# PARAMETER ESTIMATION BASED ON CUMULATIVE KULLBACK-LEIBLER DIVERGENCE *

Authors: Yaser Mehrali
– Department of Statistics, University of Isfahan, 81744
Isfahan, Iran (y.mehrali@sci.ui.ac.ir)

Majid Asadi
– Department of Statistics, University of Isfahan, 81744
Isfahan, Iran (m.asadi@sci.ui.ac.ir)

Abstract:

- In this paper, we propose some estimators for the parameters of a statistical model based on Kullback-Leibler divergence of the survival function in continuous setting and apply it to type $I$ censored data. We prove that the proposed estimators are subclass of "generalized estimating equations" estimators. The asymptotic properties of the estimators such as consistency and asymptotic normality are investigated. Some illustrative examples are also provided. In particular, in estimating the shape parameter of generalized Pareto distribution, we show that our procedure dominates some existing methods in the sense of bias and mean squared error.

Key-Words:

- *Estimation; Generalized Estimating Equations; Information Measures; Generalized Pareto distribution; Censoring.*

AMS Subject Classification:

- 62B10, 94A15, 94A17, 62G30, 62E20, 62F10, 62F12, 62N02.

---

*The opinions expressed in this text are those of the authors and do not necessarily reflect the views of any organization.

## 1. INTRODUCTION

The Kullback-Leibler ($KL$) divergence (also known as relative entropy) is a measure of discrimination between two probability distributions. If the random variables $X$ and $Y$ have probability density functions $f$ and $g$, respectively, the $KL$ divergence of $f$ relative to $g$ is defined as

$$D(f||g) = \int_{\mathbb{R}} f(x) \log \frac{f(x)}{g(x)} dx,$$

for $x$ such that $g(x) \neq 0$. The function $D(f||g)$ is always nonnegative and it is zero if and only if $f = g$ $a.s.$

Let $f_{\boldsymbol{\theta}}$ belong to a parametric family with $p$-dimensional parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ and $f_n$ be a kernel density estimator of $f_{\boldsymbol{\theta}}$ based on $n$ random variables $\{X_1, \ldots, X_n\}$ of distribution of $X$. Basu and Lindsay [3] used $KL$ divergence of $f_n$ relative to $f_{\boldsymbol{\theta}}$ as

$$D\left(f_n||f_{\boldsymbol{\theta}}\right) = \int_{\mathbb{R}} f_n(x) \log \frac{f_n(x)}{f(x;\boldsymbol{\theta})} dx, \tag{1.1}$$

and defined the minimum $KL$ divergence estimator of $\boldsymbol{\theta}$ as

$$\widehat{\boldsymbol{\theta}} = \arg \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} D\left(f_n||f_{\boldsymbol{\theta}}\right).$$

Lindsay [19] proposed a version of (1.1) in discrete setting. In recent years, many authors such as Morales et al. [21], Jiménez and Shao [17], Broniatowski and Keziou [6], Broniatowski [5], Cherfi [7, 8, 9] studied the properties of minimum divergence estimators under different conditions. Basu et al. [4] discussed in their book about the statistical inference with the minimum distance approach.

Although the method of estimation based on $D\left(f_n||f_{\boldsymbol{\theta}}\right)$ has very interesting properties, the definition is based on $f$ which, in general, may not exist.

Let $X$ be a random variable with cumulative distribution function ($c.d.f$) $F(x) = P(X \leq x)$ and survival function ($s.f$) $\bar{F}(x) = 1 - F(x)$. Based on $n$ observations $\{x_1, \ldots, x_n\}$ of distribution $F$, define the empirical cumulative distribution and survival functions, respectively, by

$$F_n(x) = \sum_{i=1}^{n} \frac{i}{n} I_{\left[x_{(i)}, x_{(i+1)}\right)}(x), \tag{1.2}$$

and

$$\bar{F}_n(x) = \sum_{i=0}^{n-1} \left(1 - \frac{i}{n}\right) I_{\left[x_{(i)}, x_{(i+1)}\right)}(x), \tag{1.3}$$

where $I$ is the indicator function and $(-\infty = x_{(0)} \leq)\ x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ $(\leq x_{(n+1)} = \infty)$ are the order observations corresponding to the sample. The function $F_n$ $(\bar{F}_n)$ is known in the literature as "empirical estimator" of $F$ $(\bar{F})$.

In the case when $X$ and $Y$ are continuous nonnegative random variables with $s.f$'s $\bar{F}$ and $\bar{G}$, respectively, a version of $KL$ divergence in terms of $s.f$'s $\bar{F}$ and $\bar{G}$ can be given as follows:

$$KLS\left(\bar{F}||\bar{G}\right) = \int_0^\infty \bar{F}(x) \log \frac{\bar{F}(x)}{\bar{G}(x)} dx - \left[E(X) - E(Y)\right].$$

The properties of this divergence measure are studied by some authors such as Liu [20] and Baratpour and Habibi Rad [1].

In order to estimate the parameters of a statistical model $F_{\boldsymbol{\theta}}$, Liu [20] proposed cumulative $KL$ divergence between the empirical survival function $\bar{F}_n$ and survival function $\bar{F}_{\boldsymbol{\theta}}$ (we call it $CKL\left(\bar{F}_n||\bar{F}_{\boldsymbol{\theta}}\right)$) as

$$\begin{aligned} CKL\left(\bar{F}_n||\bar{F}_{\boldsymbol{\theta}}\right) &= \int_0^\infty \left(\bar{F}_n(x) \log \frac{\bar{F}_n(x)}{\bar{F}(x;\boldsymbol{\theta})} - \left[\bar{F}_n(x) - \bar{F}(x;\boldsymbol{\theta})\right]\right) dx \\ &= \int_0^\infty \bar{F}_n(x) \log \bar{F}_n(x)\, dx - \int_0^\infty \bar{F}_n(x) \log \bar{F}(x;\boldsymbol{\theta})\, dx \\ &\quad - \left[\bar{x} - E_{\boldsymbol{\theta}}(X)\right], \end{aligned}$$

where $\overline{x}$ is the observed sample mean. The cited author defined minimum $CKL$ divergence estimator $(MCKLE)$ of $\boldsymbol{\theta}$ as

$$\widehat{\boldsymbol{\theta}} = \arg \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} CKL\left(\bar{F}_n(x)\,||\bar{F}_{\boldsymbol{\theta}}\right).$$

If we consider the parts of $CKL\left(\bar{F}_n||\bar{F}\right)$ that depends on $\boldsymbol{\theta}$ and define

$$g(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(X) - \int_0^\infty \bar{F}_n(x) \log \bar{F}(x;\boldsymbol{\theta})\, dx, \tag{1.4}$$

then the $MCKLE$ of $\boldsymbol{\theta}$ can equivalently be defined by

$$\widehat{\boldsymbol{\theta}} = \arg \inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} g(\boldsymbol{\theta}).$$

Two important advantages of this estimator are that one does not need to have the density function and that for large values of $n$ the empirical estimator $F_n$ tends to the distribution function $F$. Liu [20] applied this estimator in uniform and exponential models and Yari and Saghafi [35] and Yari et al. [34] used it for estimating parameters of Weibull distribution; see also Park et al. [26] and Hwang and Park [16]. Yari et al. [34] found a simple form of (1.4) as

$$g(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(X) - \frac{1}{n} \sum_{i=1}^n h(x_i) = E_{\boldsymbol{\theta}}(X) - \overline{h(x)}, \tag{1.5}$$

where $\overline{h(x)} = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$, and

$$h(x) = \int_0^x \log \bar{F}(y;\boldsymbol{\theta})\, dy. \tag{1.6}$$

They also proved that

$$E(h(X)) = \int_0^\infty \bar{F}(x;\boldsymbol{\theta}) \log \bar{F}(x;\boldsymbol{\theta})\, dx,$$

which shows that if $n$ tends to infinity, then $CKL\left(\bar{F}_n || \bar{F}_{\boldsymbol{\theta}}\right)$ converges to zero.

The aim of the present paper is to extend the definition of $MCKLE$ to the case that the random variable of interest has support in whole real line. In the process of doing so we also investigate asymptotic properties of $MCKLE$ and provide some examples.

Recently Park et al. [24] extended the cumulative Kullback-Leibler information to the whole real line as

$$CRKL(F:G) = \int_{-\infty}^{\infty} \bar{F}(x) \log \frac{\bar{F}(x)}{\bar{G}(x)} dx - [E(X) - E(Y)],$$

and

$$CKL(F:G) = \int_{-\infty}^{\infty} F(x) \log \frac{F(x)}{G(x)} dx - [E(Y) - E(X)].$$

They proposed a general cumulative Kullback-Leibler information as

$$GCKL_\alpha(F:G) = \alpha CKL(F:G) + (1-\alpha) CRKL(F:G), \quad 0 \le \alpha \le 1,$$

and studied its application to a test for normality in comparison with some competing test statistics based on the empirical distribution function.

The rest of the paper is organized as follows: In Section 2, we propose an extension of the $MCKLE$ in the case when the support of the distribution is real line and present some illustrative examples. In Section 3, we show that the proposed estimator belongs to the class of generalized estimating equations ($GEE$). Asymptotic properties of $MCKLE$ such as consistency, normality are investigated in this section. Several examples are given in this section. We have shown, among other examples, that when the underlying distribution is generalized Pareto one can employ $MCKLE$ to estimate the shape parameter of the model, for a subset of parameter space, while the $MLE$ does not exist in that subset. In Section 4, we extend the results to the type $I$ censored data.

## 2. AN EXTENSION OF $MCKLE$

In this section, we propose an extension of the $MCKLE$ for the case when $X$ is assumed to be a continuous random variable with support $\mathbb{R}$. It is known that [30]

$$E_{\boldsymbol{\theta}}|X| = \int_{-\infty}^{0} F(x)\,dx + \int_{0}^{\infty} \bar{F}(x)\,dx.$$

We first give an extension of $CKL$ divergence for the case that the random variables are distributed over real line $\mathbb{R}$.

**Definition 2.1.** Let $X$ and $Y$ be random variables on $\mathbb{R}$ with *c.d.f*'s $F$ and $G$, *s.f*'s $\bar{F}$ and $\bar{G}$ and finite means $E(X)$ and $E(Y)$, respectively. The $CKL$ divergence of $\bar{F}$ relative to $\bar{G}$ is defined as

$$\begin{aligned} CKL\left(\bar{F}||\bar{G}\right) &= \int_{-\infty}^{0} \left\{F(x)\log\frac{F(x)}{G(x)} - [F(x) - G(x)]\right\} dx \\ &\quad + \int_{0}^{\infty} \left\{\bar{F}(x)\log\frac{\bar{F}(x)}{\bar{G}(x)} - \left[\bar{F}(x) - \bar{G}(x)\right]\right\} dx \\ &= \int_{-\infty}^{0} F(x)\log\frac{F(x)}{G(x)}dx + \int_{0}^{\infty} \bar{F}(x)\log\frac{\bar{F}(x)}{\bar{G}(x)}dx - [E|X| - E|Y|]. \end{aligned}$$

An application of the log-sum inequality and the fact that, for all $x, y > 0$ $x\log\frac{x}{y} \geq x - y$, (equality holds if and only if $x = y$) show that the $CKL$ is non-negative. Using the fact that in log-sum inequality, equality holds if and only if $F = G$, *a.s.*, one gets that $CKL\left(\bar{F}||\bar{G}\right) = 0$ if and only if $F = G$, *a.s.*

Let $F_{\boldsymbol{\theta}}$ be the population *c.d.f.* with unknown parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ and $F_n$ be the empirical *c.d.f.* based on a random sample $X_1, X_2, \ldots, X_n$ from $F_{\boldsymbol{\theta}}$. Based on the above definition, the $CKL$ divergence of $\bar{F}_n$ relative to $\bar{F}_{\boldsymbol{\theta}}$ is defined as

$$\begin{aligned} CKL\left(\bar{F}_n||\bar{F}_{\boldsymbol{\theta}}\right) &= \int_{-\infty}^{0} F_n(x)\log\frac{F_n(x)}{F(x;\boldsymbol{\theta})}dx + \int_{0}^{\infty} \bar{F}_n(x)\log\frac{\bar{F}_n(x)}{\bar{F}(x;\boldsymbol{\theta})}dx \\ &\quad - \left[\overline{|x|} - E_{\boldsymbol{\theta}}|X|\right], \end{aligned}$$

where $\overline{|x|}$ is the mean of absolute values of the observations. Let us also define

$$g(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}|X| - \int_{-\infty}^{0} F_n(x)\log F(x;\boldsymbol{\theta})\,dx - \int_{0}^{\infty} \bar{F}_n(x)\log\bar{F}(x;\boldsymbol{\theta})\,dx. \tag{2.1}$$

Now, we have the following definition which is an extension of $CKL$ estimator in Liu approach:

**Definition 2.2.** Assume that $E_{\boldsymbol{\theta}}|X| < \infty$ and $g''(\boldsymbol{\theta})$ is positive definite. Then, under the existence, we define $MCKLE$ of $\boldsymbol{\theta}$ to be a value in the parameter space $\boldsymbol{\Theta}$ which minimizes $g(\boldsymbol{\theta})$.

If $X$ is nonnegative, then $g(\boldsymbol{\theta})$ in (2.1) reduces to (1.4). So the results of Liu [20], Yari and Saghafi [35], Yari et al. [34], Park et al. [26] and Hwang and Park

[16] yield as special cases. It should be noted that by the law of large numbers $F_n$ converges to $F_{\boldsymbol{\theta}}$ and $\bar{F}_n$ converges to $\bar{F}_{\boldsymbol{\theta}}$ as $n$ tends to infinity. Consequently $CKL\left(\bar{F}_n||\bar{F}_{\boldsymbol{\theta}}\right)$ converges to zero as $n$ tends to infinity.

In order to study the properties of the estimator, we first find a simple form of (2.1). Let us introduce the following notations:

$$u(x) = \int_x^0 \log F(y;\boldsymbol{\theta})\, dy,$$

for $x < 0$, and

$$s(x) = I_{(-\infty,0)}(x)\, u(x) + I_{[0,\infty)}(x)\, h(x), \tag{2.2}$$

for $x \in \mathbb{R}$, where $h$ is defined in (1.6). Assuming that $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ denote the ordered observed values of the sample and that $x_{(k)} < 0 \leq x_{(k+1)}$, for some value of $k$, $k = 0, \ldots, n$ ($x_{(0)} = -\infty$), then by (1.2) and (1.3), we have

$$\begin{aligned}
\int_{-\infty}^0 F_n(x) \log F(x;\boldsymbol{\theta})\, dx &= \sum_{i=1}^{k-1} \frac{i}{n} \int_{x_{(i)}}^{x_{(i+1)}} \log F(x;\boldsymbol{\theta})\, dx + \frac{k}{n} \int_{x_{(k)}}^{0} \log F(x;\boldsymbol{\theta})\, dx \\
&= \frac{1}{n} \sum_{i=1}^{k-1} i\left[u\left(x_{(i)}\right) - u\left(x_{(i+1)}\right)\right] + \frac{k}{n} u\left(x_{(k)}\right) \\
&= \frac{1}{n} \sum_{i=1}^{k} u\left(x_{(i)}\right).
\end{aligned}$$

Using the same steps, we have

$$\int_0^\infty \bar{F}_n(x) \log \bar{F}(x;\boldsymbol{\theta})\, dx = \frac{1}{n} \sum_{i=k+1}^{n} h\left(x_{(i)}\right).$$

So, $g(\boldsymbol{\theta})$ in (2.1) gets the simple form

$$\begin{aligned}
g(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}|X| - \frac{1}{n} \sum_{i=1}^{k} u\left(x_{(i)}\right) - \frac{1}{n} \sum_{i=k+1}^{n} h\left(x_{(i)}\right) \\
&= E_{\boldsymbol{\theta}}|X| - \frac{1}{n} \sum_{i=1}^{n} s(x_i) = E_{\boldsymbol{\theta}}|X| - \overline{s(x)}. \qquad (2.3)
\end{aligned}$$

If $k = 0$ (i.e., $X$ is nonnegative), then $g(\boldsymbol{\theta})$ in (2.3) reduces to (1.5). It can be easily seen that

$$E(s(X)) = \int_{-\infty}^0 F(x;\boldsymbol{\theta}) \log F(x;\boldsymbol{\theta})\, dx + \int_0^\infty \bar{F}(x;\boldsymbol{\theta}) \log \bar{F}(x;\boldsymbol{\theta})\, dx,$$

In the following, we give some examples.

**Example 2.1.** Let $\{X_1, \ldots, X_n\}$ be *i.i.d.* Normal random variables with probability density function

$$\phi\left(x;\mu,\sigma\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right),\quad x\in\mathbb{R},\quad \mu\in\mathbb{R},\quad \sigma>0.$$

In this case $E\left(|X|\right)=\mu\left[2\Phi\left(\frac{\mu}{\sigma}\right)-1\right]+2\sigma\phi\left(\frac{\mu}{\sigma}\right)$, where $\Phi$ denotes the distribution function of standard normal. For this distribution, $h\left(x\right)$, $u\left(x\right)$ and $g\left(\mu,\sigma\right)$ do not have closed forms. The zeros of the gradient of $g\left(\mu,\sigma\right)$ with respect to $\mu$ and $\sigma$ give respectively

$$2n\Phi\left(\frac{\mu}{\sigma}\right)-n-\sum_{\substack{i=1\\x_i<0}}^{k}\log\Phi\left(\frac{x_i-\mu}{\sigma}\right)+k\log\Phi\left(-\frac{\mu}{\sigma}\right)$$
$$+\sum_{\substack{i=k+1\\x_i\geq 0}}^{n}\log\Phi\left(\frac{\mu-x_i}{\sigma}\right)-\left(n-k\right)\log\Phi\left(\frac{\mu}{\sigma}\right)=0,$$

and

$$2n\phi\left(\frac{\mu}{\sigma}\right)+\sum_{\substack{i=1\\x_i<0}}^{k}\int_{\frac{x_i-\mu}{\sigma}}^{-\frac{\mu}{\sigma}}\frac{z\phi\left(z\right)}{\Phi\left(z\right)}dz-\sum_{\substack{i=k+1\\x_i\geq 0}}^{n}\int_{-\frac{\mu}{\sigma}}^{\frac{x_i-\mu}{\sigma}}\frac{z\phi\left(z\right)}{1-\Phi\left(z\right)}dz=0. \tag{2.4}$$

To obtain our estimators, we need to solve these equations numerically. For computational purposes, the following equivalent equation can be solved instead of (2.4).

$$2\phi\left(\frac{\mu}{\sigma}\right)+\int_{\frac{x_{(1)}-\mu}{\sigma}}^{-\frac{\mu}{\sigma}}F_n\left(\mu+\sigma z\right)\frac{z\phi\left(z\right)}{\Phi\left(z\right)}dz-\int_{-\frac{\mu}{\sigma}}^{\frac{x_{(n)}-\mu}{\sigma}}\bar{F}_n\left(\mu+\sigma z\right)\frac{z\phi\left(z\right)}{1-\Phi\left(z\right)}dz=0.$$

Figure 1 compares these estimators with the corresponding $MLE$'s. In order to compare our estimators and the $MLE$'s we made a simulation study in which we used samples of sizes 10 to 55 by 5 with 10000 repeats, where we assume that the true values of the model parameters are $\mu_{true}=2$ and $\sigma_{true}=3$. It is evident from the plots that the $MCKLE$ approximately coincides with the $MLE$ in both cases.

**Example 2.2.** Let $\{X_1, \ldots, X_n\}$ be *i.i.d.* Laplace random variables with probability density function

$$f\left(x;\theta\right)=\frac{1}{2\theta}\exp\left(-\left|\frac{x}{\theta}\right|\right),\quad x\in\mathbb{R},\quad \theta>0.$$

We simply have $MCKLE$ of $\theta$ as

$$\widehat{\theta}=\sqrt{\frac{\overline{X^2}}{2}}.$$

This is exactly the moment estimator of $\theta$.

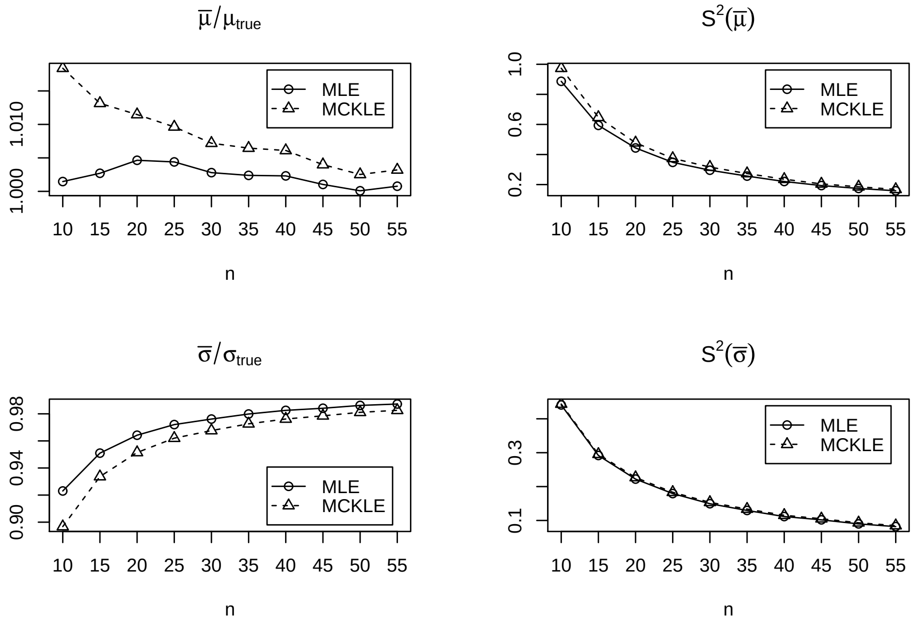


**Figure 1**: $\bar{\mu}/\mu_{true}$, $S^2\left(\bar{\mu}\right)$, $\bar{\sigma}/\sigma_{true}$ and $S^2\left(\bar{\sigma}\right)$ as functions of sample size

## 3. ASYMPTOTIC PROPERTIES OF ESTIMATORS

In this section we study asymptotic properties of $MCKLE$'s. For this purpose, first we give a brief review on $GEE$. Some related references on $GEE$ are Huber [13], Serfling [31], Qin and Lawless [29], van der Vaart [33], Pawitan [28], Shao [32], Huber and Ronchetti [15] and Hampel et al. [12].

Throughout this section, we use the terminology from Shao [32]. We assume that $X_1, ..., X_n$ represents independent random vectors, in which the dimension of $X_i$ is $d_i$, $i = 1, ..., n$ $(\sup_i d_i < \infty)$. We also assume that in the population model the vector $\boldsymbol{\theta}$ is a $p$-vector of unknown parameters. The $GEE$ method is a general method in statistical inference for deriving point estimators. Let $\boldsymbol{\Theta} \subset \mathbb{R}^p$ be the range of $\boldsymbol{\theta}$, $\boldsymbol{\psi}_i$ be a Borel function from $\mathbb{R}^{d_i} \times \boldsymbol{\Theta}$ to $\mathbb{R}^p$, $i = 1, ..., n$, and

$$s_n(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \boldsymbol{\psi}_i\left(X_i, \boldsymbol{\gamma}\right),\ \boldsymbol{\gamma} \in \boldsymbol{\Theta}.$$

If $\widehat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ is an estimator of $\boldsymbol{\theta}$ which satisfies $s_n(\widehat{\boldsymbol{\theta}}) = 0$, then $\widehat{\boldsymbol{\theta}}$ is called a $GEE$ estimator. The equation $s_n\left(\boldsymbol{\gamma}\right) = 0$ is called a $GEE$. Most of the estimation methods such as likelihood estimators, moment estimators and M-estimators are special cases of $GEE$ estimators. Usually $GEE$'s are chosen such that

$$E\left[s_n\left(\boldsymbol{\theta}\right)\right] = \sum_{i=1}^{n} E\left[\boldsymbol{\psi}_i\left(X_i, \boldsymbol{\theta}\right)\right] = 0. \tag{3.1}$$

If the exact expectation does not exist, then the expectation $E$ may be replaced by an asymptotic expectation. The consistency and asymptotic normality of the $GEE$ are studied under different conditions (see, for example Shao [32]).

## 3.1. Consistency and asymptotic normality of the $MCKLE$

Let $\widehat{\boldsymbol{\theta}}_n$ be $MCKLE$ which minimizes $g$ in (2.3) with $s$ as defined in (2.2). Here, we show that the $MCKLE$'s are special cases of $GEE$. Using this, we show the consistency and asymptotic normality of $MCKLE$'s.

**Theorem 3.1.** *$MCKLE$'s, by minimizing $g$ in (2.3), are special cases of $GEE$ estimators.*

**Proof:** In order to minimize $g$ in (2.3), we get the derivative of $g$, under the assumption that it exists,

$$\frac{\partial}{\partial \boldsymbol{\theta}} g\left(\boldsymbol{\theta}\right) = \frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}}\left|X\right| - \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} s\left(x_i\right) = 0,$$

which is equivalent to $GEE$ $s_n\left(\boldsymbol{\theta}\right) = 0$ where

$$s_n\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{n}\left[\frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}}\left|X\right| - \frac{\partial}{\partial \boldsymbol{\theta}} s\left(x_i\right)\right] = \sum_{i=1}^{n} \boldsymbol{\psi}\left(x_i, \boldsymbol{\theta}\right), \tag{3.2}$$

with

$$\boldsymbol{\psi}\left(x, \boldsymbol{\theta}\right) = \frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}}\left|X\right| - \frac{\partial}{\partial \boldsymbol{\theta}} s\left(x\right). \tag{3.3}$$

Now $E\left[s_n\left(\boldsymbol{\theta}\right)\right] = 0$, since

$$E\left[\frac{\partial}{\partial \boldsymbol{\theta}} s\left(X\right)\right] = \frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}}\left|X\right|, \tag{3.4}$$

that can be proven by some simple algebra. This proves the result. □

**Corollary 3.1.** *In the special case when the support of $X$ is $\mathbb{R}^+$, $MCKLE$ is an special case of $GEE$ estimators, where*

$$s_n(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[\frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(X) - \frac{\partial}{\partial\boldsymbol{\theta}}h(x_i)\right] = \sum_{i=1}^{n}\boldsymbol{\psi}(x_i,\boldsymbol{\theta}), \tag{3.5}$$

*with*

$$\boldsymbol{\psi}(x,\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(X) - \frac{\partial}{\partial\boldsymbol{\theta}}h(x). \tag{3.6}$$

The $MCKLE$'s are consistent estimators under mild conditions. To see this, let for each $n$ $\widehat{\boldsymbol{\theta}}_n$ be an $MCKLE$ or equivalently a $GEE$ estimator, i.e., $s_n\left(\widehat{\boldsymbol{\theta}}_n\right) = 0$, where $s_n$ is defined as (3.2) or (3.5). Suppose that $\boldsymbol{\psi}$ defined in (3.3) or (3.6) is a bounded and continuous function of $\boldsymbol{\theta}$. Let also

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = E\left[\boldsymbol{\psi}(X,\boldsymbol{\theta})\right],$$

where we assume that $\boldsymbol{\Psi}'(\boldsymbol{\theta})$ exists and is full rank. Then, from Proposition 5.2 of Shao [32] and using the fact that (3.1) holds, $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$.

Asymptotic normality of a consistent sequence of $MCKLE$'s can be established under some conditions. We first consider the special case where $\boldsymbol{\theta}$ is scalar and $X_1, ..., X_n$ are *i.i.d.*

**Theorem 3.2.** *Let $\widehat{\theta}_n$ be a consistent MCKLE of $\theta$. Then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N\left(0, \sigma_F^2\right),$$

*where $\sigma_F^2 = A/B^2$, with*

$$A = E\left[\frac{\partial}{\partial\boldsymbol{\theta}}s(X)\right]^2 - \left[\frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}|X|\right]^2,$$

*and*

$$B = \int_{-\infty}^{0}\frac{\left[\frac{\partial}{\partial\boldsymbol{\theta}}F(x;\boldsymbol{\theta})\right]^2}{F(x;\boldsymbol{\theta})}dx + \int_{0}^{\infty}\frac{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\bar{F}(x;\boldsymbol{\theta})\right]^2}{\bar{F}(x;\boldsymbol{\theta})}dx.$$

**Proof:** Using Theorem 3.1 we have $E\left[\boldsymbol{\psi}(X,\boldsymbol{\theta})\right] = 0$. So if we consider $\boldsymbol{\psi}$ defined in (3.3), we have

$$\begin{aligned} E\left[\boldsymbol{\psi}(X,\boldsymbol{\theta})\right]^2 &= Var\left[\boldsymbol{\psi}(X,\boldsymbol{\theta})\right] \\ &= Var\left[\frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}|X| - \frac{\partial}{\partial\boldsymbol{\theta}}s(X)\right] \\ &= Var\left[\frac{\partial}{\partial\boldsymbol{\theta}}s(X)\right] \\ &= E\left[\frac{\partial}{\partial\boldsymbol{\theta}}s(X)\right]^2 - \left[\frac{\partial}{\partial\boldsymbol{\theta}}E_{\boldsymbol{\theta}}|X|\right]^2, \end{aligned}$$

where the last equality follows from (3.4). On the other hand

$$\mathbf{\Psi}'(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta}^2} E_{\boldsymbol{\theta}} |X| - E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} s(X)\right],$$

and

$$\begin{aligned} E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} s(X)\right] &= \int_{-\infty}^{0}\int_{x}^{0} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log F(y;\boldsymbol{\theta})\, dy f(x;\boldsymbol{\theta})\, dx \\ &+ \int_{0}^{\infty}\int_{0}^{x} \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log \bar{F}(y;\boldsymbol{\theta})\, dy f(x;\boldsymbol{\theta})\, dx \\ &= \int_{-\infty}^{0} \left\{ \frac{\frac{\partial^2}{\partial \boldsymbol{\theta}^2} F(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})} - \left[\frac{\frac{\partial}{\partial \boldsymbol{\theta}} F(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})}\right]^2 \right\} F(y;\boldsymbol{\theta})\, dy \\ &+ \int_{0}^{\infty} \left\{ \frac{\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \bar{F}(y;\boldsymbol{\theta})}{\bar{F}(y;\boldsymbol{\theta})} - \left[\frac{\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(y;\boldsymbol{\theta})}{\bar{F}(y;\boldsymbol{\theta})}\right]^2 \right\} \bar{F}(y;\boldsymbol{\theta})\, dy \\ &= \frac{\partial^2}{\partial \boldsymbol{\theta}^2} E_{\boldsymbol{\theta}} |X| - \int_{-\infty}^{0} \frac{\left[\frac{\partial}{\partial \boldsymbol{\theta}} F(x;\boldsymbol{\theta})\right]^2}{F(x;\boldsymbol{\theta})} dx - \int_{0}^{\infty} \frac{\left[\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(x;\boldsymbol{\theta})\right]^2}{\bar{F}(x;\boldsymbol{\theta})} dx. \end{aligned}$$

So

$$\mathbf{\Psi}'(\boldsymbol{\theta}) = \int_{-\infty}^{0} \frac{\left[\frac{\partial}{\partial \boldsymbol{\theta}} F(x;\boldsymbol{\theta})\right]^2}{F(x;\boldsymbol{\theta})} dx + \int_{0}^{\infty} \frac{\left[\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(x;\boldsymbol{\theta})\right]^2}{\bar{F}(x;\boldsymbol{\theta})} dx.$$

Now, using Theorem 5.13 of Shao [32], $\sigma_F^2$ is given as

$$\sigma_F^2 = \frac{E(\psi^2(X,\boldsymbol{\theta}))}{[\Psi'(\boldsymbol{\theta})]^2}.$$

□

Similar to Theorem 3.2 it can be shown in the case that $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ is vector and $X_1, ..., X_n$ are *i.i.d.*, under the conditions of Theorem 5.14 of Shao [32],

$$V_n^{-1/2}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N_p(0, I_p),$$

where $V_n = \frac{1}{n} B^{-1} A B^{-1}$ with

$$A = \left[\frac{\partial}{\partial \boldsymbol{\theta}} s(X)\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}} s(X)\right]^T - \left[\frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}} |X|\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}} |X|\right]^T,$$

and

$$B = \int_{-\infty}^{0} \frac{\left[\frac{\partial}{\partial \boldsymbol{\theta}} F(x;\boldsymbol{\theta})\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}} F(x;\boldsymbol{\theta})\right]^T}{F(x;\boldsymbol{\theta})} dx + \int_{0}^{\infty} \frac{\left[\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(x;\boldsymbol{\theta})\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(x;\boldsymbol{\theta})\right]^T}{\bar{F}(x;\boldsymbol{\theta})} dx,$$

provided that $B$ is invertible matrix.

**Remark 3.1.** In Theorem 3.2 (and the result stated just after that for $p$ dimensional parameter) if we assume that the support of $X$ is nonnegative $A$ and $B$ are given, respectively, by

$$A = E\left[\frac{\partial}{\partial \boldsymbol{\theta}} h(X)\right]^2 - \left[\frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}}(X)\right]^2, \tag{3.7}$$

$$B = \int_0^\infty \frac{\left[\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(x;\boldsymbol{\theta})\right]^2}{\bar{F}(x;\boldsymbol{\theta})} dx,$$

and

$$A = E\left[\frac{\partial}{\partial \boldsymbol{\theta}} h(X)\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}} h(X)\right]^T - \left[\frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}}(X)\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}}(X)\right]^T, \tag{3.8}$$

$$B = \int_0^\infty \frac{\left[\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(x;\boldsymbol{\theta})\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(x;\boldsymbol{\theta})\right]^T}{\bar{F}(x;\boldsymbol{\theta})} dx.$$

Now, following Pawitan [28], we can find sample version of the variance formula for the $MCKLE$ as follows. Given $x_1, ..., x_n$ let

$$\begin{aligned} J &= \widehat{E}\left[\boldsymbol{\psi}(X, \boldsymbol{\theta})\right]^2 \\ &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}\left(x_i, \widehat{\boldsymbol{\theta}}\right)\boldsymbol{\psi}^T\left(x_i, \widehat{\boldsymbol{\theta}}\right) \\ &= \overline{\left\{\frac{\partial}{\partial \boldsymbol{\theta}} s(x)\right\}\left\{\frac{\partial}{\partial \boldsymbol{\theta}} s(x)\right\}^T}\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} - \left\{\frac{\partial}{\partial \boldsymbol{\theta}} \overline{s(x)}\right\}\left\{\frac{\partial}{\partial \boldsymbol{\theta}} \overline{s(x)}\right\}^T\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} I &= -\widehat{E}\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\psi}(X, \boldsymbol{\theta}) \\ &= -\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\psi}\left(x_i, \widehat{\boldsymbol{\theta}}\right) \\ &= -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} E_{\boldsymbol{\theta}}|X|\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} + \frac{\partial^2}{\partial \boldsymbol{\theta}^2}\overline{s(x)}\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \end{aligned} \tag{3.10}$$

Using notations defined in (3.9) and (3.10) we have

$$\widehat{V}_n^{-1/2}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N_p(0, I_p),$$

where

$$\widehat{V}_n = \frac{1}{n} I^{-1} J I^{-1}, \tag{3.11}$$

provided that $I$ is invertible matrix, or equivalently $g(\boldsymbol{\theta})$ has infimum value on parameter space $\Theta$. In particular when the support of $X$ is $\mathbb{R}^{+}$, $J$ and $I$ are given, respectively, by

$$J = \overline{\left\{\frac{\partial}{\partial \boldsymbol{\theta}} h(x)\right\}\left\{\frac{\partial}{\partial \boldsymbol{\theta}} h(x)\right\}^{T}}\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} - \left\{\overline{\frac{\partial}{\partial \boldsymbol{\theta}} h(x)}\right\}\left\{\overline{\frac{\partial}{\partial \boldsymbol{\theta}} h(x)}\right\}^{T}\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \tag{3.12}$$

and

$$I = -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} E_{\boldsymbol{\theta}}(X)\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} + \overline{\frac{\partial^2}{\partial \boldsymbol{\theta}^2} h(x)}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \tag{3.13}$$

In Theorem 3.2, the estimator $\widehat{V}_n$ is a sample version of $V_n$, see also Basu and Lindsay [3]. It is also known that the sample variance (3.11) is a robust estimator which is known as the 'sandwich' estimator, with $I^{-1}$ as the bread and $J$ as the filling [14]. In likelihood approach, the quantity $I$ is the usual observed Fisher information.

**Example 3.1.** Let $\{X_1, \ldots, X_n\}$ be *i.i.d.* exponential random variables with probability density function

$$f(x;\lambda) = \lambda e^{-\lambda x}, \qquad x > 0,\ \lambda > 0.$$

We simply have $MCKLE$ of $\lambda$ as

$$\widehat{\lambda} = \sqrt{\frac{2}{\overline{X^2}}}.$$

This estimator is a function of linear combinations of $X_i^2$'s, and so by strong law of large numbers ($SLLN$), $\widehat{\lambda}$ is strongly consistent for $\lambda$.

Now, using the central limit theorem ($CLT$) and delta method or using Theorem 3.2, one can show that

$$\sqrt{n}\left(\widehat{\lambda} - \lambda\right) \xrightarrow{d} N\left(0, \frac{5\lambda^2}{4}\right),$$

and the asymptotic bias of $\widehat{\lambda}$ is of order $\frac{1}{n}$: $E\left(\widehat{\lambda} - \lambda\right) = \frac{15\lambda}{8n}$. It is well known that the $MLE$ of $\lambda$ is $\widehat{\lambda}_m = 1/\bar{X}$ with asymptotic distribution

$$\sqrt{n}\left(\widehat{\lambda}_m - \lambda\right) \xrightarrow{d} N\left(0, \lambda^2\right),$$

and the asymptotic bias of $\widehat{\lambda}_m$ is of order $\frac{1}{n}$: $E\left(\widehat{\lambda}_m - \lambda\right) = \frac{\lambda}{n}$.

Notice that using asymptotic bias of $\widehat{\lambda}$, we can find some unbiasing factors to improve our estimator. Since the $MLE$ has inverse Gamma distribution,

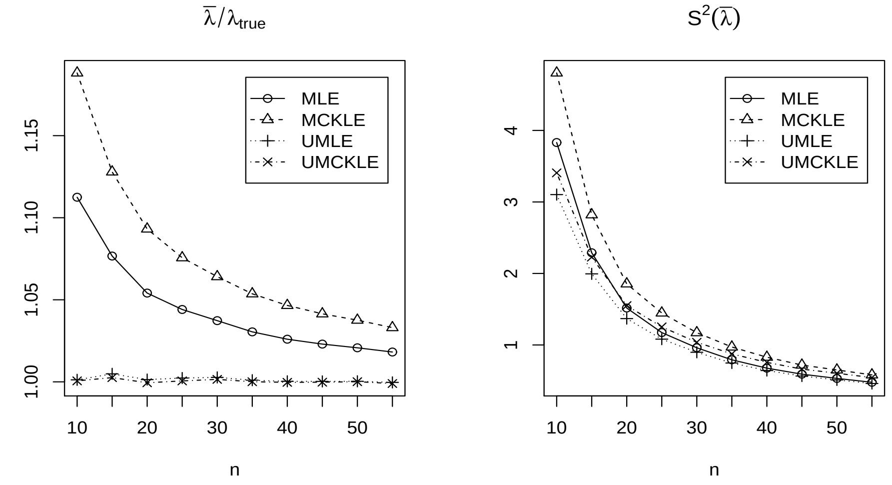


**Figure 2**: $\bar{\lambda}/\lambda_{true}$ and $S^2\left(\bar{\lambda}\right)$ as functions of sample size

the unbiased estimator of $\lambda$ is $\widehat{\lambda}_{um} = (n-1)/n\bar{X}$ [10]. In Liu approach an approximately unbiased estimator of $\lambda$ is

$$\widehat{\lambda}_u = \frac{8n}{8n+15}\sqrt{\frac{2}{\overline{X^2}}}. \tag{3.14}$$

Figure 2 compares these estimators. In order to compare our estimator and the $MLE$, we made a simulation study in which we used samples of sizes 10 to 55 by 5 with 10000 repeats, where we assumed that the true value of the model parameter is $\lambda_{true} = 5$. The plots in Figure 2 show that the $MCKLE$ has more bias than the $MLE$. It is evident from the plots that the $MCKLE$ in (3.14) which is approximately unbiased is very close to the unbiased $MLE$ in the sense of biased and variance.

**Remark 3.2.** In Example 2.2, note that $|X|$ has exponential distribution. So, using Example 3.1, one can easily find asymptotic properties of $\widehat{\theta}$ in Laplace distribution.

**Example 3.2.** Let $\{X_1, \ldots, X_n\}$ be *i.i.d.* two parameter exponential random variables with probability density function

$$f(x;\mu,\sigma) = \frac{1}{\sigma}e^{-(x-\mu)/\sigma}, \qquad x \geq \mu,\ \mu \in \mathbb{R},\ \sigma > 0.$$

If $\mu \geq 0$, then we have

$$g(\mu,\sigma) = \mu + \sigma + \frac{1}{2n\sigma}\sum_{i=1}^{n}(x_i-\mu)^2$$

and $MCKLE$ of $\mu$ and $\sigma$ are, respectively,

$$\widehat{\mu} = \overline{X} - \sqrt{\overline{X^2} - \overline{X}^2},\ \widehat{\sigma} = \sqrt{\overline{X^2} - \overline{X}^2},$$

which are also $ME$'s of $(\mu, \sigma)$. These estimators are functions of linear combinations of $X_i$'s and $X_i^2$'s, and hence by SLLN, $(\widehat{\mu}, \widehat{\sigma})$ are strongly consistent for $(\mu, \sigma)$.

Now, by $CLT$ and delta method or using Theorem 3.2, one can show that

$$V_n^{-1/2} \begin{pmatrix} \widehat{\mu} - \mu \\ \widehat{\sigma} - \sigma \end{pmatrix} \xrightarrow{d} N_2(0, I_2),$$

where

$$V_n = \frac{\sigma^2}{n} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}.$$

On the other hand if $\mu < 0$, then we get

$$g(\mu, \sigma) = 2\sigma \exp\left(\frac{\mu}{\sigma}\right) - \mu - \sigma + \frac{1}{n\sigma}\left[\sum_{\substack{i=k+1 \\ x_i \geq 0}}^{n} \frac{x_i^2}{2} - \mu \sum_{\substack{i=k+1 \\ x_i \geq 0}}^{n} x_i\right] + \frac{\sigma}{n}\left[\sum_{\substack{i=1 \\ x_i > 0}}^{k} Li_2\left(\exp\left(-\frac{x_i - \mu}{\sigma}\right)\right) - k \cdot Li_2\left(\exp\left(\frac{\mu}{\sigma}\right)\right)\right],$$

where $Li_2(\cdot)$ is the dilogarithm function. In this case, the $MCKLE$ of $\mu$ and $\sigma$ can be found numerically.

In the following example, we show that in generalized Pareto distribution while the $MLE$ of the shape parameter of the model does not exist one can use $MCKLE$ to estimate the shape parameter.

**Example 3.3.** Suppose that $\{X_1, \ldots, X_n\}$ are *iid* from generalized Pareto distribution $(GPD)$ with *c.d.f.*

$$F(x; \sigma, k) = \begin{cases} 1 - (1 - kx/\sigma)^{1/k}, & \text{if } k \neq 0, \\ 1 - e^{-x/\sigma}, & \text{if } k = 0, \end{cases}$$

where $\sigma > 0$, $k \in \mathbb{R}$, $0 \leq x < \infty$ for $k \leq 0$ and $0 \leq x \leq \sigma/k$ for $k > 0$. For this distribution the $MLE$ of the shape parameter $k$ does not exist for $k \in (1, \infty)$ [11]. Let $\sigma$ be fixed. After some algebra we get

$$g_n(k) = \frac{\sigma}{k+1} - \frac{1}{n}\sum_{i=1}^{n} h(x_i),\quad -1 < k \leq \sigma/x_{(n)},$$

where

$$h(x) = \begin{cases} -\frac{\sigma}{k^2}\left[\frac{kx}{\sigma} + \left(1-\frac{kx}{\sigma}\right)\log\left(1-\frac{kx}{\sigma}\right)\right], & k \neq 0, \frac{\sigma}{x}, \\ -\frac{x^2}{2\sigma}, & k = 0, \\ -\frac{x^2}{\sigma}, & k = \frac{\sigma}{x}, \end{cases}$$

and $MCKLE$ estimator $\widehat{k}$ can be found numerically. It should be noted that in this case, for $k \leq -1$, $\widehat{k}$ does not exist. Recently Zhang [37] considered the estimation of for $k$ based on the likelihood method and empirical Bayesian [36], [38]. Denoting the Zhang's estimator by $\widehat{k}_{Zhang}$, the cited author shows that the performance of $\widehat{k}_{Zhang}$ is better than other existing methods for $-6 \leq k \leq 1/2$. In order to compare our estimator ($\widehat{k}_{MCKLE}$) and Zhang's estimator $\widehat{k}_{Zhang}$, we evaluated them using simulated samples of sizes $15, 20, 50, 100, 200, 500$ and $1000$ with 10000 replicates, considering different true values of the population parameter as $k = -0.75, -0.5, -0.25, 0, 0.25, 0.5, 1, 3, 5$ and $7$. Tables 3.3 and 3.3 compare bias and root mean squared error ($RMSE$) of estimators, respectively. It is evident from Table 3.3 that for all values $k > 0.25$, $\widehat{k}_{MCKLE}$ has less bias than $\widehat{k}_{Zhang}$. Also for $k = 0.25$, $n = 15, 20, 500, 1000$, the performance of our estimator is better than the Zhang's estimator. On the other hand, it is seen from table 3.3 that except for $k = -0.75$, $n = 100, 200, 500, 1000$, and $k = -0.5$, $n = 500, 1000$, for all values of $k$, $\widehat{k}_{MCKLE}$ has less $RMSE$ than $\widehat{k}_{Zhang}$.

| $k$ | -0.75 | | -0.5 | | -0.25 | | 0 | | 0.25 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE |
| 15 | 0.0478 | 0.3084 | 0.0271 | 0.2136 | -0.0002 | 0.1472 | -0.0401 | 0.1041 | -0.1005 | 0.0761 |
| 20 | 0.0185 | 0.2714 | 0.0055 | 0.1801 | -0.0113 | 0.1189 | -0.0366 | 0.0810 | -0.0789 | 0.0573 |
| 50 | 0.0126 | 0.1840 | 0.0066 | 0.1039 | -0.0003 | 0.0581 | -0.0086 | 0.0346 | -0.0217 | 0.0219 |
| 100 | 0.0051 | 0.1420 | 0.0023 | 0.0698 | -0.0012 | 0.0337 | -0.0054 | 0.0180 | -0.0097 | 0.0103 |
| 200 | 0.0044 | 0.1135 | 0.0025 | 0.0490 | 0.0002 | 0.0209 | -0.0028 | 0.0103 | -0.0052 | 0.0056 |
| 500 | 0.0014 | 0.0845 | 0.0008 | 0.0293 | -0.0001 | 0.0100 | -0.0013 | 0.0043 | -0.0024 | 0.0021 |
| 1000 | 0.0010 | 0.0687 | 0.0007 | 0.0200 | 0.0002 | 0.0057 | -0.0006 | 0.0023 | -0.0012 | 0.0010 |
| $k$ | 0.5 | | 1 | | 3 | | 5 | | 7 | |
| $n$ | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE |
| 15 | -0.1852 | 0.0566 | -0.4162 | 0.0306 | -1.8133 | 0.0014 | -3.5561 | 0.0001 | -5.4191 | $2\times10^{-5}$ |
| 20 | -0.1452 | 0.0412 | -0.3430 | 0.0201 | -1.6568 | 0.0002 | -3.3632 | $-6\times10^{-6}$ | -5.2066 | $6\times10^{-6}$ |
| 50 | -0.0499 | 0.0136 | -0.1687 | 0.0033 | -1.2339 | -0.0004 | -2.8083 | $-1\times10^{-5}$ | -4.5742 | $3\times10^{-8}$ |
| 100 | -0.0208 | 0.0055 | -0.0979 | -0.0004 | -0.9988 | -0.0002 | -2.4627 | $-6\times10^{-7}$ | -4.1576 | $2\times10^{-10}$ |
| 200 | -0.0089 | 0.0025 | -0.0620 | -0.0012 | -0.8251 | -0.0001 | -2.1764 | $2\times10^{-9}$ | -3.7953 | $3\times10^{-12}$ |
| 500 | -0.0025 | 0.0005 | -0.0396 | -0.0012 | -0.6514 | $-8\times10^{-6}$ | -1.8621 | $2\times10^{-11}$ | -3.3789 | $4\times10^{-15}$ |
| 1000 | -0.0008 | 0.0001 | -0.0303 | -0.0010 | -0.5518 | $-2\times10^{-7}$ | -1.6659 | $5\times10^{-13}$ | -3.1068 | $< 10^{-16}$ |

**Table 1**: Biases of $\widehat{k}_{MCKLE}$ and $\widehat{k}_{Zhang}$ for the $GPD$

## 4. AN EXTENSION OF $MCKLE$ TO THE TYPE $I$ CENSORED DATA

In this section, we extend $MCKLE$ for the case when the data are collected in censored type $I$ scheme, in continuous case. Some authors such as Lim and Park [18], Cherfi [8], Baratpour and Habibi Rad [2], Park and Shin [27], Park et

| $k$ | -0.75 | | -0.5 | | -0.25 | | 0 | | 0.25 | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE |
| 15 | 0.4672 | 0.3968 | 0.4040 | 0.3267 | 0.3425 | 0.2730 | 0.2893 | 0.2264 | 0.2618 | 0.1852 |
| 20 | 0.4071 | 0.3496 | 0.3543 | 0.2826 | 0.3030 | 0.2324 | 0.2565 | 0.1893 | 0.2272 | 0.1516 |
| 50 | 0.2504 | 0.2382 | 0.2167 | 0.1808 | 0.1851 | 0.1409 | 0.1573 | 0.1074 | 0.1352 | 0.0803 |
| 100 | 0.1753 | 0.1863 | 0.1510 | 0.1354 | 0.1278 | 0.1014 | 0.1073 | 0.0736 | 0.0919 | 0.0527 |
| 200 | 0.1235 | 0.1501 | 0.1060 | 0.1043 | 0.0889 | 0.0743 | 0.0732 | 0.0514 | 0.0616 | 0.0356 |
| 500 | 0.0785 | 0.1154 | 0.0674 | 0.0758 | 0.0565 | 0.0498 | 0.0460 | 0.0322 | 0.0374 | 0.0216 |
| 1000 | 0.0550 | 0.0957 | 0.0472 | 0.0597 | 0.0395 | 0.0364 | 0.0319 | 0.0227 | 0.0255 | 0.0149 |
| $k$ | 0.5 | | 1 | | 3 | | 5 | | 7 | |
| $n$ | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE | Zhang | MCKLE |
| 15 | 0.2824 | 0.1498 | 0.4592 | 0.0948 | 1.8238 | 0.0131 | 3.5606 | 0.0021 | 5.4216 | 0.0004 |
| 20 | 0.2363 | 0.1198 | 0.3837 | 0.0715 | 1.6671 | 0.0077 | 3.3676 | 0.0010 | 5.2091 | 0.0001 |
| 50 | 0.1277 | 0.0587 | 0.2060 | 0.0287 | 1.2436 | 0.0016 | 2.8124 | 0.0001 | 4.5764 | $9 \times 10^{-7}$ |
| 100 | 0.0842 | 0.0367 | 0.1313 | 0.0158 | 1.0073 | 0.0008 | 2.4662 | $2 \times 10^{-5}$ | 4.1595 | $3 \times 10^{-9}$ |
| 200 | 0.0564 | 0.0239 | 0.0889 | 0.0093 | 0.8321 | 0.0003 | 2.1794 | $3 \times 10^{-8}$ | 3.7969 | $1 \times 10^{-10}$ |
| 500 | 0.0336 | 0.0139 | 0.0568 | 0.0049 | 0.6561 | 0.0001 | 1.8641 | $2 \times 10^{-10}$ | 3.3800 | $1 \times 10^{-13}$ |
| 1000 | 0.0228 | 0.0093 | 0.0422 | 0.0031 | 0.5550 | $8 \times 10^{-6}$ | 1.6673 | $2 \times 10^{-10}$ | 3.1075 | $6 \times 10^{-16}$ |

**Table 2**: $RMSE$'s of $\widehat{k}_{MCKLE}$ and $\widehat{k}_{Zhang}$ for the $GPD$

al. [22] Park and Lim [23] and Park and Pakyari [25] studied some forms of $KL$ divergences in different censored data cases. Let $T_1, ..., T_n$ be $i.i.d.$ nonnegative continuous random variables from a c.d.f. $F$, $p.d.f.$ $f$ and survival function $\bar{F}$. In a variety of applications in biostatistics and life testing, we are only able to observe $X = \min(T, C)$ where $C$ is the constant censoring point. The density function of $X$ can be written as

$$f_C(x) = \begin{cases} f(x), & 0 < x < C, \\ \bar{F}(C), & x = C, \\ 0, & o.w. \end{cases}$$

It is known that

$$E_{\boldsymbol{\theta}}(X) = \int_0^C \bar{F}(x)\,dx. \tag{4.1}$$

The authors in Lim and Park [18] and Park and Shin [27] presented two censored versions of $KL$ divergence of density $g_C$ relative to $f_C$ , respectively, by

$$I^*(g, f : C) = \int_{-\infty}^{C} g(x) \log \frac{g(x)}{f(x)} dx + F(C) - G(C),$$

and

$$K_{(-\infty,C)}(g : f) = \int_{-\infty}^{C} g(x) \log \frac{g(x)}{f(x)} dx + (1 - G(C)) \log \frac{1 - G(C)}{1 - F(C)},$$

which is nonnegative and is monotone in $C$. Park and Lim [23] defined $CKL$ for censored data as

$$CKL_C(\bar{G}||\bar{F}) = \int_0^C \bar{G}(x) \log \frac{\bar{G}(x)}{\bar{F}(x)} - \left[\bar{G}(x) - \bar{F}(x)\right] dx.$$

They also defined the $CKL_C$ of $F_n$ relative to $F$ as

$$\begin{aligned} CKL_C\left(\bar{F}_n||\bar{F}_{\boldsymbol{\theta}}\right) &= \int_0^C \bar{F}_n(x)\log\frac{\bar{F}_n(x)}{\bar{F}(x;\boldsymbol{\theta})} - \left[\bar{F}_n(x) - \bar{F}(x;\boldsymbol{\theta})\right]dx \\ &= \int_0^C \bar{F}_n(x)\log\bar{F}_n(x)\,dx - \int_0^C \bar{F}_n(x)\log\bar{F}(x;\boldsymbol{\theta})\,dx \\ &\quad + \int_0^C \bar{F}(x;\boldsymbol{\theta})\,dx - \int_0^C \bar{F}_n(x)\,dx, \end{aligned}$$

and considered it in type $II$ censorship. Here we apply $CKL_C$ for type $I$ censored data. Using (4.1) we get

$$CKL_C\left(\bar{F}_n||\bar{F}_{\boldsymbol{\theta}}\right) = \int_0^C \bar{F}_n(x)\log\bar{F}_n(x)\,dx - \int_0^C \bar{F}_n(x)\log\bar{F}(x;\boldsymbol{\theta})\,dx + E_{\boldsymbol{\theta}}(X) - \bar{x}.$$

Consider the parts of $CKL_C\left(\bar{F}_n||\bar{F}_{\boldsymbol{\theta}}\right)$ that depends on $\boldsymbol{\theta}$ and define

$$g(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(X) - \int_0^C \bar{F}_n(x)\log\bar{F}(x;\boldsymbol{\theta})\,dx. \tag{4.2}$$

Then the $MCKLE$ of $\boldsymbol{\theta}$ is defined as

$$\widehat{\boldsymbol{\theta}} = \arg\inf_{\boldsymbol{\theta}\in\Theta} CKL_C\left(\bar{F}_n||\bar{F}_{\boldsymbol{\theta}}\right) = \arg\inf_{\boldsymbol{\theta}\in\Theta} g(\boldsymbol{\theta}),$$

provided that $E_{\boldsymbol{\theta}}(X) < \infty$ and $g''(\boldsymbol{\theta})$ is positive definite; see also Park and Lim [23].

If $C \to \infty$, then $g(\boldsymbol{\theta})$ in (4.2) reduces to (1.4) and results in non-censored case yield as special case.

In order to study the properties of the estimator, following non-censored case, we have simple form of $g(\boldsymbol{\theta})$ as (1.5), with $h$ as (1.6).

Let $\widehat{\boldsymbol{\theta}}_n$ be $MCKLE$ in censored case by minimizing $g$ in (4.2). Here, $MCKLE$ is also an special case of $GEE$ with $\boldsymbol{\psi}(x,\boldsymbol{\theta})$ as (3.6), and under the conditions given in non-censored case the $MCKLE$ in censored case is also consistent. Asymptotic normality of a consistent sequence of $MCKLE$ can be established under the conditions imposed in non-censored case. We first consider the special case where $\boldsymbol{\theta}$ is scalar and $X_1, ..., X_n$ are *i.i.d.* continuous random variables.

**Theorem 4.1.** *For each $n$, let $\widehat{\boldsymbol{\theta}}_n$ be an $MCKLE$ or equivalently a $GEE$ estimator. Then*

$$\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N\left(0, \sigma_F^2\right),$$

*where* $\sigma_F^2 = A/B^2$, *with* $A$ *as* (3.7) *and*

$$B = \int_0^C \frac{\left[\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(x;\boldsymbol{\theta})\right]^2}{\bar{F}(x;\boldsymbol{\theta})} dx.$$

**Proof:** The proof is similar to non-censored case. □

The next theorem shows asymptotic normality of $MCKLE$, when $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ is vector and $X_1, ..., X_n$ are *i.i.d.* and continuous.

**Theorem 4.2.** *Under conditions of Theorem 5.14 of Shao [32],*

$$V_n^{-1/2}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N_p(0, I_p),$$

*where* $V_n = B^{-1}AB^{-1}$, *with* $A$ *as* (3.8) *and*

$$B = \int_0^C \frac{\left[\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(x;\boldsymbol{\theta})\right]\left[\frac{\partial}{\partial \boldsymbol{\theta}} \bar{F}(x;\boldsymbol{\theta})\right]^T}{\bar{F}(x;\boldsymbol{\theta})} dx,$$

*provided that* $B$ *is invertible matrix.*

**Proof:** The proof is similar to non-censored case and hence it is omitted. □

**Remark 4.1.** In Theorems 4.1 and 4.2, if $C \to \infty$ (no censoring), then results in non-censored case yield as special cases.

Now, following Pawitan [28], similar to non-censored case the sample version of the variance formula for the $MCKLE$ in censored case is as (3.11), with $I$ and $J$ as (3.12) and (3.13).

**Example 4.1.** Let $\{X_1, \ldots, X_n\}$ be *i.i.d.* type $I$ censored Exponential random variables with probability density function

$$f_C(x) = \begin{cases} \lambda e^{-\lambda x}, & 0 < x < C, \\ e^{-\lambda C}, & x = C, \\ 0, & o.w. \end{cases}$$

where $\lambda > 0$. After some algebra, we have

$$\begin{aligned} g(\lambda) &= \frac{1}{\lambda}\left(1 - e^{-\lambda C}\right) + \frac{\lambda(n-r)}{2n} C^2 + \frac{\lambda}{2n}\sum_{i=1}^{r} x_{(i)}^2 \\ &= \frac{1}{\lambda}\left(1 - e^{-\lambda C}\right) + \frac{\lambda}{2}\overline{x^2}, \end{aligned}$$

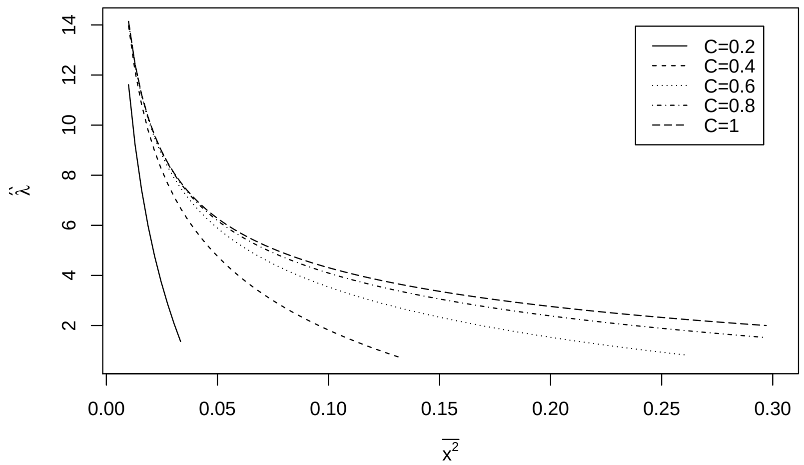


**Figure 3**: $\widehat{\lambda}$ as a decreasing function of $\overline{x^2}$

and $\widehat{\lambda}$ can be found numerically as a decreasing function of $\overline{x^2}$, and hence, by using strong law of large numbers ($SLLN$), it is strongly consistent. Figure 3 shows $\widehat{\lambda}$ as a decreasing function of $\overline{x^2}$.

Now, using Theorem 4.1, one can show that

$$\sqrt{n}\left(\widehat{\lambda}-\lambda\right) \xrightarrow{d} N\left(0,\sigma_F^2\right),$$

where

$$\sigma_F^2 = \frac{\lambda^2\left(5-e^{-2\lambda C}\left(\lambda C+1\right)^2-e^{-\lambda C}\left(\lambda^3C^3+3\lambda^2C^2+4\lambda C+4\right)\right)}{\left(2-e^{-\lambda C}\left(\lambda^2C^2+2\lambda C+2\right)\right)^2}.$$

If $C\to\infty$ (no censoring), then we obtain the results in non-censored case.

## ACKNOWLEDGMENTS

The authors would like to thank the Editor and an anonymous reviewer for useful comments which led to improving the exposition of this paper.

## REFERENCES


[1] Baratpour, S. and Habibi Rad, A. (2012). Testing goodness-of-fit for exponential distribution based on cumulative residual entropy. *Communications in Statistics-Theory & Methods*, **41**, 8, 1387-1396.

[2] Baratpour, S. and Habibi Rad, A. (2016). Exponentiality test based on the progressive type II censoring via cumulative entropy. *Communications in Statistics-Simulation & Computation*, **45**, 7, 2625-2637.

[3] Basu, A. and Lindsay, B. G. (1994). Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, **46**, 4, 683-705.

[4] Basu, A.; Shioya, H. and Park, C. (2011). *Statistical Inference: the Minimum Distance Approach.* CRC Press.

[5] Broniatowski, M. (2014). Minimum divergence estimators, maximum likelihood and exponential families. *Statistics & Probability Letters*, **93**, 27-33.

[6] Broniatowski, M. and Keziou, A. (2009). Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis,* **100**, 1, 16-36.

[7] Cherfi, M. (2011). Dual $\phi$-divergences estimation in normal models. *arXiv preprint arXiv:*1108.2999.

[8] Cherfi, M. (2012). Dual divergences estimation for censored survival data. *Journal of Statistical Planning and Inference,* **142**, 7, 1746-1756.

[9] Cherfi, M. (2014). On bayesian estimation via divergences. *Comptes Rendus Mathematique*, **352**, 9, 749-754.

[10] Forbes, C.; Evans, M., Hastings, N. and Peacock, B. (2011). *Statistical Distributions.* John Wiley & Sons.

[11] Grimshaw, S. D. (1993). Computing maximum likelihood estimates for the generalized Pareto distribution. *Technometrics*, **35**, 2, 185-191.

[12] Hampel, F. R.; Ronchetti, E. M.; Rousseeuw, P. J. & Stahel, W. A. (2011). *Robust Statistics: the Approach Based on Influence Functions.* John Wiley & Sons.

[13] Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, **35**, 1, 73-101.

[14] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics & Probability.* 221-233.

[15] Huber, P. and Ronchetti, E. (2009). *Robust Statistics*, Wiley: New York.

[16] Hwang, I. and Park, S. (2013). On scaled cumulative residual Kullback-Leibler information. *Journal of the Korean Data & Information Science Society*, **24**, 6, 1497-1501.

[17] Jiménez, R. and Shao, Y. (2001). On robustness and efficiency of minimum divergence estimators. *Test*, **10**, 2, 241-248.

[18] Lim, J. and Park, S. (2007). Censored Kullback-Leibler information and goodness-of-fit test with type II censored data. *Journal of Applied Statistics*, **34**, 9, 1051-1064.

[19] Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The Annals of Statistics*, **22**, 2, 1081-1114.

[20] Liu, J. (2007). *Information Theoretic Content and Probability.* Ph.D. Thesis, University of Florida.

[21] Morales, D.; Pardo, L. and Vajda, I. (1995). Asymptotic divergence of estimates of discrete distributions. *Journal of Statistical Planning & Inference*, **48**, 3, 347-369.

[22] Park, S.; Choi, D. and Jung, S. (2014). Kullback-Leibler information of the equilibrium distribution function and its application to goodness of fit test. *Communications for Statistical Applications & Methods*, **21**, 2, 125-134.

[23] Park, S. and Lim, J. (2015). On censored cumulative residual Kullback-Leibler information and goodness-of-fit test with type II censored data. *Statistical Papers*, **56**, 1, 247-256.

[24] Park, S.; Noughabi, H. A. and Kim, I. (2018). General cumulative Kullback-Leibler information. *Communications in Statistics-Theory & Methods*, **47**, 7, 1551-1560.

[25] Park, S. and Pakyari, R. (2015). Cumulative residual Kullback-Leibler information with the progressively Type-II censored data. *Statistics & Probability Letters*, **106**, 287-294.

[26] Park, S.; Rao, M. and Shin, D. W. (2012). On cumulative residual Kullback-Leibler information. *Statistics & Probability Letters,* **82**, 11, 2025-2032.

[27] Park, S. and Shin, M. (2014). Kullback-Leibler information of a censored variable and its applications. *Statistics*, **48**, 4, 756-765.

[28] Pawitan, Y. (2001). *In all Likelihood: Statistical Modelling and Inference Using Likelihood*: Oxford University Press.

[29] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, **22**, 1, 300-325.

[30] Rohatgi, V. K. and Ehsanes Saleh, A. K. Md. (2015). *An Introduction to Probability and Statistics (2 ed.).* John Wiley. New York.

[31] Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics* John Wiley. New York.

[32] Shao, J. (2003). *Mathematical Statistics (2 ed.).* Springer, New York, USA.

[33] van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge university press.

[34] Yari, G.; Mirhabibi, A. and Saghafi, A. (2013). Estimation of the Weibull parameters by Kullback-Leibler divergence of Survival functions. *Appl. Math. Inf. Sci.*, **7**, 1, 187-192.

[35] Yari, G. and Saghafi, A. (2012). Unbiased Weibull Modulus Estimation Using Differential Cumulative Entropy. *Communications in Statistics-Simulation & Computation*, **41**, 8, 1372-1378.

[36] Zhang, J. (2007). Likelihood moment estimation for the generalized Pareto distribution. *Australian & New Zealand Journal of Statistics*, **49**, 1, 69-77.

[37] Zhang, J. (2010). Improving on Estimation for the Generalized Pareto Distribution. *Technometrics*, **52**, 3, 335-339.

[38] Zhang, J. and Stephens, M. A. (2009). A new and efficient estimation method for the Generalized Pareto Distribution. *Technometrics*, **51**, 3, 316-325.