985), "Stress-Rupture Life of Kevlar 49/Epoxy Spherical Pressure Vessels," in *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, eds. D. F. Andrews and A. M. Herzberg, New York: Springer, pp. 181–186. [537]
- Castellanos, E. M., and Cabras, S. (2007), "A Default Bayesian Procedure for the Generalized Pareto Distribution," *Journal of Statistical Planning and Inference*, 137, 473–483. [529,533]
- Castillo, E., and Hadi, A. S. (1997), "Fitting the Generalized Pareto Distribution to Data," *Journal of the American Statistical Association*, 92, 1609–1620. [528,529,531,532,536,537]
- Castillo, E., Hadi, A. S., Balakrishnan, N., and Sarabia, J.-M. (2005), *Extreme Value and Related Models with Applications in Engineering and Science*, Hoboken, NJ: Wiley. [528,529,534]
- Chaudhury, K. N. (2013), "On the Convergence of the IRLS Algorithm in Non-Local Patch Regression," *IEEE Signal Processing Letters*, 20, 815–818. [531]
- Chen, P., and Ye, Z.-S. (2017), "Estimation of Field Reliability Based on Aggregate Lifetime Data," *Technometrics*, 59, 115–125. [531]
- Choulakian, V., and Stephens, M. A. (2001), "Goodness-of-Fit Tests for the Generalized Pareto Distribution," *Technometrics*, 43, 478–484. [528,537]
- Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001), *An Introduction to Statistical Modeling of Extreme Values*, New York: Springer. [534]
- Davis, A., and Smith, R. (1990), "Models for Exceedances Over High Thresholds" (with discussion), *Journal of the Royal Statistical Society, Series B*, 52, 393–442. [528]
- Davis, H. T., and Feldstein, M. L. (1979), "The Generalized Pareto Law as a Model for Progressively Censored Survival Data," *Biometrika*, 66, 299–306. [528]
- De Haan, L., and Ferreira, A. (2007), *Extreme Value Theory: An Introduction*, New York: Springer. [534]
- Del Castillo, J., and Serra, I. (2015), "Likelihood Inference for Generalized Pareto Distribution," *Computational Statistics and Data Analysis*, 83, 116–128. [536]
- De Zea Bermudez, P., and Kotz, S. (2010), "Parameter Estimation of the Generalized Pareto Distribution—Part II," *Journal of Statistical Planning and Inference*, 140, 1374–1388. [539]
- Efron, B., and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press. [531]
- Einmahl, J. (1989), "On the Standardized Empirical Process," *Statistica Neerlandica*, 43, 175–179. [540]
- Ferger, D. (2010), "Minimum Distance Estimation in Normed Linear Spaces with Donsker-Classes," *Mathematical Methods of Statistics*, 19, 246–266. [530]
- Finkenstadt, B., and Rootzén, H. (2003), *Extreme Values in Finance, Telecommunications, and the Environment*, Boca Raton, FL: CRC Press. [528,534]
- Green, P. J. (1984), "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives," *Journal of the Royal Statistical Society, Series B*, 46, 149–192. [531]
- Grimshaw, S. D. (1993), "Computing Maximum Likelihood Estimates for the Generalized Pareto Distribution," *Technometrics*, 35, 185–191. [529]
- Gumbel, E. (1958), *Statistics of Extremes*, New York: Columbia University Press. [528]
- Hart, J. D., and Cañette, I. (2011), "Nonparametric Estimation of Distributions in Random Effects Models," *Journal of Computational and Graphical Statistics*, 20, 461–478. [530]
- Hosking, J., and Wallis, J. (1987), "Parameter and Quantile Estimation for the Generalized Pareto Distribution," *Technometrics*, 29, 339–349. [529]
- Huber, P. J. (1972), "The 1972 Wald Lecture Robust Statistics: A Review," *The Annals of Mathematical Statistics*, 43, 1041–1067. [530]
- (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821. [530]
- (2011), *Robust Statistics*, New York: Springer. [530]
- Jaeckel, L. A. (1972), "Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals," *The Annals of Mathematical Statistics*, 43, 1449–1458. [530]
- Krehbiel, T., and Adkins, L. C. (2008), "Extreme Daily Changes in US Dollar London Inter-Bank Offer Rates," *International Review of Economics & Finance*, 17, 397–411. [528]
- Luceño, A. (2006), "Fitting the Generalized Pareto Distribution to Data using Maximum Goodness-of-Fit Estimators," *Computational Statistics and Data Analysis*, 51, 904–917. [536]
- Meeker, W. Q., and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, Hoboken, NJ: Wiley. [535]
- Moisello, U. (2007), "On the Use of Partial Probability Weighted Moments in the Analysis of Hydrological Extremes," *Hydrological Processes*, 21, 1265–1279. [528]
- Parr, W. C., and Schucany, W. R. (1980), "Minimum Distance and Robust Estimation," *Journal of the American Statistical Association*, 75, 616–624. [530]
- Pickands III, J. (1975), "Statistical Inference Using Extreme Order Statistics," *The Annals of Statistics*, 3, 119–131. [528]
- Politis, D. N. (2013), "Model-Free Model-Fitting and Predictive Distributions," *Test*, 22, 183–221. [530]
- Rey, W. (2012), *Introduction to Robust and Quasi-Robust Statistical Methods*, New York: Springer. [530,531]
- Rootzén, H., and Zholud, D. (2016), "Tail Estimation for Window-Censored Processes," *Technometrics*, 58, 95–103. [528]
- Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880. [530]
- Smith, R. L. (1984), "Threshold Methods for Sample Extremes," in *Statistical Extremes and Applications*, ed. J. Tiago de Oliveira, New York: Springer, pp. 621–638. [528,529]
- Song, J., and Song, S. (2012), "A Quantile Estimation for Massive Data with Generalized Pareto Distribution," *Computational Statistics and Data Analysis*, 56, 143–150. [530]
- Van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press. [530]
- Van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence*, New York: Springer. [539]

- Wolfowitz, J. (1957), "The Minimum Distance Method," *The Annals of Mathematical Statistics*, 28, 75–88. [530]
- Yohai, V. J., and Zamar, R. H. (1988), "High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale," *Journal of the American Statistical Association*, 83, 406–413. [530]
- Zagorski, M., and Wnek, M. (2007), "Analysis of the Turbine Steady-State Data by Means of Generalized Pareto Distribution," *Mechanical Systems and Signal Processing*, 21, 2546–2559. [528]
- Zhang, J. (2002), "Powerful Goodness-of-Fit Tests Based on the Likelihood Ratio," *Journal of the Royal Statistical Society, Series B*, 64, 281–294. [530]
- (2010), "Improving on Estimation for the Generalized Pareto Distribution," *Technometrics*, 52, 335–339. [529]
- Zhang, J., and Stephens, M. A. (2009), "A New and Efficient Estimation Method for the Generalized Pareto Distribution," *Technometrics*, 51, 316–325. [528,529,532,533,536,539]