VOLUME/CİLT 14 NUMBER/SAYI 2 DECEMBER/ARALIK 2022

İSTATİSTİK

JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION

TÜRK İSTATİSTİK DERNEĞİ DERGİSİ

ISSN: 1300-4077

SUPPORTED AND PRINTED BY

IZMIR UNIVERSITY OF ECONOMICS

İSTATİSTİK

JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION

ISSN: 1300-4077
https://dergipark.org.tr/tr/pub/ijtsa

SUPPORTED AND PRINTED BY

İZMİR UNIVERSITY OF ECONOMICS

Izmir University of Economics Sakarya Cad. No:156 Balcova - Izmir/TURKEY

Phone: +90 (232) 488-8139 Fax: + 90 (232) 279-2626 www.ieu.edu.tr

AIM AND SCOPE

İSTATİSTİK, Journal of the Turkish Statistical Association is a refereed journal which publishes papers containing original contributions to probability, statistics and the interface of them with other disciplines where prosperity of scientific thought emerges through innovation, vitality and communication on interdisciplinary grounds. The Journal is published four-monthly. The media of Journal is English. The main areas of publication are probability and stochastic processes, theory of statistics, applied statistics, statistical computing and simulation, and interdisciplinary applications in social, demographic, physical, medical, biological, agricultural studies, engineering, computer science, management science, econometrics, etc.

In addition, the journal contains original research reports, authoritative review papers, discussed papers and occasional special issues or relevant conference proceedings.

ABSTRACTED/INDEXED: İSTATİSTİK, Journal of the Turkish Statistical Association is indexed in MathSciNet, Zentralblatt MATH and EBSCO.

SUBMISSION OF MANUSCRIPTS:

Authors should submit their papers electronically by using online submission system (https://dergipark.org.tr/tr/pub/ijtsa).

As part of the submission process, authors are required to check their submission's compliance with all of the following items. Submissions that do not adhere to these guidelines may be returned to authors.

1. The submission has not been previously published, or is being considered for publication in another journal.

2. The submission file is in Portable Document Format (PDF) prepared in LaTeX (TEX) document file format.

3. Where available, URLs for the references have been provided.

4. The text is single-spaced; uses a 12-point font; employs italics, rather than underlining (except with URL addresses); and all illustrations, figures, and tables are placed within the text at the appropriate points, rather than at the end.

5. References should be listed in alphabetical order and should be in the following formats:

Cai, T. and Low, M. (2005). Non-quadratic estimators of a quadratic functional. *The Annals of Statistics*, 33, 2930-2956.

Meyer, Y. (1992). *Wavelets and Operators.* Cambridge University Press, Cambridge.

Cox, D. (1969). Some sampling problems in technology. *In New Developments in Survey Sampling* (N.L. Johnson and H. Smith, Jr., eds.). Wiley, New York, 506-527.

İSTATİSTİK: JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION
Vol. 14, No. 2, December 2022, pp. 38–50
ISSN 1300-4077 | 22 | 2 | 38 | 50

İSTATİSTİK

A GRAPHICAL TOOL FOR EXTREME VALUE COPULA SELECTION BASED ON THE PICKANDS DEPENDENCE FUNCTION

Selim Orhun Susam*
Department of Econometric,
Munzur University,
Tunceli, Turkey

Abstract: We present a graphical tool that was primarily proposed by Michiels et al. [18] and later modified by Durante et al. [4]. We also improve this method to select the better fit of the given data among some extreme value copulas based on the Pickands dependence function. We conduct a Monte Carlo simulation study to investigate its performance. Also, the graphical method is illustrated by a real data example.

Key words: Copula, Pickands function, Extreme value copula.

1. Introduction

A copula is a joint distribution of the random variables U and V, each of which is marginally uniformly distributed as $U(0,1)$. Sklar's [20] theorem states that for any bivariate random variables X,Y with a cumulative distribution function (CDF)

$$H(x,y) = P(X \leq x, Y \leq y)$$

and the marginal CDF $F(x) = P(X \leq x)$ and $G(y) = P(Y \leq y)$ then there exist a copula such as:

$$H(x,y) = C(F(x), G(y)) = C(u,v),$$

where $u = F(x)$ and $v = g(y)$. From the modelling perspective, Sklar's Theorem allows us to separate the modelling of the marginal distributions $F(x), G(y)$ from the dependence structure, which is expressed in C.

One of the most important fields of statistics is the extreme value (EV) theory. The estimation of the events outside the range of data should be estimated by the EV distributions such as the daily maximum air temperature, and annual maximum sea levels. The EV distribution is the limiting distribution for the minimum or the maximum of random observations. Pickands [19] states that the pair (X,Y) has an EV dependence if and only if its copula C can be expressed for all $u,v \in (0,1)$

$$C(u,v) = \exp\Big(\log(uv)A(\frac{\log(v)}{\log(uv)})\Big),$$

where $A(.)$ is the Pickands dependence function defined on $[0,1] \to [1/2,1]$. The Pickand's dependence function has some properties as follows:

- $A(0) = A(1) = 1$.
- A is the convex function.
- $\max(1-t,t) \leq A(t) \leq 1$ for all $t \in [0,1]$.

* Corresponding author. E-mail address:orhunsusam@munzur.edu.tr

The non-parametric estimation of the Pickands dependence is an important issue when dealing with extreme events. Let $\{X_i, Y_i\}_{i=1}^n$ be a n random observation from the random variables X and Y with the joint distribution function $H(x,y)$, copula $C(u,v)$ and the marginal distribution functions $F(x)$ and $G(y)$. Also, let $U_i = F(x_i)$ and $V_i = G(y_i)$ then put $S_i = -\log(U_i)$ and $T_i = -\log(V_i)$. For every $t \in [0,1]$

$$\xi_i(t) = \min(\frac{S_i}{1-t}, \frac{T_i}{t}).$$

Pickands [19] introduced the non-parametric Pickands dependence function estimator as follows:

$$\hat{A}_P = \Big(\frac{1}{n}\sum_{i=1}^{n} \xi_i(t)\Big)^{-1}.$$

This estimator does not satisfy the conditions of the Pickands dependence function $A(.)$. Capéraá et al. [3] proposed an estimator called the CFG as follows:

$$\hat{A}_{CFG} = \exp\Big(-\gamma - \frac{1}{n}\sum_{i=1}^{n} \xi_i(t)\Big)^{-1},$$

where γ is Euler's constant that is $\gamma = -\int_0^{\inf} \log(x)\exp(x)dx$. In practice, marginals are rarely known. Thus, F and G should be estimated by their empirical counterparts $\hat{F}_n$ and $\hat{G}_n$ (Genest et al. [9]). In this paper, we use the corrected estimator $\hat{A}_{CFG}$ that is studied in Gudendorf et al. [10].

In the past few years, a certain number of papers have emerged which use Bernstein polynomials for the modelling of the extremal dependence, i.e. (Marcon et al. [16]; Guillotte et al. [11]; Marcon et al. [17]), to name a few. Also, Ahmadabadi et al. [1] investigated a new nonparametric approach using the Bernstein copula approximation. They used the Kernel regression method in order to derive an intrinsic estimator satisfying all the properties of the Pickands dependence function. See (Vettori et al. [21]) for a review.

The selection of EV copulas is an important issue when dealing with extreme situations. For this reason, many authors developed a tool for EV copulas selection. Michiels et al. [18] introduced a graphical tool for copula selection, based on the principal coordinate analysis. The main idea of this paper is that calculating the distance between the empirical copula and the parametric families of copulas then the calculated distances are visualized in 2D space via principal coordinate analysis. Also, Durante et al. [4] proposed the graphical tool in order to detect which families of copulas are closer to the empirical copula in tail dependence behavior.

In this study, we present a graphical tool that was firstly proposed by Michiels et al. [18] and later modified by Durante et al. [4]. We also improve this method to select the better fit of the given data among some extreme value copulas based on the Pickands dependence function. The EV copulas exhibit a similar upper tail dependence structure in terms of the tail concentration function. Thus, the tail concentration function proposed in Durante et al. [4] may fail to detect the tail dependence structure for the extreme value copulas for the same dependence level. In Figure 1, tail concentration functions of five EV copulas with the same Kendall's tau ($\tau = 0.5$) are presented. From this figure, it is hard to distinguish tail concentration function visually for the same dependence level among all EV copulas. For this reason, we prefer using the Pickands function in order to select the best suited extremely distributed random variables in the graphical method proposed in Durante et al. [4]. The extreme value copula is characterized by the Pickands dependence function; therefore, it can be useful in determining the best-fitted model for the bivariate extreme events. Although the test statistic proposed by Genest et al. [8] are consistent and effective tools for distinguishing between the symmetric and asymmetric extreme value copulas, the processing time is drawn out when dealing with big data because the test procedure involves the bootstrap method for estimating the p-values of the test statistic. For all these reasons, the graphical method based on the Pickands function can be used for determining the best-fitted EV copula for underlying data.

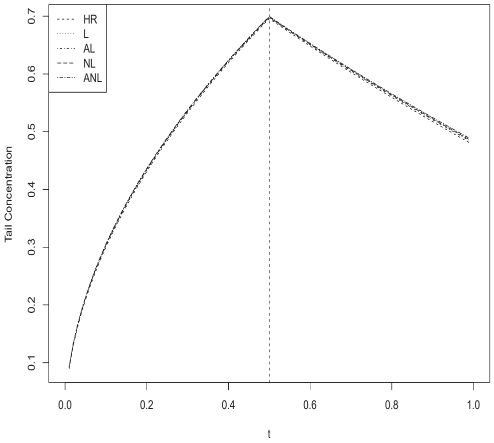


Figure 1. Tail concentration function for EV copulas with $\tau = 0.5$

The remainder of the study is organized as follows. In section 2, some EV copulas with Pickands dependence functions are introduced. In section 3, a graphical method to select EV copulas is presented. Some advantages of the proposed methods are discussed. Also, we performed a graphical method to show how accurately it works for a simulated data set from the EV copulas. In section 4, we apply the proposed method to the Danube dataset. Finally, the conclusion is given in the last section.

2. Some parametric extreme value copulas

Constructing a Pickands dependence function is one of the popular methods to obtain an EV copula. In this section, five EV copulas are introduced. Logistic (L) or Gumbel-Hougaard copula dating back to Gumbel [12] and Hougaard [13] can be considered as one of the oldest bivariate extreme value models. The logistic copula is the only copula that is at the same time as the extreme value and Archimedean copula. The Pickands dependence function of the Logistic copula with the dependence parameter θ given by:

$$A_L(t) = (t^\theta + (1-t)^\theta)^{\frac{1}{\theta}}.$$

The bivariate Asymmetric Logistic (AL) copulas Pickands dependence function with the dependence parameter $1 \leq \theta < \infty$ and asymmetry parameters α, β is given by

$$A_{AL}(t) = (1-\alpha)(1-t) + (1-\beta)t + \left((\alpha t)^\theta + (\beta(1-t))^\theta\right)^{\frac{1}{\theta}},$$

where $0 \leq \alpha$, $\beta \leq 1$. the Asymmetric Logistic copula adds further exibility to the Logistic copula. Note that by taking $\alpha = \beta = 1$, we can obtain the Logistic model, and by allowing $\alpha = \beta$, the Asymmetric Logistic copula is symmetric. The complete dependence is obtained. The complete dependence is obtained when $\alpha=\beta=1$ and $\theta\to 0$. And, also the independence is obtained when $\theta=1$ and $\alpha=0$ or $\beta=0$.

The bivariate Pickands function of the Negative Logistic (NL) model dating back to Galambos [7] is given by

$$A_{NL}(t)=1-(t^{-\theta}+(1-t)^{-\theta})^{-\frac{1}{\theta}},$$

where $\theta\in[0,\infty)$. Independence is obtained as $\theta=0$ and complete dependence is obtained when $\theta\to\infty$.

The bivariate Asymmetric Negative Logistic copula dating back to Joe [15] is an extension of the Negative Logistic copula. The Joe copula has two parameters α and β which allow the model to be asymmetric. Pickands dependence function of Asymmetric Negative Logistic copula is given by

$$A_{ANL}(t)=1-((\alpha t)^{-\theta}+(\beta(1-t))^{-\theta})^{-\frac{1}{\theta}},$$

where $0\leq\alpha,\ \beta\leq 1$ and $\theta\in(0,\infty)$. Note that if $\alpha=\beta=1$, we obatain the Negative Logistic copula. If $\alpha=\beta$, then the Asymmetric Negative Logistic copula is symmetric. Independence is obtained as $\alpha=\beta=0$ or $\theta\to 0$ and complete dependence is obtained when $\alpha=\beta=1$ and $\theta\to\infty$.

The Pickands dependence function of the bivariate Húsler-Reiss copula with parameter $\theta>0$ is

$$A_{HR}(t)=(1-t)\phi(Z_{1-t})+t\phi(Z_t),$$

where $\phi(.)$ is the standard normal distribution function and $Z_t=\frac{1}{\theta}+\frac{\theta}{2}\log(\frac{t}{1-t})$. Independence is obtained as $\theta\to 0$ and complete dependence is obtained when $\theta\to\infty$. For more details, see Húsler [14].

For the basics of the multivariate extreme value distributions and their Pickands dependence function see Dutfoy et al. [5] and Breachmann [2].

3. Graphical tool to select extreme value copula

In this section, we present a graphical tool that can help in the selection of the appropriate EV copula for underlying data set. Let $(X_i,Y_i)_{i=1}^n$ be a random sample from the EV copulas and $(U_i,V_i)_{i=1}^n$ be associated with the pseudo-observations. Consider a set of m EV copula's Pickands dependence function $A_1(.),\ldots,A_m(.)$ which belong to a different EV copula. A dissimilarity between the empirical estimate of the Pickands function $A_n(.)$ and the parametric Pickands function $A_i(.)$ for $i=1,\ldots,m$ can be defined by

$$d(A_n,A_i)=\int_0^1|A_n(t)-A_i(t)|^2dt,\ i=1,\ldots,m. \tag{3.1}$$

Similarly, the dissimilarity between the i-th and the j-th Pickands function is computed as

$$d(A_i,A_j)=\int_0^1|A_i(t)-A_j(t)|^2dt,\ 1<i\neq j<m. \tag{3.2}$$

Let us give the procedure of graphical tool for selection of appropriate extreme value copula for the given data set. The procedure can be provided by following:

- For $i=1,\ldots,m$ estimate dependence parameter(s) of a Pickands dependence function $A_i(.)$ from the family of the EV copula.
- For $i=1,\ldots,m$ compute the dissimilarity between $A_i(.)$ and the corrected empirical estimate $\Delta_{(emp,i)}=d(\hat{A}_{CFG},A_i)$ by using Eq. (3.1).
- For the m EV copulas Pickands function $A_1,\ldots,A_m$ compute mutual dissimilarities between $\Delta_{(i,j)}=d(A_i,A_j)$ by using Eq. (3.2).

- Symmetric square matrix of the dimension $m+1$, $D=\sigma_{(i,j)}$ can be defined as the following:

$$\begin{aligned}
&\sigma_{(1,j)}=\Delta_{(emp,j+1)},\ j=2,\dots,m+1,\\
&\sigma_{(i,j)}=\Delta_{(i-1,j-1)},\ i,j=2,\dots,m+1,\ i<j,\\
&\sigma_{(i,i)}=0,\ i=1,\dots,m+1.
\end{aligned}$$

- Using the dissimilarity matrix D, a non-metric multidimensional scaling (NMDS) technique can be performed.

Dissimilarity matrix D contains L^2- type distances which contain the information about the relation among the $A_n(.)$ (empirical Pickands function), $A_1(.),\dots,A_m(.)$. In order to obtain a two-dimensional representation through the ranking of distances between A_n and $A_1,\dots,A_m$, a a non-metric multidimensional scaling (MDS) technique can be performed on D. Finally, the m points $p_i=(x_i,y_i)$ corresponding to Pickands function A_i and $p_{emp}=(x_{emp},y_{emp})$ corresponding to the empirical Pickands dependence function estimation A_n can be visualized in a two dimensional graph.

For Figures 1-5, we apply the NMDS method based on the Pickands function for each generated data sets from EV copulas. The procedure provides a graphical representation of the empirical Pickands function and the five fitted EV copulas (L: Logistic, AL: Asymmetric logistic, NL: Negative logistic, ANL: Asymmetric negative logistic, and HR: Husler-Reiss) in two dimensions for a stress a value lower than 100th of a percent of 0.05. As can be seen from Figures 1-5, the charts are often useful to determine the true data generating process except for asymmetric EV copulas.

Now, in order to assess the performance of graphical method for EV copulas, we conduct simulation study. Let the five points $p_i=(x_i,y_i)$, $i=1,\dots,5$ be corresponding to Pickands dependence function $A_i(.)$ of five EV copulas and $p_{emp}=(x_{emp},y_{emp})$ be corresponding to empirical estimation of Pickands dependence function, which are obtained by NMDS method in a 2D graph. We may define an Euclidean distances d_i^2 from the points p_i, $i=1,\dots,5$ to p_{emp} given by following:

$$d_i^2=(x_{i;1}-x_{emp,1})^2+(x_{i;2}-x_{emp,2})^2,\ i=1,\dots,5.$$

Thus, the point p_i, corresponding to Pickands function A_i, with smallest distance d_i^2 is the best choice for given data among all possible five EV copulas. By repeating this process K times for the randomly generated EV copula then we can measure the performance of the graphical method. Let $(X_{i,k},Y_{i,k})_{i=1,\dots,n}^{k=1,\dots,K}$ be K Monte Carlo samples of size n from EV copula. Also, $P_{i,k}=(x_{i,k},y_{i,k})_{i=1,\dots,5}^{k=1,\dots,K}$ and $P_{emp,k}=(x_{emp,k},y_{emp,k})^{k=1,\dots,K}$ be the points obtained by NMDS method in a 2D graph. The simulation procedure goes as follows. We can define Euclidean distances in 2D space for K Monte Carlo samples from EV copula as following:

$$d_{i,k}^2=(x_{i,k;1}-x_{emp,k;1})^2+(x_{i,k;2}-x_{emp,k;2})^2,\ i=1,\dots,5,\ k=1,\dots,K.$$

We can calculate the ranks of $d_{i,k}^2$ associated to index i for all K Monte Carlo samples given by $r_{i,k}$. Hence, the smallest rank of $r_{i,k}$, $k=1,\dots K$ indicates that the Pickands dependence function $A_i(.)$ is as close as to empirical Pickands dependence function $A_n(.)$ than other Pickands dependence function for the Monte Carlo samples of $k=1,\dots,K$ in 2D graph. For the overall performance, we define the mean of ranks $r_{i,k}$ as $\overline{r}_i=\sum_{k=1}^{K}r_{i,k}/K,\ i=1,\dots,5$ for all EV copulas.

Let us consider the bivariate random data from the EV copulas. We simulate the bivariate 1000 Monte sample of sizes 250 and 500 from the Logistic, Asymmetric logistic, Negative logistic, Asymmetric negative logistic, and Húsler-Reiss EV copula models by using the following combinations:

TABLE 1. Mean of the ranks for different EV copulas with $n = 250$

True Copula	$\overline{r}_L$	$\overline{r}_{AL}$	$\overline{r}_{HR}$	$\overline{r}_{NL}$	$\overline{r}_{ANL}$
$L(\theta = 0.1)$	**1.4713**	2.9713	1.7459	3.8504	4.9610
$L(\theta = 0.9)$	**2.4090**	3.4545	2.8363	2.5727	3.7272
$AL(\theta = 0.1, \alpha = 0.2, \beta = 0.8)$	2.19375	4.5687	2.1375	**1.6750**	4.4250
$AL(\theta = 0.1, \alpha = 0.8, \beta = 0.2)$	2.2083	4.5416	2.5000	**1.5833**	4.1666
$AL(\theta = 0.1, \alpha = 0.5, \beta = 0.5)$	2.870	**2.118**	4.538	3.223	2.251
$AL(\theta = 0.9, \alpha = 0.2, \beta = 0.8)$	**2.73**	2.99	2.92	2.98	3.38
$AL(\theta = 0.9, \alpha = 0.8, \beta = 0.2)$	2.9629	3.0740	**2.6913**	2.9135	3.3580
$AL(\theta = 0.9, \alpha = 0.5, \beta = 0.5)$	2.4375	**2.3125**	3.1250	3.1875	3.9375
$NL(\theta = 10)$	3.8619	2.9079	2.0083	**1.2887**	4.9330
$NL(\theta = 1)$	2.4814	4.0370	2.2592	**1.9629**	4.2592
$ANL(\theta = 1, \alpha = 0.2, \beta = 0.8)$	**2.2777**	3.5222	2.6222	2.3888	4.1888
$ANL(\theta = 1, \alpha = 0.8, \beta = 0.2)$	2.42	3.52	2.46	**2.30**	4.30
$ANL(\theta = 1, \alpha = 0.5, \beta = 0.5)$	**1.8401**	3.3848	3.5555	2.4986	3.7208
$ANL(\theta = 10, \alpha = 0.2, \beta = 0.8)$	2.1645	4.6195	2.1413	**1.7146**	4.3598
$ANL(\theta = 10, \alpha = 0.8, \beta = 0.2)$	2.181	4.614	2.156	**1.703**	4.346
$ANL(\theta = 10, \alpha = 0.5, \beta = 0.5)$	2.341	2.909	4.529	3.175	**2.046**
$HR(\theta = 0.1)$	3.1612	3.2096	**2.6935**	2.7580	3.1774
$HR(\theta = 0.9)$	2.5116	4.2558	**2.1395**	2.2558	3.8372

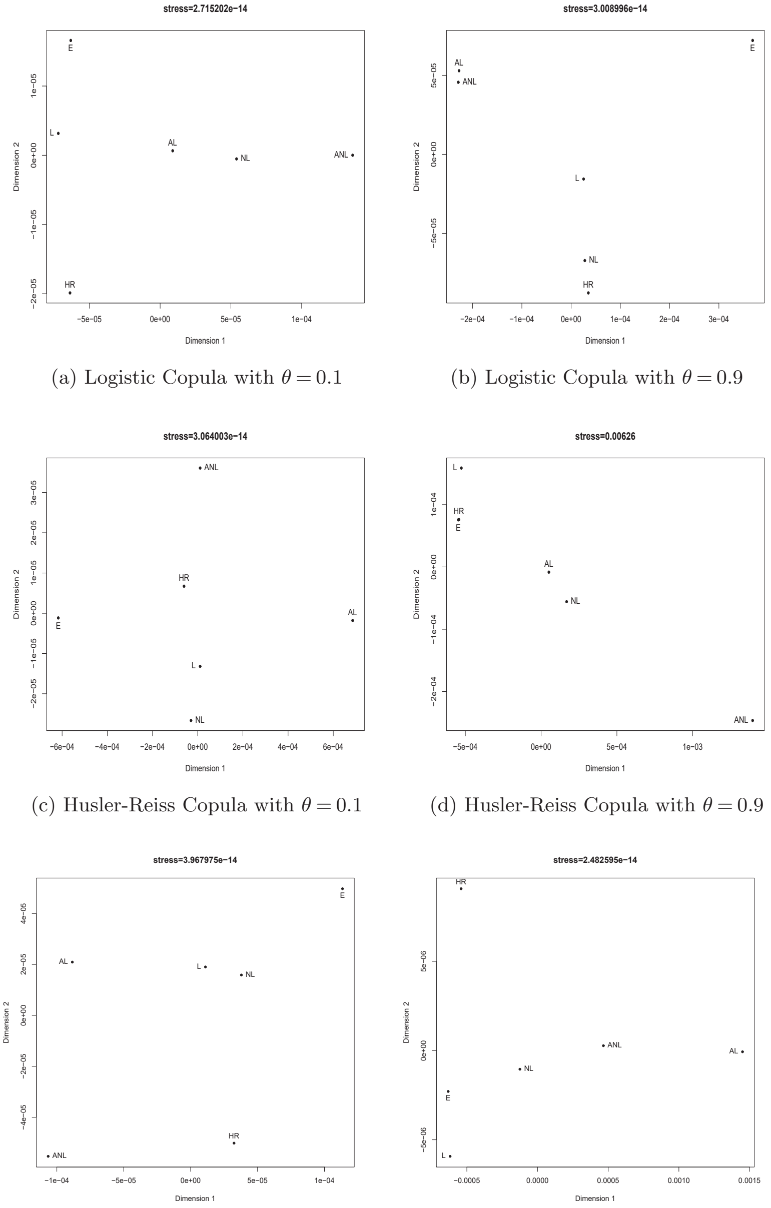
1. Logistic copula with dependence parameters $\theta = 0.1$ (Strong dependence), $\theta = 0.9$ (Mild dependence)
2. Asymmetric logistic copula
   (a) $\theta = 0.1, \alpha = 0.2, \beta = 0.8$ (Strong dependence and asymmetric Pickands function with $\alpha < \beta$)
   (b) $\theta = 0.1, \alpha = 0.8, \beta = 0.2$ (Strong dependence and asymmetric Pickands function with $\alpha > \beta$)
   (c) $\theta = 0.1, \alpha = 0.5, \beta = 0.5$ (Strong dependence and asymmetric Pickands function with $\alpha = \beta$)
   (d) $\theta = 0.9, \alpha = 0.2, \beta = 0.8$ (Mild dependence and asymmetric Pickands function with $\alpha < \beta$)
   (e) $\theta = 0.9, \alpha = 0.8, \beta = 0.2$ (Mild dependence and asymmetric Pickands function with $\alpha > \beta$)
   (f) $\theta = 0.9, \alpha = 0.5, \beta = 0.5$ (Mild dependence and asymmetric Pickands function with $\alpha = \beta$)
3. Negative logistic copula with the dependence parameters $\theta = 10$ (Strong dependence), $\theta = 1$ (Mild dependence)
4. Asymmetric negative logistic copula
   (a) $\theta = 10, \alpha = 0.2, \beta = 0.8$ (Strong dependence and asymmetric Pickands function with $\alpha < \beta$)
   (b) $\theta = 10, \alpha = 0.8, \beta = 0.2$ (Strong dependence and asymmetric Pickands function with $\alpha > \beta$)
   (c) $\theta = 10, \alpha = 0.5, \beta = 0.5$ (Strong dependence and asymmetric Pickands function with $\alpha = \beta$)
   (d) $\theta = 1, \alpha = 0.2, \beta = 0.8$ (Mild dependence and asymmetric Pickands function with $\alpha < \beta$)
   (e) $\theta = 1, \alpha = 0.8, \beta = 0.2$ (Mild dependence and asymmetric Pickands function with $\alpha > \beta$)
   (f) $\theta = 1, \alpha = 0.5, \beta = 0.5$ (Mild dependence and asymmetric Pickands function with $\alpha > \beta$)

TABLE 2. Mean of the ranks for different EV copulas with $n = 500$

True Copula	$\overline{r}_L$	$\overline{r}_{AL}$	$\overline{r}_{HR}$	$\overline{r}_{NL}$	$\overline{r}_{ANL}$
$L(\theta = 0.1)$	**1.360**	2.784	1.973	3.934	4.949
$L(\theta = 0.9)$	**2.0458**	4.0917	2.8990	2.2385	3.7247
$AL(\theta = 0.1, \alpha = 0.2, \beta = 0.8)$	2.2248	4.5574	2.2129	**1.5741**	4.4306
$AL(\theta = 0.1, \alpha = 0.8, \beta = 0.2)$	2.2299	4.5328	2.1934	**1.5839**	4.4598
$AL(\theta = 0.1, \alpha = 0.5, \beta = 0.5)$	2.509	**1.771**	4.800	3.567	2.353
$AL(\theta = 0.9, \alpha = 0.2, \beta = 0.8)$	**2.4117**	3.3176	3.1529	2.7058	3.4117
$AL(\theta = 0.9, \alpha = 0.8, \beta = 0.2)$	**2.5022**	3.3452	3.1748	2.6905	3.2869
$AL(\theta = 0.9, \alpha = 0.5, \beta = 0.5)$	3.6692	**2.4307**	2.9000	2.6230	3.3769
$NL(\theta = 10)$	3.9509	2.6666	2.3039	**1.0980**	4.9803
$NL(\theta = 1)$	2.3617	4.0265	2.6648	**1.8510**	4.0957
$ANL(\theta = 1, \alpha = 0.2, \beta = 0.8)$	**1.820**	4.454	2.508	1.857	4.361
$ANL(\theta = 1, \alpha = 0.8, \beta = 0.2)$	1.9	4.3	2.4	**1.7**	4.7
$ANL(\theta = 1, \alpha = 0.5, \beta = 0.5)$	**1.6170**	3.4361	4.0212	2.6276	3.2978
$ANL(\theta = 10, \alpha = 0.2, \beta = 0.8)$	2.1415	4.5752	4.5752	**1.6106**	4.4247
$ANL(\theta = 10, \alpha = 0.8, \beta = 0.2)$	2.2788	4.5576	2.1538	**1.5769**	4.4326
$ANL(\theta = 10, \alpha = 0.5, \beta = 0.5)$	2.462	2.475	4.797	3.544	**1.722**
$HR(\theta = 0.1)$	3.2142	3.0357	**2.9285**	3.0000	2.8214
$HR(\theta = 0.9)$	2.6736	4.621	**1.7568**	1.9847	3.8736

5. Húsler-Reiss copula with dependence parameters $\theta = 0.9$ (Strong dependence), $\theta = 0.1$ (Mild dependence)

Tables 1-2 represent $\overline{r}_L$, $\overline{r}_{AL}$, $\overline{r}_{HR}$, $\overline{r}_{NL}$ and $\overline{r}_{ANL}$ which are obtained from 1000 Monte Carlo samples with 250 and 500 sizes, respectively. From these tables, if the True EV copula possesses to symmetric dependence structure, graphical method performs well. As an example, when the data is generated from Logistic copula with $\theta = 0.1$ and $n = 250$, smallest value of $\overline{r}_i$, $i = 1, \ldots, 5$ is $\overline{r}_L = 1.4713$ among five EV copulas. This means that the points which correspond to Pickands dependence function of Logistic copula $(A_L(.))$ in 2D graph is the closest to points correspond to A_n in 2D graph. Also, we can conclude from the Table 1-2, if the true EV copula possesses to asymmetric dependence structure, the graphical method does not perform well except for EV copula with equal asymmetry parameters. Also, mean of the ranks is decreased when the sample of size is increased.



(a) Logistic Copula with $\theta = 0.1$

(b) Logistic Copula with $\theta = 0.9$

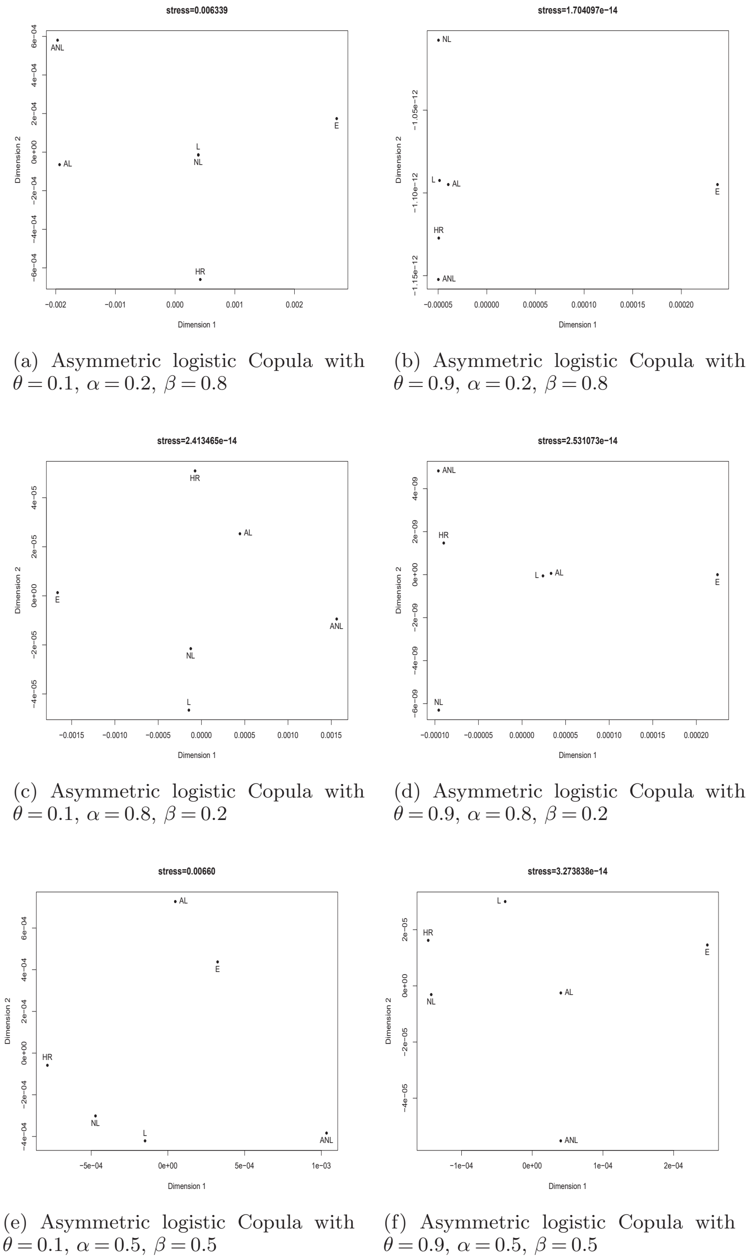
(c) Husler-Reiss Copula with $\theta = 0.1$

(d) Husler-Reiss Copula with $\theta = 0.9$

(e) Negative logistic Copula with $\theta = 1$

(f) Negative logistic Copula with $\theta = 10$

FIGURE 2. Graphical representation based on the two dimensional NMDS with data generated by Logistic, Husler-Reiss and Negative logistic copula



(a) Asymmetric logistic Copula with $\theta = 0.1$, $\alpha = 0.2$, $\beta = 0.8$

(b) Asymmetric logistic Copula with $\theta = 0.9$, $\alpha = 0.2$, $\beta = 0.8$

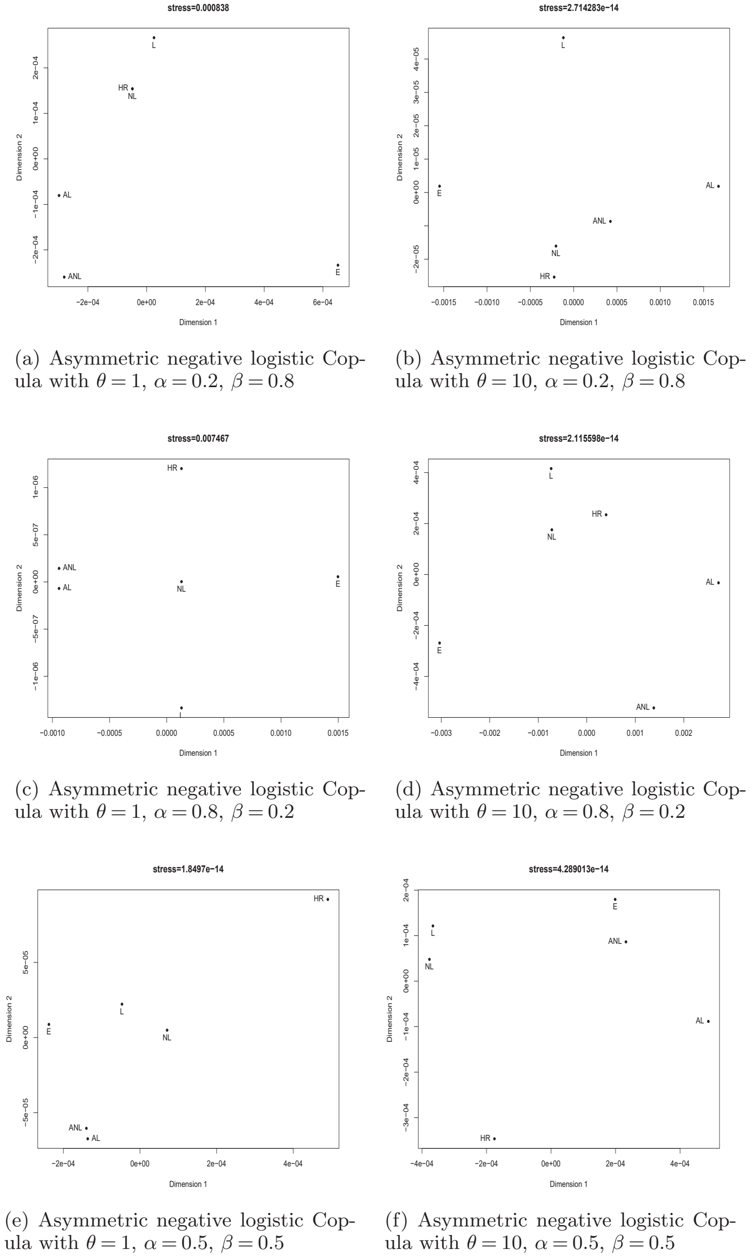
(c) Asymmetric logistic Copula with $\theta = 0.1$, $\alpha = 0.8$, $\beta = 0.2$

(d) Asymmetric logistic Copula with $\theta = 0.9$, $\alpha = 0.8$, $\beta = 0.2$

(e) Asymmetric logistic Copula with $\theta = 0.1$, $\alpha = 0.5$, $\beta = 0.5$

(f) Asymmetric logistic Copula with $\theta = 0.9$, $\alpha = 0.5$, $\beta = 0.5$

FIGURE 3. Graphical representation based on the two dimensional NMDS with the data generated by the asymmetric logistic copula



(a) Asymmetric negative logistic Copula with $\theta = 1$, $\alpha = 0.2$, $\beta = 0.8$

(b) Asymmetric negative logistic Copula with $\theta = 10$, $\alpha = 0.2$, $\beta = 0.8$

(c) Asymmetric negative logistic Copula with $\theta = 1$, $\alpha = 0.8$, $\beta = 0.2$

(d) Asymmetric negative logistic Copula with $\theta = 10$, $\alpha = 0.8$, $\beta = 0.2$

(e) Asymmetric negative logistic Copula with $\theta = 1$, $\alpha = 0.5$, $\beta = 0.5$

(f) Asymmetric negative logistic Copula with $\theta = 10$, $\alpha = 0.5$, $\beta = 0.5$

FIGURE 4. Graphical representation based on the two dimensional NMDS with the data generated by the asymmetric negative logistic copula

4. Real data example

To demonstrate the graphical method for the selection of the best-fitted EV copulas, in this section, we fit the EV copulas to the Danube data set which is available in the R package copula. According to this package, the Danube dataset contains ranks of base of observations from the Global River Discharge project of the Oak Ridge National Laboratory Distributed Active Archive Centre (ORNL DAAC), a NASA data centre. The measurements are the monthly average of rate for two stations situated at Scharding (Austria) on the Inn River and Nagymaros (Hungary) on the Danube.

Table 3. Estimation of dependence and asymmetry parameters for five EV copulas

EV Copula	CvM	α	β	θ
L	8.233616×10^{-5}	—	—	0.4872
AL	0.000333	0.8582	0.9992	0.4445
NL	9.257569×10^{-5}	—	—	1.3332
ANL	0.001962	0.9158	0.9995	1.0034
HR	0.000152	—	—	1.7981

The scatter plot of the pseudo-observations of the Danube data set is displayed in Figure 5. From Figure 5, symmetrical dependence structures are observed. Also, the Danube data set has a heavy right tail dependence structure.

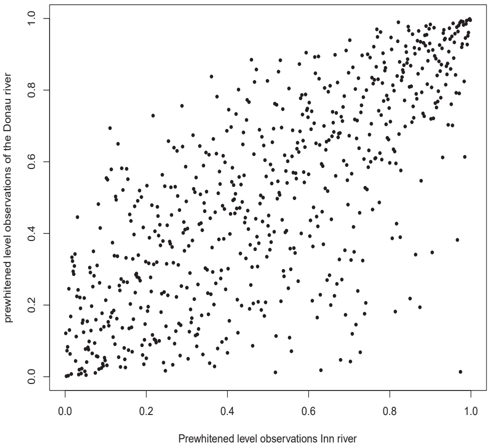
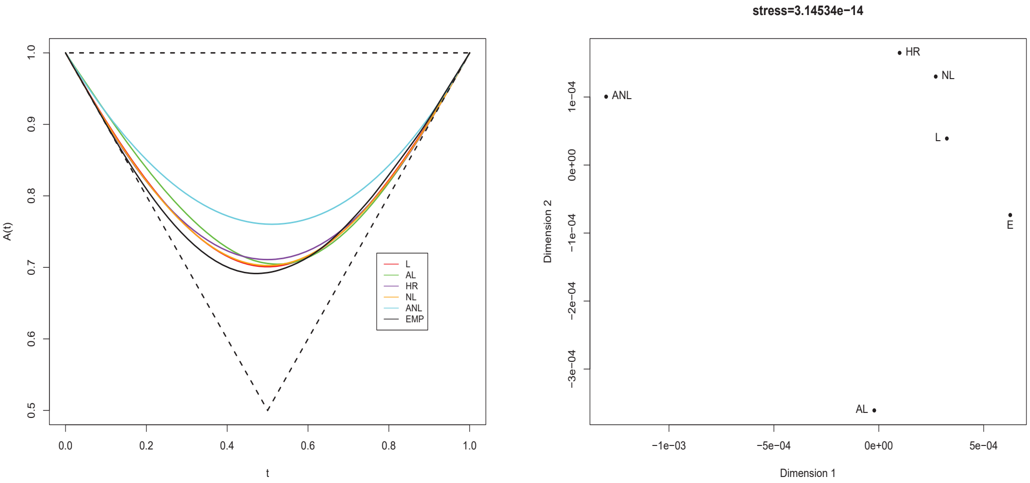


Figure 5. Scatter plot of danube dataset



(a) Parametric and Non-parametric estimation of the Pickands function for the danube data

(b) Graphical representation based on two dimensional NMDS for danube dataset

Figure 6. Fiiting results for danube dataset

In Figure 6(a), the parametric and non-parametric estimation of the Pickands dependence functions is displayed. From this figure, we can observe that the Logistic copula's Pickands dependence the function is much closer to the empirical Pickands dependence function for the Danube dataset. Also Table 1 represents the CvM distances between $A_n, A_L, \ldots, A_{HR}$, and estimation of dependence parameters for five EV copulas. On the other hand, Figure 6(b) displays the two-dimensional representation of the EV copula test spaces with the Danube dataset based the on NMDS method. When Figure 6(b) is examined; it can be concluded that the Logistic copulas are the most appropriate EV copulas for the Danube data set.

5. Conclusions

In this study, we proposed a graphical method based on NMDS to select the best-fitted EV copulas for underlying data. Also, we discussed some advantages of the proposed methods. If practitioners are interested in modelling for extreme situations which consist of a big data size, the graphical method can be useful to select the EV copulas. We performed the graphical method to see how accurately it works for the simulated data set from EV copulas. From the simulation study, when the dependence structure is symmetric, the procedure is useful to identify the true EV copula which is data generated. On the other hand when the data has asymmetric dependence the structure of the graphical procedure fails except for with the Asymmetric EV copula with equal asymmetry parameters. This problem can be overcome by using the Bernstein polynomial based on the Pickands dependence function estimator in the procedure of the graphical method. The main advantage of Bernstein polynomials is their flexibility against data that has a complex structure. So, Bernstein polynomials can take on an extremely wider range of shapes than simple estimators. Also, to demonstrate the graphical method for the selection of the best-fitted EV copulas, we fitted the EV copulas to the real data set. We have shown that the graphical procedure can lead to acceptable results.

References

[1] Ahmadabadi, A. and Ucer, B.H. (2017). Bivariate nonparametric estimation of the Pickands dependence function using Bernstein copula with kernel regression approach. *Computitional Statistics*, 32, 1515-1532.

[2] Brechmann, E.C. (2013). *Properties of extreme-value copulas.* [Thesis, Technical University of Munich].

[3] Capéraá, P., Fougres, A.L. and Genest, C. (1997). A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, 84, 567-577.

[4] Durante, F., Fernández-Sánchez, J. and Pappadá, R. (2015). Copulas, diagonals and tail dependence. *Fuzzy Sets and Systems*, 264, 22-41.

[5] Dutfoy, A., Parey, S. and Roche, N. (2017). Multivariate extreme value theory-A tutorial with applications to hydrology and meteorology. *Dependence Modeling*, 2, 30-48.

[6] Frees, E. and Valdez, E. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2, 1-25.

[7] Galambos, J. (1975). Order statistics of samples from multivariate distributions. *Journal of the American Statistical Association*, 70, 674-680.

[8] Genest, C., Kojadinovic, I., Néslehová, J. and Yan, J. (2011). A goodness-of-fit test for bivariate extreme-value copulas. *Bernoulli*, 17(1), 253-275.

[9] Genest, C. and Segers, J. (2009). Rank-Based inference for bivariate extreme value copulas. *The Annals of Statistics*, 37, 2990-3022.

[10] Gudendorf, G. and Segers, J. (2011). Nonparametric estimation of an extreme-value copula in arbitrary dimensions. *Journal of Multivariate Analysis*, 102(1), 37-47.

[11] Guillotte, S. and Perron, F. (2016). Polynomial pickands functions. *Bernoulli*, 22(1), 213-241.

[12] Gumbel, E.J. (1960). Distributios des valeurs extrémes en plusiers dimensions. *Publications de l'Institut de Statistique de l'Université de Paris*, 9, 171-173.

[13] Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika*, 73, 671-678.

[14] Husler, J. (1986). Extreme values of non-stationary random sequences. *Journal of Applied Probability*, 23, 937-950.

[15] Joe, H. (1990). Multivariate concordance. *Journal of Multivariate Analysis*, 35, 12-30.

[16] Marcon, G., Padoan, S.A. and Antoniano-Villalobos, I. (2016). Bayesian inference for the extremal dependence. *Electronic Journal of Statistics*, 10(2), 3310-3337.

[17] Marcon, G., Padoan, S.A., Naveau, P., Muliere, P. and Segers, J. (2017). Multivariate nonparametric estimation of the Pickands dependence function using Bernstein polynomials. *Journal of Statistical Planning and Inference*, 183, 1-17.

[18] Michiels, F. and De Schepper, A. (2013). A new graphical tool for copula selection. *Journal of Computational and Graphical Statistics*, 22(2), 471-493.

[19] Pickands, J. (1981). Multivariate extreme value distribution. *In Proceedings of the International Statistical Institute*, 859-878.

[20] Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *Publications de lInstitut de Statistique de lUniversite de Paris*, 8, 229-231.

[21] Vettori, S., Huser, R. and Genton, M.G. (2018). A comparison of dependence function estimators in multivariate extremes. *Statistics and Computing*, 28(3), 525-538.

İSTATİSTİK: JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION
Vol. 14, No. 2, December 2022, pp. 51–73
ISSN 1300-4077 | 22 | 2 | 51 | 73

İSTATİSTİK

A LIFETIME REGRESSION ANALYSIS WITH UNIT LINDLEY-WEIBULL DISTRIBUTION

Ahmet Pekgör
Department of Statistics,
Necmettin Erbakan University,
42090, Konya, Turkey

Coşkun Kuş, Kadir Karakaya*, Buğra Saraçoğlu
Department of Statistics,
Selcuk University,
42250, Konya, Turkey

İsmail Kınacı
Department of Actuarial Science,
Selcuk University,
42250, Konya, Turkey

Abstract: In this paper, a new lifetime distribution is introduced. Motivation is provided to obtain this distribution. The closed-form expressions of probability density and cumulative distribution functions are provided. Several distributional properties are obtained and the statistical inference are discussed on unknown parameters. The most important novelty of this study is to bring a lifetime regression analysis with the re-parameterized log-transform of the new distribution.

Key words: Unit-Lindley distribution, Weibull distribution, Monte Carlo simulation, Estimation, Lifetime regression, Confidence interval.

1. Introduction

In the last two decades, many statistical distributions have been introduced. Most of them are derived from various compounding methods. [8] introduced a Beta-normal distribution with cumulative distribution function (cdf) $R(G(x))$, where R is beta cdf, and G is normal cdf [15], [16], and [17] have introduced new distributions by using Gumbel, Fréchet and exponential cdfs for G in $R(G(x))$. All these distributions belong to Beta-G family. However, all these works don't give cdf in explicit form because of the structure of beta cdf.

In order to get an explicit cdf, [5] considers the Kumaraswamy cdf for R in [8]'s formula and they obtained different distributions by changing cdf G. The distribution family in [5] is called ”Kw-G family” [7] and [19] introduced new distributions by using Kw-G family.

[2] introduced a new family of distribution by getting inspired by [8]'s formula. They consider cdf $R(W(G(x)))$, where R and G are any cdf of continuous random variables and W is a function that satisfies certain conditions. It is noted that, If $W(x)=x$ and R and G are assigned as beta and normal cdfs, respectively, then the distribution in [8] is achieved.

In this paper, we introduce a new distribution, which is a member of [2] family. Some general distributional and inferential properties of the introduced distribution are studied. Here, there are two crucial discussions on statistical inference.

The first discussion is related to the confidence intervals (CIs) for unknown parameters. In general, the CIs for unknown parameters are discussed through asymptotical normality of maximum likelihood estimates (MLEs). Here, the CIs based on asymptotical normality of MLEs are denoted

* Corresponding author. E-mail address:kkarakaya@selcuk.edu.tr

by AN CIs. However, the limits of AN CIs sometimes turn out to be outside of the parameter space. It is an undesired outcome in practice. It should also be remembered that a large sample is needed to good approximation to the normality of MLEs when the number of the parameter is more than two. Furthermore, it is also needed a large sample to get true coverage probabilities (CPs) of AN CIs. In Subsection 3.2, uncorrected likelihood ratio (ULR) type CIs for the unknown parameters are discussed as an alternative to AN CIs. It is pointed out that the ULR type CIs have wonderful properties: The limits of ULR type CIs are always within parameter space. In the simulation given in Subsection 3.2, it is also observed that the CPs of ULR CIs are better than the CPs of AN CIs.

The second discussion is focused on the lifetime regression issue in the survival data analysis: In the lifetime regression analysis, a functional relationship between the dependent variable (lifetime or log-lifetime) and covariates are studied. A common assumption that there is a linear relationship between the location parameter and covariates in the models. These models can be used to determine the sign and magnitudes of covariates on the log-lifetimes through the location parameter. In practice, the survival data obeys to distribution, which has various types of failure rate functions. From this point of view, there is a demand for new lifetime distributions in the survival analysis.

In this study, a new lifetime distribution is introduced by using the [2]'s method. In order to obtain a new distribution with explicit cdf, W, R and G are assigned by an identity function, Unit-Lindley cdf and Weibull cdf, respectively. In Section 2, a new distribution is described with motivation and exact moments are obtained. In addition, the properties of hazard function and stochastic ordering are studied. An accepting rejecting algorithm is also provided to generate data from the new distribution. In Section 3, the several point estimators and CIs of unknown parameters are discussed through Monte Carlo simulation studies. In Section 4, a lifetime regression analysis based on introduced distribution is studied, and an extensive simulation study is performed. A practical real data set is given to illustrate the applicability of the new distribution in Section 5.

2. Unit-Lindley-Weibull distribution

In this section, we introduce a new distribution and discuss its distributional properties. Recently, Unit-Lindley (UL) distribution is introduced by [14]. If T is UL random variable, the pdf and cdf of T are given, respectively, by

$$r(t;\theta)=\left(\frac{\theta^2}{(1+\theta)(1-t)^3}\right)\exp\left(\frac{\theta t}{t-1}\right)\mathbb{I}_{(0,1)}(t)$$

and

$$R(t;\theta)=1-\left(1-\frac{\theta t}{(1+\theta)(t-1)}\right)\exp\left(\frac{\theta t}{t-1}\right),$$

where $\theta>0$ is a parameter and $\mathbb{I}_A(\cdot)$ is an indicator function on A. Let us also consider a Weibull random variable Y with pdf and cdf

$$g(y;\alpha,\beta)=\frac{\beta}{\alpha}\left(\frac{y}{\alpha}\right)^{\beta-1}\exp\left(-\left(\frac{y}{\alpha}\right)^{\beta}\right)\mathbb{I}_{\mathbb{R}_+}(y) \tag{2.1}$$

and

$$G(y;\alpha,\beta)=1-\exp\left(-\left(\frac{y}{\alpha}\right)^{\beta}\right),$$

respectively. Let us assign W is an identity function and consider UL cdf and Weibull cdf for R and G in $F(x)=R(W(G(x)))$, a valid cdf is obtained by

$$F(x;\Xi)=1-\left(1+\frac{\theta\left(1-\exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right)\right)}{(1+\theta)\exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right)}\right)\exp\left(-\frac{\left(1-\exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right)\right)}{\exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right)}\right), \tag{2.2}$$

where $\Xi = (\alpha, \beta, \theta) \in \mathbb{R}_+^3$ is the parameter vector, α is a scale, β and θ are shape parameters. A distribution with cdf (2.2) is called unit-Lindley Weibull (ULW) and it is denoted by ULW(Ξ). Let X be the ULW(Ξ) random variable with cdf (2.2). Then, the pdf of X is given by

$$f(x; \Xi) = \frac{\beta \theta^2 x^{\beta-1}}{\alpha^\beta (1+\theta)} \exp\left\{-\theta \exp\left(\left(\frac{x}{\alpha}\right)^\beta\right) + \theta + 2\left(\frac{x}{\alpha}\right)^\beta\right\} \mathbb{I}_{\mathbb{R}_+}(x). \tag{2.3}$$

For some selected values of parameters, the pdf plots of ULW distribution are given in Figure 1. It is concluded from Figure 1, the pdf of ULW distribution can be unimodal or decreasing. It is also observed that the pdf can be skewed at right or left.

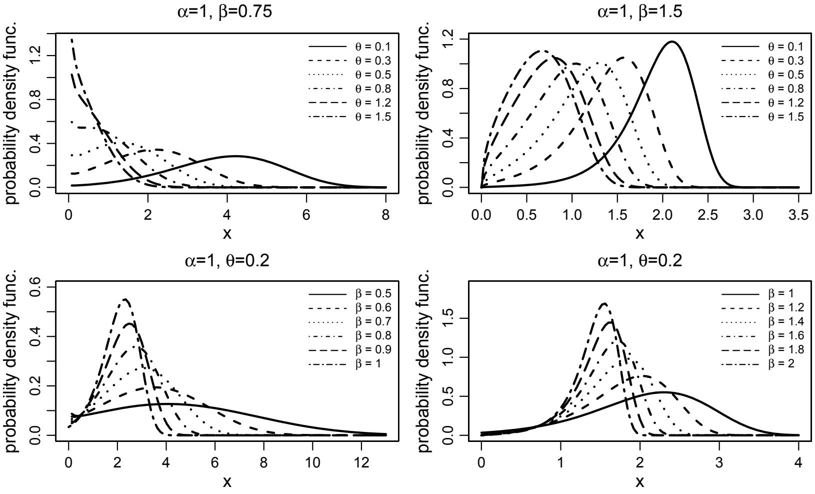


FIGURE 1. Probability density function plots for ULW distribution

2.1. Hazard function

The hazard function (hf) of ULW(Ξ) distribution can be written by

$$h(x; \Xi) = \frac{\beta x^{\beta-1}\theta^2}{\alpha^\beta \exp\left(-\left(\frac{x}{\alpha}\right)^\beta\right)\left(\exp\left(-\left(\frac{x}{\alpha}\right)^\beta\right) + \theta\right)} \mathbb{I}_{\mathbb{R}_+}(x).$$

For some selected values of parameters, the hf of ULW distribution is plotted in Figure 2. From Figure 2, it is observed that the hf of ULW distribution has increasing or bathtub shapes. In the following, we discuss these properties of hf. Let us consider the first-order derivative of hf

$$h'(x; \Xi) = \frac{\beta \theta^2 x^{\beta-2}\left(2\beta\left(\frac{x}{\alpha}\right)^\beta + \beta - 1\right)\exp\left(-\left(\frac{x}{\alpha}\right)^\beta\right) + \theta\left(\beta\left(\frac{x}{\alpha}\right)^\beta + \beta - 1\right)}{\alpha^\beta \exp\left(-\left(\frac{x}{\alpha}\right)^\beta\right)\left(\exp\left(-\left(\frac{x}{\alpha}\right)^\beta\right) + \theta\right)^2}.$$

It can be easily seen that $h'(x; \Xi) > 0$ for $\beta > 1$ and hence hf is increasing. In addition, $h'(x; \Xi) < 0$ for $x < \alpha\left(\frac{1-\beta}{2\beta}\right)^{1/\beta}$ and $h'(x; \Xi) > 0$ for $x > \alpha\left(\frac{1-\beta}{\beta}\right)^{1/\beta}$ under the condition $\beta < 1$. According to this discussion, it can be observed that hf decrease at first and increases as time progress for $\beta < 1$. Furthermore, from Figure 2, it is observed that the hf exhibits bath-tube type when $\beta < 1$.

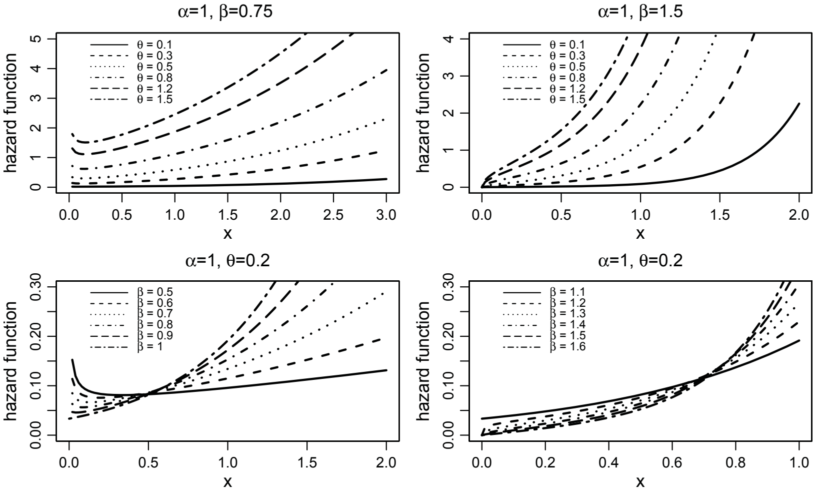


FIGURE 2. Hazard function plots for ULW distribution

2.2. Motivation for the ULW distribution

The pdf of the ULW distribution given in Eq. (2.3) can be obtained in a different ways using the method of [18]. Let Y be the Weibull random variable with pdf $g(x;\alpha,\beta)$ given in Eq. (2.1). According to [18], the pdf of the weighted random variable Y^w is defined by

$$f_w(y)=\frac{w(y;\alpha,\beta,\theta)}{E(w(Y;\alpha,\beta,\theta))}g(y;\alpha,\beta)\mathbb{I}_{\mathbb{R}_+}(y). \tag{2.4}$$

Let us consider $w(y;\alpha,\beta,\theta)=\exp\left\{-\theta\exp\left(\left(\frac{y}{\alpha}\right)^{\beta}\right)+\theta+3\left(\frac{y}{\alpha}\right)^{\beta}\right\}$ in Eq. (2.4) and we get

$$E(w(Y;\alpha,\beta,\theta))=\frac{\theta+1}{\theta^2}.$$

Thus, the pdf of Y^w is identical to the pdf of introduced ULW(Ξ) distribution with pdf (2.3).

2.3. Moments

In this subsection, exact moments of ULW(Ξ) distribution under a certain conditon. Let us consider the result of [10] given by

$$\int_1^{\infty}(\log(x))^m x^{v-1}\exp(-\mu x)dx=\frac{\partial^m}{\partial v^m}\left\{\theta^{-v}\Gamma(v,\theta)\right\}.$$

Under the condition $r/\beta\in\mathbb{Z}^+$, the r^{th} moment of a random variable X having ULW(Ξ) is obtained by

$$\begin{aligned}E(X^r)&=\int_0^{\infty}x^r f(x)dx\\&=\int_0^{\infty}x^r\frac{\theta^2\beta\alpha^{-\beta}x^{\beta-1}}{1+\theta}\exp\left(-\theta\exp\left(\left(\frac{x}{\alpha}\right)^{\beta}\right)+\theta+2\left(\frac{x}{\alpha}\right)^{\beta}\right)dx\\&=\frac{\theta^2\alpha^r e^{\theta}}{1+\theta}\int_1^{\infty}(\log(t))^{r/\beta}t\exp(-\theta t)dx\end{aligned}$$

$$= \frac{\theta^2 \alpha^r e^\theta}{1+\theta} \times \frac{\partial^m}{\partial v^m} \left\{ \theta^{-v} \Gamma(v,\theta) \right\} \Big|_{v=2}, \ r \in \mathbb{N}_+$$
$$= \frac{\alpha^r e^\theta}{1+\theta} MeijerG([[1,1],[]],[[2],[0,0]],\theta),$$

where $\Gamma(v,\theta)$ is incomplete gamma function, $m = r/\beta$ and $MeijerG$ is the well-known Meijer G function which is available in Maple software. Some numerical values of first four moments are presented in Table 1.

Table 1. The first four moments of the ULW distribution

r	β	α	θ	$E(X^r)$
1	1	3	2	1.3613
2				2.8004
3				7.1059
4				20.6759
1	0.5	3	2	0.9334
2				2.2973
3				8.5639
4				41.3767

2.4. Stochastic ordering

For a positive continuous random variable, stochastic ordering and the other ordering are important tools for judging the comparative behavior. The following theorem shows that the ULW random variables can be ordered with respect to the likelihood ratio.

Theorem 1. *Let $X \sim ULW(\alpha,\beta,\theta_1)$ and $Y \sim ULW(\alpha,\beta,\theta_2)$. If $\theta_1 > \theta_2$ then X is smaller than Y in the likelihood ratio order, i.e., the ratio function of the corresponding pdfs is decreasing in x.*

Corollary 1. *It follows from [21] that X is also smaller than Y in the hazard ratio, mean residual life and stochastic orders under the conditions given in Theorem 1.*

2.5. Data generating algorithm

In this subsection, we give an algorithm to generate data from ULW(Ξ) distribution. Since the inverse transformation method does not give an explicit formula, we propose an acceptance-rejection (AR) sampling algorithm. In this algorithm, the Weibull distribution is chosen as a proposal distribution. The AR algorithm is given as follows:

Algorithm 1.

A1. Generate data on random variable Y from Weibull distribution with pdf g given in Eq. (2.1)

A2. Generate U from standard uniform distribution (independent of Y).

A3. If
$$U < \frac{f(Y;\Xi)}{k \times g(Y;\alpha,\beta)},$$
then set $X = Y$ ("accept"); otherwise go back to A1 ("reject"), where pdf f is given as in Eq. (2.3) and
$$k = \max_{z \in \mathbb{R}_+} \frac{f(z;\Xi)}{g(z;\alpha,\beta)}.$$

The output of this algorithm suggest a random data on X from ULW(Ξ) distribution. It is noted that Algorithm 1 is used for all simulations in the paper.

3. Statistical inference on distribution parameters

In this section, we propose several estimators for estimating the unknown parameters of the ULW(Ξ) distribution. We discuss the maximum likelihood, least-squares, weighted least squares, Cramér-von Mises type, and Anderson-Darling type estimation methods. Furthermore, the two types of CIs for the parameters are discussed. Simulation studies are also performed to observe the performances of the methods discussed here.

3.1. Point estimation

Let $X_1, X_2, \ldots, X_n$ be a random sample from the ULW(Ξ) distribution and $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ denotes the corresponding order statistics. Furthermore, $x_{(i)}$ denotes the observed value of $X_{(i)}$ for $i = 1, 2, \ldots, n$. Based on this sample, the log-likelihood function is given by

$$\begin{aligned}\ell(\Xi) = n\log\left(\beta\theta^2\right) - n\log\left(\alpha\left(1+\theta\right)\right) + (\beta-1)\sum_{i=1}^{n}\log\left(\frac{x_i}{\alpha}\right) \\ + \sum_{i=1}^{n}\log\left(\exp\left(-\theta\exp\left(\left(\frac{x_i}{\alpha}\right)^{\beta}\right) + \theta + 2\left(\frac{x_i}{\alpha}\right)^{\beta}\right)\right). \end{aligned} \quad (3.1)$$

Then, the MLEs of α, β and θ are given by

$$\widehat{\Xi}_1 = \arg\max_{\Xi}\left\{\ell\left(\Xi\right)\right\}, \quad (3.2)$$

where $\Xi = (\alpha, \beta, \theta)$ and $\widehat{\Xi}_1 = \left(\widehat{\alpha}, \widehat{\beta}, \widehat{\theta}\right)$. Let us define the following functions which are used to define the different type of estimators:

$$\begin{aligned} Q_{LS}\left(\Xi\right) &= \sum_{i=1}^{n}\left(F\left(x_{(i)}\right) - \frac{i}{n+1}\right)^2, \\ Q_{WLS}\left(\Xi\right) &= \sum_{i=1}^{n}\frac{(n+2)(n+1)^2}{i(n-i+1)}\left(F\left(x_{(i)}\right) - \frac{i}{n+1}\right)^2, \\ Q_{CvM}\left(\Xi\right) &= \frac{1}{12n} + \sum_{i=1}^{n}\left(F\left(x_{(i)}\right) - \frac{2i-1}{2n}\right)^2 \end{aligned}$$

and

$$Q_{AD}\left(\Xi\right) = -n - \frac{1}{n}\sum_{i=1}^{n}\left\{(2i-1)\log\left(F\left(x_{(i)}\right)\right)\right\} + \frac{1}{n}\sum_{i=1}^{n}\log\left\{1 - F\left(x_{(i)}\right)\right\},$$

where $F(\cdot)$ is cdf of ULW(Ξ) distribution given in Eq. (2.2). Then, the least squares estimator (LSE), weighted least squares estimator (WLSE), Anderson Darling estimator (ADE) and the Cramér-von Mises estimator (CvME) of Ξ are given, respectively, by

$$\widehat{\Xi}_2 = \arg\min_{\Xi}\left\{Q_{LS}\left(\Xi\right)\right\}, \quad (3.3)$$

$$\widehat{\Xi}_3 = \arg\min_{\Xi}\left\{Q_{WLS}\left(\Xi\right)\right\}, \quad (3.4)$$

$$\widehat{\Xi}_4 = \arg\min_{\Xi}\left\{Q_{AD}\left(\Xi\right)\right\}, \quad (3.5)$$

$$\widehat{\Xi}_5 = \arg\min_{\Xi}\left\{Q_{CvM}\left(\Xi\right)\right\}. \quad (3.6)$$

It is noted that these estimates are discussed before in [12], [13], and [23]. All maximization and minimization problems can be solved by some numerical methods such as Nelder-Mead, BFGS, or CG. These methods can be easily conducted by **optim** function in R.

3.2. Interval estimation

In this subsection, the CIs are discussed for the parameters θ, β and α. In the statistical literature, CIs are usually constructed by using a pivotal quantity based on MLEs of parameters. However, an exact CIs can not be obtained since the MLEs are usually obtained by a numerical method to optimize the likelihood. Consequently, asymptotic CIs based on the asymptotic normality of MLEs are most popular in the all fields of statistics and it has widespread usage. It is well-known that the AN of MLEs can be stated by

$$\widehat{\boldsymbol{\Xi}}_1 \overset{d}{\to} N_3\left(\boldsymbol{\Xi}, \mathbb{I}^{-1}\left(\boldsymbol{\Xi}\right)\right),$$

where $\widehat{\boldsymbol{\Xi}}_1$ is MLE of $\boldsymbol{\Xi}$ given in Eq. (3.2) and $\mathbb{I}\left(\boldsymbol{\Xi}\right)$ is Fisher information matrix. Using this result, the $100 \times (1-\alpha)\,\%$ AN CIs of parameters α, β and θ are constructed, respectively, by

$$\begin{aligned} &\widehat{\alpha} \pm z_{1-\frac{\alpha}{2}} \times se\left(\widehat{\alpha}\right), \\ &\widehat{\beta} \pm z_{1-\frac{\alpha}{2}} \times se\left(\widehat{\beta}\right), \\ &\widehat{\theta} \pm z_{1-\frac{\alpha}{2}} \times se\left(\widehat{\theta}\right), \end{aligned}$$

where z_a, is the a^{th} quantile of standard normal distribution, $se\left(\widehat{\alpha}\right)$, $se\left(\widehat{\beta}\right)$ and $se\left(\widehat{\theta}\right)$ are the roots of the diagonal member of $\mathbb{I}^{-1}\left(\widehat{\boldsymbol{\Xi}}_1\right)$ which is a consistent estimate of $\mathbb{I}^{-1}\left(\boldsymbol{\Xi}\right)$ and the $se\left(\cdot\right)$ stands for standard error.

By the way, there is another method called ULR, which is not used in most of statistical software, but it has interesting properties. AN and ULR CIs are asymptotically equivalent [9]. The ULR CIs are transformation invariant. It is range preserving that means the CIs it produces will always be inside of the parameter space. There is no need to compute/estimate the variance of the estimates, unlike to AN. In addition, the ULR method doesn't necessarily give symmetric intervals around MLE.

Under usual regularity assumptions on the likelihood function, if the θ is true parameter, then $-2\log\left(\ell\left(\theta, \widetilde{\boldsymbol{\lambda}}\right) - \ell\left(\widehat{\boldsymbol{\Xi}}_1\right)\right)$ distributed χ^2 with degrees of freedom 1, where $\boldsymbol{\lambda} = (\alpha, \beta)$ are the nuisance parameters, ℓ is the log-likelihood function as in Eq. (3.1), $\widehat{\boldsymbol{\Xi}}_1$ is the joint MLEs of (θ, β, α) given in Eq. (3.2), $\widetilde{\boldsymbol{\lambda}} = \left(\widetilde{\alpha}, \widetilde{\beta}\right)$ is the restricted MLEs of $\boldsymbol{\lambda}$ given a fixed value of θ. Using this fact, $100 \times (1-\alpha)\,\%$ ULR CI limits (θ_L, θ_U) that satisfy

$$\ell\left(\theta, \widetilde{\boldsymbol{\lambda}}\right) = \underbrace{\ell\left(\widehat{\boldsymbol{\Xi}}_1\right) - \frac{1}{2}\chi^2_{(1)}\left(1-\alpha\right)}_{\text{LR Bound}},$$

with $\theta_L < \theta$ and $\theta_U > \theta$, where $\chi^2_{(1)}\left(a\right)$ is the a^{th} quantile of the χ^2 distribution with 1 degrees of freedom. The $100 \times (1-\alpha)\,\%$ ULR CIs can be produced in the same manner for the other parameters α and β.

3.3. Simulation study for point estimates

In the simulation study, 5000 trials are used to estimates the bias and mean squared errors (MSEs) of the MLEs, LSEs, WLSEs, ADEs and CVMEs estimates. Different sample sizes are considered in the study. Two parameter settings are considered. The results are given in the Tables 2-4.

The simulation study is performed based on the following algorithm (for one cycle):

Algorithm 2.

A1. Given true parameters, generate the data from ULW($\boldsymbol{\Xi}$) distribution by using AR sampling given in Algorithm 1.

A2. True parameters are used as initial values in optimization.

A3. The numerical method BFGS is used for the optimization problem given in Eq.'s (3.2), (3.3)-(3.6).

A4. If there is no solution or there is an estimate out of the parameter space, go to **A1**.

From the Tables 2-4, it is observed that the bias and MSEs of the all estimates decrease to zero as expected. The MLEs are best estimates in terms of MSEs. In general, the CVMEs have smaller bias than the others.

3.4. Simulation study for CIs

In the simulation study, 5000 trials are used to predict the CPs of the AN and ULR CIs. The nominal level is fixed at 0.95. In order to get CPs of ULR CIs, there is no need to obtain the CIs limits. It is possible that the CPs of ULR CIs can be simulated by a likelihood ratio test on the true parameter. The simulated CPs of these intervals are given in Table 5. Let us discuss the case $\Xi = (0.5, 1, 1)$. From Table 5, it is observed that the CPs of ULR reach to desired level when the sample of size greater than 100 for all parameters. However, the CPs of AN can not reach the desired level even if a large sample of size is available. In the case of $\Xi = (2.5, 1, 1)$, CIs of AN CIs reach to nominal level when the sample of size greater than 300 for parameters α and β. However, more than 800 sample of size is needed to achive nominal level for parameter θ. The CPs of ULR of CIs reach to nominal level for the all parameters when the sample of size greater than 200.

Under discussion given here, it is indicated that ULR CIs powerful tool to construct the CIs for the ULW parameters.

TABLE 2. Average bias and MSEs of the estimates for the true parameters $\Xi = (0.5, 1, 1)$

		Bias			MSE		
n		θ	α	β	θ	α	β
MLEs	100	-0.1635	-0.3226	-0.1792	0.0419	0.1601	0.0520
	250	-0.1585	-0.3016	-0.1683	0.0384	0.1412	0.0443
	500	-0.1455	-0.2711	-0.1485	0.0319	0.1141	0.0349
	750	-0.1329	-0.2436	-0.1351	0.0263	0.0911	0.0279
	1000	-0.1146	-0.2080	-0.1149	0.0211	0.0711	0.0216
	1250	-0.1106	-0.1987	-0.1098	0.0195	0.0649	0.0197
	1500	-0.0988	-0.1746	-0.0975	0.0156	0.0504	0.0154
	2000	-0.0917	-0.1617	-0.0886	0.0135	0.0433	0.0130
LSEs	100	-0.0776	-0.2175	-0.1277	0.0966	0.2323	0.0741
	250	-0.0834	-0.2039	-0.1151	0.0682	0.1927	0.0611
	500	-0.0474	-0.1452	-0.0771	0.0755	0.1645	0.0511
	750	-0.0630	-0.1528	-0.0830	0.0491	0.1340	0.0421
	1000	-0.0484	-0.1277	-0.0686	0.0471	0.1237	0.0390
	1250	-0.0413	-0.1139	-0.0608	0.0465	0.1158	0.0369
	1500	-0.0372	-0.1009	-0.0541	0.0406	0.1028	0.0329
	2000	-0.0469	-0.1151	-0.0609	0.0382	0.0958	0.0304
WLSEs	100	-0.1061	-0.2618	-0.1515	0.0935	0.2098	0.0703
	250	-0.0927	-0.2271	-0.1304	0.1047	0.1828	0.0563
	500	-0.0892	-0.1994	-0.1092	0.0545	0.1403	0.0440
	750	-0.0981	-0.2003	-0.1115	0.0388	0.1154	0.0362
	1000	-0.0836	-0.1732	-0.0960	0.0354	0.1007	0.0311
	1250	-0.0599	-0.1377	-0.0757	0.0411	0.0992	0.0312
	1500	-0.0751	-0.1503	-0.0839	0.028	0.0802	0.0252
	2000	-0.0798	-0.1561	-0.0855	0.0251	0.0747	0.0230
ADEs	100	-0.1330	-0.3050	-0.1767	0.0998	0.2137	0.0684
	250	-0.1305	-0.2735	-0.1553	0.0595	0.1751	0.0557
	500	-0.0992	-0.2137	-0.1175	0.0527	0.1391	0.0432
	750	-0.1107	-0.2189	-0.1221	0.0368	0.1152	0.0358
	1000	-0.0973	-0.1923	-0.1069	0.0315	0.0991	0.0305
	1250	-0.0819	-0.1668	-0.0924	0.0324	0.0945	0.0294
	1500	-0.0806	-0.1581	-0.0883	0.0269	0.0793	0.0247
	2000	-0.0852	-0.1639	-0.0899	0.0242	0.0743	0.0229
CvMEs	100	-0.0611	-0.1960	-0.1056	0.1125	0.2445	0.0774
	250	-0.0760	-0.1940	-0.1053	0.0731	0.1984	0.0627
	500	-0.0453	-0.1412	-0.0726	0.0746	0.1654	0.0514
	750	-0.0580	-0.1462	-0.0778	0.0516	0.1376	0.0433
	1000	-0.0444	-0.1228	-0.0646	0.0493	0.1269	0.0401
	1250	-0.0385	-0.1101	-0.0577	0.0477	0.1175	0.0375
	1500	-0.0343	-0.0972	-0.0512	0.0420	0.1049	0.0336
	2000	-0.0443	-0.1119	-0.0584	0.0396	0.0977	0.0310

TABLE 3. Average bias and MSEs of the estimates for the true parameters $\Xi = (2.5, 1, 1)$

		Bias			MSE		
n		θ	α	β	θ	α	β
MLEs	100	-0.9446	-0.3368	-0.1242	1.2437	0.1612	0.0389
	250	-0.7282	-0.2465	-0.0928	0.7688	0.0920	0.0207
	500	-0.5585	-0.1843	-0.0640	0.4591	0.0504	0.0090
	750	-0.4759	-0.1525	-0.0524	0.3332	0.0343	0.0055
	1000	-0.4052	-0.1280	-0.0429	0.2417	0.0243	0.0039
	1250	-0.3778	-0.1186	-0.0395	0.2176	0.0217	0.0033
	1500	-0.3417	-0.1068	-0.036	0.1811	0.0177	0.0028
	2000	-0.2916	-0.0899	-0.0292	0.1321	0.0126	0.0018
LSEs	100	-1.1644	-0.4158	-0.2291	1.9877	0.2767	0.0802
	250	-0.5859	-0.2103	-0.1212	1.4339	0.1713	0.0368
	500	-0.5411	-0.1856	-0.0870	0.8559	0.0970	0.0190
	750	-0.3531	-0.1216	-0.0593	0.7588	0.0770	0.0119
	1000	-0.3449	-0.1161	-0.0516	0.5621	0.0570	0.0091
	1250	-0.3094	-0.1019	-0.0458	0.489	0.0491	0.0071
	1500	-0.2766	-0.0903	-0.0406	0.4242	0.0423	0.0058
	2000	-0.2576	-0.0823	-0.0330	0.3002	0.0301	0.0038
WLSEs	100	-1.1623	-0.4183	-0.2125	1.8529	0.2536	0.0736
	250	-0.7235	-0.2451	-0.1111	0.9373	0.1120	0.0266
	500	-0.5608	-0.1853	-0.0743	0.5703	0.0630	0.0121
	750	-0.4401	-0.1420	-0.0547	0.4080	0.0423	0.0072
	1000	-0.4043	-0.1289	-0.048	0.3226	0.0329	0.0056
	1250	-0.3526	-0.1112	-0.0415	0.2766	0.0277	0.0043
	1500	-0.3257	-0.1023	-0.0381	0.2347	0.0234	0.0036
	2000	-0.2845	-0.0881	-0.0307	0.1723	0.0169	0.0023
ADEs	100	-1.1124	-0.3982	-0.1929	1.7112	0.2314	0.0642
	250	-0.7076	-0.2397	-0.1063	0.9011	0.108	0.0250
	500	-0.5514	-0.1823	-0.0721	0.5528	0.0613	0.0117
	750	-0.4369	-0.1410	-0.0538	0.4007	0.0417	0.0070
	1000	-0.4001	-0.1277	-0.0473	0.3188	0.0326	0.0056
	1250	-0.3492	-0.1101	-0.0409	0.2711	0.0273	0.0042
	1500	-0.3202	-0.1006	-0.0374	0.2302	0.0230	0.0036
	2000	-0.2802	-0.0868	-0.0301	0.1698	0.0167	0.0023
CvMEs	100	-1.085	-0.3904	-0.2027	1.8796	0.2588	0.0703
	250	-0.5378	-0.1977	-0.109	1.4186	0.1654	0.0338
	500	-0.5173	-0.1795	-0.0809	0.8367	0.0941	0.0178
	750	-0.3367	-0.1178	-0.0554	0.7528	0.0757	0.0113
	1000	-0.3328	-0.1133	-0.0488	0.5568	0.0562	0.0088
	1250	-0.2999	-0.0998	-0.0436	0.485	0.0485	0.0069
	1500	-0.2686	-0.0885	-0.0387	0.4212	0.0419	0.0056
	2000	-0.2517	-0.0810	-0.0317	0.2978	0.0298	0.0037

TABLE 4. Average bias and MSEs of the estimates for the true parameters $\Xi = (0.9, 1, 0.7)$

		Bias			MSE		
		θ	α	β	θ	α	β
MLEs	100	-0.2621	-0.3343	-0.0937	0.1162	0.1918	0.0203
	250	-0.2273	-0.2897	-0.0830	0.0949	0.1552	0.0155
	500	-0.1898	-0.2437	-0.0718	0.0748	0.1237	0.0122
	750	-0.1538	-0.1953	-0.0570	0.0542	0.0882	0.0084
	1000	-0.1395	-0.1767	-0.0509	0.0478	0.0774	0.0072
	1250	-0.1239	-0.1566	-0.0449	0.0405	0.0650	0.0059
	1500	-0.1104	-0.1393	-0.0396	0.0339	0.0541	0.0048
	2000	-0.0919	-0.1161	-0.0331	0.0270	0.0428	0.0037
LSEs	100	0.2081	0.1441	0.0232	0.3023	0.0469	0.0075
	250	0.1805	0.1407	0.0307	0.3155	0.0373	0.0047
	500	0.3649	0.1114	0.0169	0.6810	0.0269	0.0039
	750	0.2607	0.1226	0.0237	0.4100	0.0287	0.0039
	1000	0.2486	0.1264	0.0287	0.4820	0.0307	0.0048
	1250	0.4196	0.0962	0.0125	0.7386	0.0179	0.0030
	1500	0.4022	0.0967	0.0133	0.6763	0.0175	0.0030
	2000	0.3701	0.0963	0.0134	0.5429	0.0159	0.0024
WLSEs	100	0.3516	0.1031	0.0115	0.9506	0.0266	0.0062
	250	0.3277	0.1029	0.0160	0.6265	0.0201	0.0032
	500	0.3996	0.0880	0.0075	0.5159	0.0153	0.0024
	750	0.3087	0.0993	0.0140	0.3184	0.0167	0.0023
	1000	0.3954	0.0865	0.0083	0.4673	0.0131	0.0019
	1250	0.4015	0.0847	0.0075	0.4527	0.0122	0.0017
	1500	0.3509	0.0910	0.0114	0.3566	0.0138	0.0019
	2000	0.3584	0.0869	0.0091	0.3458	0.0114	0.0013
ADEs	100	0.4130	0.0916	0.0039	0.5834	0.0228	0.0055
	250	0.3729	0.0943	0.0105	0.5367	0.0174	0.0030
	500	0.4344	0.0820	0.0037	0.5408	0.0137	0.0022
	750	0.3527	0.0917	0.0094	0.3573	0.0145	0.0020
	1000	0.3881	0.0872	0.0083	0.4336	0.0136	0.0020
	1250	0.4220	0.0814	0.0052	0.4631	0.0115	0.0017
	1500	0.4142	0.0809	0.0054	0.4339	0.0110	0.0015
	2000	0.3807	0.0833	0.0068	0.3633	0.0106	0.0013
CVMEs	100	0.1796	0.1481	0.0387	0.3267	0.0494	0.0089
	250	0.1612	0.1442	0.0380	0.3270	0.0392	0.0054
	500	0.3530	0.1139	0.0210	0.6981	0.0281	0.0042
	750	0.2458	0.1253	0.0271	0.4097	0.0298	0.0042
	1000	0.2445	0.1276	0.0306	0.4962	0.0312	0.0049
	1250	0.4130	0.0978	0.0145	0.7506	0.0184	0.0031
	1500	0.3955	0.0982	0.0151	0.6842	0.0181	0.0031
	2000	0.3627	0.0977	0.0148	0.5434	0.0163	0.0025

Table 5. The CPs of AN and ULR CIs

		AN			ULR		
Ξ	n	α	β	θ	α	β	θ
$(0.5, 1, 1)$	100	0.8772	0.8136	0.8790	0.9566	0.9564	0.9558
	200	0.8254	0.8310	0.8830	0.9470	0.9476	0.9480
	300	0.8130	0.8210	0.8750	0.9380	0.9390	0.9382
	400	0.8268	0.8372	0.8862	0.9408	0.9402	0.9406
	500	0.8370	0.8462	0.8812	0.9434	0.9412	0.9456
	600	0.8420	0.8500	0.8834	0.9400	0.9402	0.9394
	700	0.8548	0.8612	0.8806	0.9412	0.9422	0.9424
	800	0.8450	0.8526	0.8746	0.9382	0.9362	0.9364
	900	0.8604	0.8664	0.8800	0.9414	0.9426	0.9430
	1000	0.8672	0.8712	0.8830	0.9352	0.9358	0.9344
$(2.5, 1, 1)$	100	0.9176	0.9164	0.8858	0.9602	0.9534	0.9630
	200	0.9458	0.9386	0.9156	0.9500	0.9496	0.9500
	300	0.9520	0.9450	0.9294	0.9516	0.9492	0.9492
	400	0.9584	0.9548	0.9342	0.9508	0.9536	0.9528
	500	0.9546	0.9552	0.9328	0.9520	0.9552	0.9506
	600	0.9604	0.9562	0.9410	0.9562	0.9544	0.9562
	700	0.9586	0.9602	0.9378	0.9552	0.9540	0.9538
	800	0.9626	0.9542	0.9444	0.9524	0.9522	0.9524
	900	0.9618	0.9570	0.9470	0.9520	0.9534	0.9540
	1000	0.9608	0.9588	0.9440	0.9544	0.9552	0.9532

4. ULW regression analysis

The regression models are used in different ways in survival analysis. Sometimes mean or quantiles of underlying distribution are assumed as a linear function of covariates(predictors). When the mean or quantiles have not explicit form, the location parameter is assumed as a linear function of covariates by using a suitable link function. The log-location-scale regression models are studied by several authors such as [1] and [24]. In this section, we describe the use of log-location-scale ULW regression methodology.

Let X be a ULW($\boldsymbol{\Xi}$) random variable. Let us consider are-parameterization by $\beta = 1/\sigma$ and $\alpha = \exp(\mu)$ and then, the log-lifetime $Y = \log(X)$ is a random variable with the pdf

$$h(y;\boldsymbol{\tau}) = \frac{\theta^2 \exp\left(\frac{y-\mu}{\sigma}\right) \exp\left\{-\theta \exp\left(\exp\left(\frac{y-\mu}{\sigma}\right)\right) + \theta + 2\exp\left(\frac{y-\mu}{\sigma}\right)\right\}}{(1+\theta)\,\sigma} \mathbb{I}_{\mathbb{R}}(y),$$

where $\boldsymbol{\tau} = (\mu, \sigma, \theta)$, μ and σ are location and scale parameters, respectively. The cdf of Y is also given by

$$H(y;\boldsymbol{\tau}) = 1 - \left(1 + \frac{\theta\left(1-\exp\left(-\exp\left(\frac{y-\mu}{\sigma}\right)\right)\right)}{(1+\theta)\exp\left(-\exp\left(\frac{y-\mu}{\sigma}\right)\right)}\right) \exp\left(-\frac{\left(1-\exp\left(-\exp\left(\frac{y-\mu}{\sigma}\right)\right)\right)}{\exp\left(-\exp\left(\frac{y-\mu}{\sigma}\right)\right)}\right). \tag{4.1}$$

It is noted that the random variable Y with cdf (4.1) is denoted LULW(μ, σ, θ), where LULW stands for log-ULW distribution. Let us consider the regression model

$$\mathbf{Y} = \boldsymbol{\mu} + \sigma \boldsymbol{\varepsilon}, \tag{4.2}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ and $Y_1, Y_2, \ldots, Y_n$ are independent LULW random variables with parameters $\left(\mu_i = \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}, \sigma, \theta\right)$, respectively. Furthermore, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, $\mu_i = \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}$ and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip})^{\mathrm{T}}$ $(= (1, Z_{i1}, \ldots, Z_{ip})^{\mathrm{T}}$ when a intercept is included in a model) are ith values of covariates for $i = 1, 2, \ldots, n$. In addition, $\varepsilon_i = (Y_i - \mu_i)/\sigma$ for $i = 1, 2, \ldots, n$ is a random error distributed LULW with parameters ($\mu = 0, \sigma = 1, \theta$).

Let us discuss the MLEs of parameters $\boldsymbol{\eta} = (\boldsymbol{\beta}, \sigma, \theta)$ in the model (4.2) under Type-I right censoring. Suppose that the log-lifetimes $Y_i\,(i = 1, 2, \ldots, n)$ are Type-I right censored (at $\log(c_i)$) from LULW(μ_i, σ, θ), where c_i is censoring time for lifetime X_i. Let us define

$$T_i = \min\{Y_i, \log(c_i)\},\ i = 1, 2, \ldots, n.$$

Hence, the log-likelihood function based on the Type-I right censored sample $T_1, T_2, \ldots, T_n$ is written by

$$\ell(\boldsymbol{\eta}) = \sum_{i=1}^{n} \left\{\omega_i \log\left(h\left(t_i; \left(\boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}, \sigma, \theta\right)\right)\right) + (1-\omega_i) \log\left(1 - H\left(t_i; \left(\boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}, \sigma, \theta\right)\right)\right)\right\}, \tag{4.3}$$

where

$$\omega_i = \begin{cases} 0\ ,\ T_i > \log(c_i) \\ 1\ ,\ T_i \leq \log(c_i) \end{cases}$$

is an indicator function and t_i denotes the observed value of T_i, $i = 1, 2, \ldots, n$.

The MLE of $\boldsymbol{\eta}$ can be obtained by maximizing the log-likelihood (4.3). Some numerical methods such as Nelder-Mead and BFGS can be used for a maximization problem.

4.1. Simulation study for MLEs of regression parameters

In this subsection, the bias and MSEs of MLEs are discussed for lifetime regression model parameters through a Monte Carlo simulation with 2000 trials. All simulations are run for the following model

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \beta_3 Z_{i3} + \sigma\varepsilon_i \\ &= \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\beta} + \sigma\varepsilon_i,\ i = 1, 2, \ldots, n \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$, $\mathbf{Z}_i^{\mathrm{T}} = (1, Z_{i1}, Z_{i2}, Z_{i3})$ and $\varepsilon_i \sim$LUWL$(\mu = 0, \sigma = 1, \theta)$, $i = 1, 2, \ldots, n$.

In the simulation, we consider $\boldsymbol{\beta} = (-.1, -.1, -.1, -.1)$, $\theta = 1, 1.5$ and 2. The true parameter $\boldsymbol{\beta} = (.1, .1, .1, .1)$ is also considered in the simulation, but no different patterns are observed for the other one. The covariates $\boldsymbol{Z}_i$, $(i = 1, 2, \ldots, n)$ are generated in two cases: In the first case, four levels (there are 4 categories: 1,2,3,4) are considered for Z_{i1}, Z_{i2} and Z_{i3}. The other case, (Z_{i1}, Z_{i2}, Z_{i3}) are generated from multivariate normal distribution. In addition, two correlation matrix for covariates (Z_{i1}, Z_{i2}, Z_{i3}) are considered by

$$\boldsymbol{\rho}_1 = \begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix} \text{ and } \boldsymbol{\rho}_2 = \begin{pmatrix} 1 & 0.5 & 0.5 \\ & 1 & 0.5 \\ & & 1 \end{pmatrix}.$$

Hence, it can be observed the multicollinearity effect on the URL regression analysis. The simulation study is performed based on the following algorithm for one cycle:

Algorithm 3

A1. For a fixed n, generate the covariates with a given correlation matrix ($\boldsymbol{\rho}_1$ or $\boldsymbol{\rho}_2$) and equal marginal probabilities of four levels. R function **ordsample** in the package **GenOrd** is used in our study. Otherwise, the covariates are generated from a multivariate normal distribution with mean **0** and given correlation matrix (ρ_1 or ρ_2). R function **mvrnorm** in the package **MASS** is used in our study**.**

A2. Compute $\mu_i = \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\beta}$, $i = 1, 2, \ldots, n$.

A3. Set the true parameters $\alpha_i = \exp(\mu_i)$, β and θ.

A4. For $i = 1, 2, \ldots, n$, generate the dependent variable X_i from ULW$(\boldsymbol{\Xi}_i)$ distribution with $\boldsymbol{\Xi}_i = (\alpha_i, \beta, \theta)$ using the AR sampling given in Algorithm 1 and set $Y_i = \log(X_i) \sim$LULW(μ_i, σ, θ).

A5. The numerical methods such as Nelder-Mead, BFGS and CG are used to maximize the log-likelihood given in Eq. (4.3) and the true parameters given **A3** are used as initial values.

A6. If there is no solution or estimate out of parameter space, or negative standard error, go to **A4**.

Using Algorithm 3, a simulation study is performed with 2000 trials for a sample of size $n = 100, 200, \ldots, 1000$ and the nominal level is fixed at 0.95. Figures 3-6 are produced by settings given at the beginning of this subsection.

From Figures 3-6, the discrete or continuous covariates discussed above, does not affect the properties of estimates. If the multicollinearity level increase, the MSEs of $\widehat{\beta}_1, \widehat{\beta}_2$ and $\widehat{\beta}_3$ increase. It is an interesting observation from Figures 3-4 that, MSEs of $\widehat{\beta}_0, \widehat{\sigma}$ and $\widehat{\theta}$ are not affected by the degree of multicollinearity within covariates $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3$. Although the $\widehat{\beta}_1, \widehat{\beta}_2$ and $\widehat{\beta}_3$ has a negligible bias for even if a small sample of size, the estimates $\widehat{\beta}_0, \widehat{\sigma}$ and $\widehat{\theta}$ are asymptotically unbiased. The CPs of AN CIs for β_1, β_2 and β_3 are almost equal to the nominal level for all sample size and multicollinearity cases. Furthermore, the CPs of AN CIs for θ are greater than nominal level for small sample size but it reduces to the nominal level when the sample size increases. The CPs of AN CIs for β_0 and σ are less than nominal level for small sample size, but it climbs to the nominal level when sample size increases. The mean lengths of CIs for all parameters decrease to zero when the sample size increases. The mean lengths of AN CIs for $\beta_1, \beta_2, \beta_3$ in the case multicollinearity are wider than being uncorrelated covariates. Being multicollinearity does not affect negatively on the mean lengths of AN CIs for β_0, σ and θ.

In Figures 5-6, the behaviors of estimates and CIs are also discussed according to increment in true parameter θ. When the true parameter θ is 1, bias of $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\sigma}$ and $\widehat{\theta}$ are negligible for moderate sample size. If $\theta < 1\,(> 1)$ the bias of $\widehat{\beta}_0$ are positive (negative), but it reduces (increases) to zero when the sample size increases. If $\theta < 1\,(> 1)$ the bias of $\widehat{\sigma}$ are negative (positive) but it increases (reduce) to zero when the sample size increases. When the θ increases, MSEs of $\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3$ and $\widehat{\theta}$ increase. For small sample size, if the θ increases, the MSEs of $\widehat{\beta}_0$ and $\widehat{\sigma}$ increase. For large sample size, if the θ increases, the MSEs of $\widehat{\beta}_0$ and $\widehat{\sigma}$ decrease. The CPs of AN CIs for β_1, β_2 and β_3 are almost equal to nominal level for $\theta = 0.5, 1$ and 2. The CPs of AN CIs for β_0 are less (greater) than nominal level when $\theta < 1\,(> 1)$. If $\theta = 1$, the CPs of AN CIs for β_0 tends to nominal level for $n \geq 300$. When the θ increases, CPs of AN CIs of σ are closing to nominal level, but mean lengths of AN CIs of $\beta_1, \beta_2, \beta_3$, σ and θ increase.

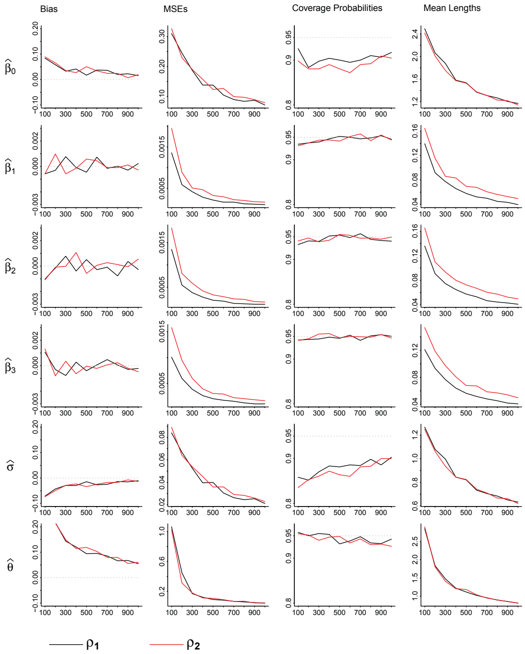


FIGURE 3. Average bias and MSEs for MLEs, CPs and mean lengths for AN CIs of ULW regression model parameters when multivariate normal covariates and $\theta = 1$

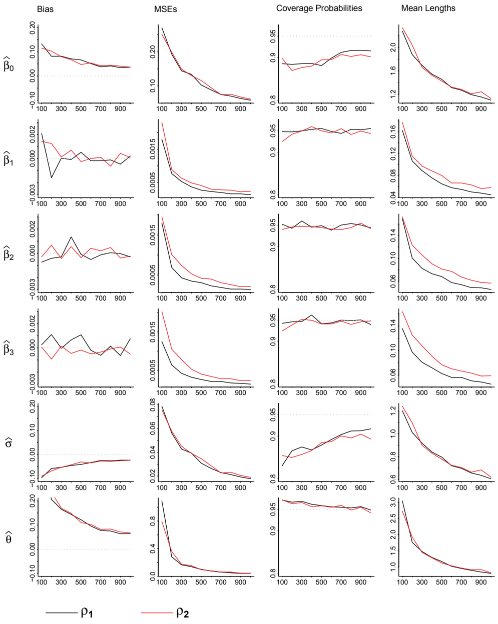


FIGURE 4. Average bias and MSEs for MLEs, CPs and mean lengths for AN CIs of ULW regression model parameters when ordinal covariates and $\theta = 1$

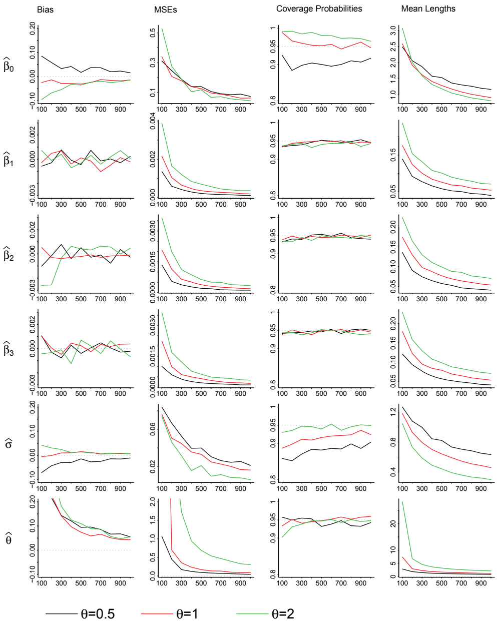


FIGURE 5. Average bias and MSEs for MLEs, CPs and mean lengths for AN CIs of ULW regression model parameters when multivariate normal covariates with correlation matrix $\boldsymbol{\rho}_1$

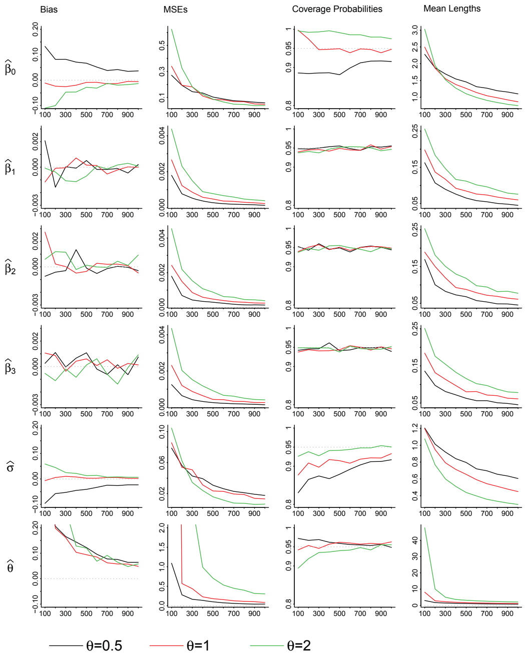


FIGURE 6. Avarage bias and MSEs for MLEs, CPs and mean lengths for AN CIs of ULW regression model parameters when ordinal covariates with correlation matrix $\boldsymbol{\rho}_1$

5. Real Data Analysis

In this section, the real data application of ULW distribution is given. The distribution fitting to total milk production data is studied.

The ULW distribution is now fitted to the data about the total milk production in the first birth of 107 cows from SINDI race. The data is taken from [3] and the data given as follow:

0.4365, 0.4260, 0.5140, 0.6907, 0.7471, 0.2605, 0.6196, 0.8781, 0.4990, 0.6058 ,0.6891, 0.5770, 0.5394, 0.1479, 0.2356, 0.6012, 0.1525, 0.5483, 0.6927, 0.7261, 0.3323, 0.0671, 0.2361, 0.4800, 0.5707, 0.7131, 0.5853, 0.6768 ,0.5350, 0.4151 ,0.6789, 0.4576, 0.3259, 0.2303, 0.7687, 0.4371 ,0.3383, 0.6114, 0.3480, 0.4564, 0.7804, 0.3406, 0.4823, 0.5912 ,0.5744, 0.5481, 0.1131, 0.7290, 0.0168, 0.5529 ,0.4530, 0.3891, 0.4752, 0.3134, 0.3175 ,0.1167, 0.6750, 0.5113,0.5447, 0.4143, 0.5627, 0.5150, 0.0776, 0.3945 ,0.4553, 0.4470, 0.5285, 0.5232, 0.6465, 0.0650, 0.8492, 0.8147, 0.3627, 0.3906, 0.4438, 0.4612, 0.3188, 0.2160, 0.6707, 0.6220, 0.5629, 0.4675, 0.6844, 0.3413,0.4332, 0.0854, 0.3821, 0.4694, 0.3635, 0.4111, 0.5349, 0.3751, 0.1546, 0.4517 ,0.2681, 0.4049, 0.5553, 0.5878 ,0.4741 ,0.3598, 0.7629, 0.5941, 0.6174, 0.6860, 0.0609, 0.6488, 0.2747.

It should be pointed out that this data is also analyzed in [6] and [20]. For the comparison, beta, Weibull (W), the Lindley Weibull (LW), unit-gamma (UG), unit-logistic (ULOG), UL distributions are considered. It is noted that LW, UG, ULOG and UL are introduced by [4], [11], [14], and [22] respectively. The pdfs of these distributions are given by

$$\begin{aligned}
f_{ULW}\left(x\right) &= 1-\left\{1+\frac{p_1\left(1-\exp\left(-\left(\frac{x}{p_3}\right)^{p_2}\right)\right)}{\left(1+p_1\right)\exp\left(-\left(\frac{x}{p_3}\right)^{p_2}\right)}\right\}\exp\left\{-\frac{p_1\left(1-\exp\left(-\left(\frac{x}{p_3}\right)^{p_2}\right)\right)}{\exp\left(-\left(\frac{x}{p_3}\right)^{p_2}\right)}\right\}\mathbb{I}_{\mathbb{R}_+}\left(x\right)\\
f_W\left(x\right) &= \frac{p_1}{p_2}-\left(\frac{x}{p_1}\right)^{p_2-1}\exp\left(-\left(\frac{x}{p_1}\right)^{p_2}\right)\mathbb{I}_{\mathbb{R}_+}\left(x\right)\\
f_{LW}\left(x\right) &= \frac{x^{p_2-1}p_3^2p_2p_1^2\exp\left(-\left(xp_1\right)^{p_2}\right)}{1+p_3}\\
&\quad\times\left(1-\log\left(\exp\left(-\left(xp_1\right)^{p_2}\right)\right)\right)\left(\exp\left(-\left(xp_1\right)^{p_2}\right)\right)^{p_3-1}\mathbb{I}_{\mathbb{R}_+}\left(x\right)\\
f_{Beta}\left(x\right) &= \frac{1}{\beta\left(p_1,p_2\right)}x^{p_1-1}\left(1-x\right)^{p_2-1}\mathbb{I}_{(0,1)}\left(x\right)\\
f_{UG}\left(x\right) &= \frac{p_2^{p_1}x^{p_2-1}\left(-\log\left(x\right)\right)^{p_1-1}}{\Gamma\left(p_1\right)}\mathbb{I}_{(0,1)}\left(x\right)\\
f_{ULOG}\left(x\right) &= \frac{p_2\exp\left(p_1\right)x^{p_2-1}\left(1-x\right)^{p_2-1}}{\left(x^{p_2}\exp\left(p_1\right)+\left(1-x\right)^{p_2}\right)^2}\mathbb{I}_{(0,1)}\left(x\right)\\
f_{UL}\left(x\right) &= \frac{p_1^2\exp\left(-\frac{xp_1}{1-x}\right)}{\left(1+p_1\right)\left(1-x\right)^3}\mathbb{I}_{(0,1)}\left(x\right)
\end{aligned}$$

The total time on test (TTT) plot is used to determine the hazard behavior of the data. TTT plot for the total milk production data is given in Figure 7 and it indicates that the total milk production comes from a distribution with the increasing failure rate. Therefore, the ULW distribution is a candidate for modeling this data (see, Section 2.1 and Figure 2).

In this section, seven distributions are fitted to the total milk production data with the likelihood principle. The MLEs of distribution parameters are obtained by numerical methods that try to maximize the log-likelihood. In most cases, we observe that the different initial values give different estimates, and one can not conclude which one is treated as a MLE. Therefore, an algorithm is used to get the almost correct MLEs of parameters given in Table 6. An algorithm is given as follows:

Algorithm 4.

A1. 1000 (it can be increased by optionally) initial values are uniformly generated from a subset of parameter space.

A2. Using initial values generated in Step **A1**, the numerical methods Nelder-Mead, BFGS, and CG are used to maximize the log-likelihood.

A3. The likelihoods for all estimates in Step **A2** are ordered from large to small.

A4. The estimates with the largest likelihoods are treated as MLEs of parameters.

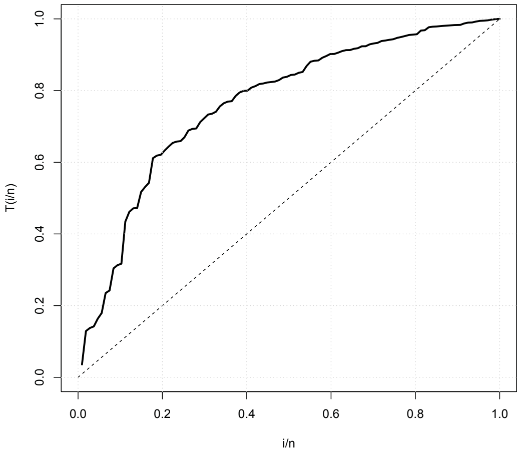


FIGURE 7. TTT plot for the total milk production data

The MLEs of parameters and related standard errors for ULW, W, LW, beta, UG, ULOG and UL distributions are given in Table 6. In this table, some comparison criteria are presented. The log-likelihood ℓ, -2ℓ, AIC, Bayesian information criterion (BIC), corrected Akaike' s information criterion (CAIC), Hannan–Quinn information criterion (HQIC), Kolmogorov-Smirnov statistic (KS), Anderson-Darling statistic(AD), Cramér von Mises statistic(CvM) and related p-values(KS p-value, AD p-value and CvM p-value), the MLE $\widehat{p_i}$ $(i=1,2,3)$ of parameter p_i with standard error $se\left(\widehat{p_i}\right)$ and AN intervals (LB_{p_i}, UB_{p_i}) are calculated and they are presented in Table 6. It is noted that some lower limit of AN CI are below the lower bound of parameter space. It can be corrected with lower bound of parameter space. In the Table 6, initial parameters, and the numerical methods are given to get MLEs for all models in the analysis. From the Table 6, the ULW distribution has the smallest values of -2ℓ, AIC, BIC, CAIC, HQIC, KS, AD and CvM. Furthermore, goodness of fit tests KS, AD and CvM confirm the ULW model validity (p values>0.05). From these results, it is concluded that the ULW distribution is better than the others in terms of all criteria. Figure 9 presents the overlapping of the fitted ULW cdf on the empirical cdf. From Figure 9, it is observed that fitted cdf of ULW distribution exhibits better than the others.

Using discussion in Subsection 3.2, 95% ULR CIs for θ, α and β are calculated by (0.0560,1.7914), (0.1515,0.7142) and (0.5473,2.1491), respectively. Figure 8 represents the 95% ULR CI of parameter θ. A logarithmic scale is used for x-axis to improve the quality of graphical view.

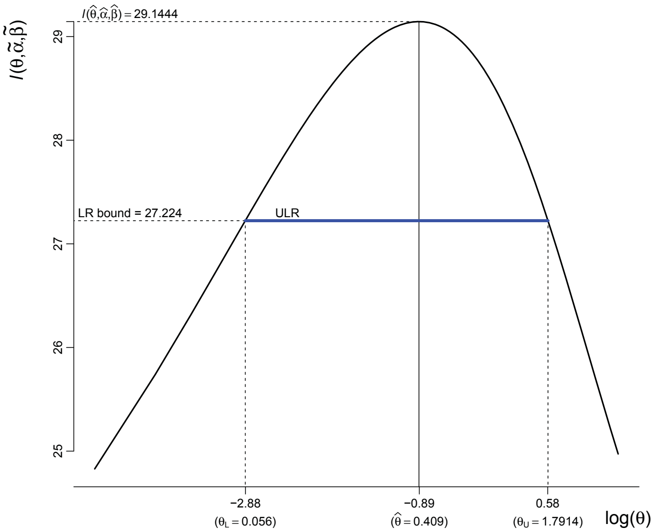


FIGURE 8. Confidence limits for parameter θ based on ULR

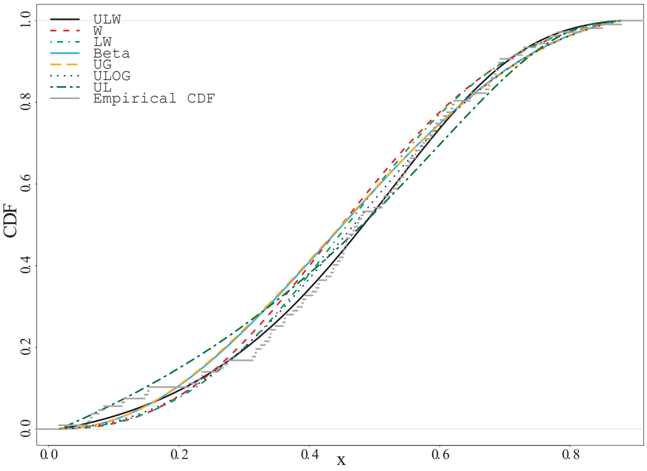


FIGURE 9. Fitted and empirical cdf plots for the total milk production data

TABLE 6. Data analysis results for the total milk production data

	ULW	W	LW	Beta	UG	ULOG	UL
ℓ	29.1444	21.3475	23.6708	23.7772	23.0467	24.8400	25.3805
-2ℓ	-58.2888	-42.6950	-47.3417	-47.5545	-46.0934	-49.6800	-50.7609
AIC	-52.2888	-38.6950	-41.3417	-43.5545	-42.0934	-45.6800	-48.7609
BIC	-44.2703	-33.3494	-33.3232	-38.2088	-36.7477	-40.3343	-46.0881
CAIC	-52.0557	-38.5796	-41.1087	-43.4391	-41.9780	-45.5646	-48.7229
HQIC	-49.0382	-36.5280	-38.0911	-41.3874	-39.9263	-43.5129	-47.6774
KS	0.0459	0.0832	0.0653	0.0910	0.0939	0.0571	0.1096
AD	0.2332	1.4841	1.0403	1.3853	1.4997	0.8646	1.3116
CVM	0.0292	0.1895	0.1104	0.2282	0.2450	0.0771	0.2286
KS p value	0.9778	0.4487	0.7518	0.3384	0.3021	0.8767	0.1532
AD p value	0.9785	0.1804	0.3366	0.2064	0.1766	0.4364	0.2286
CVM p value	0.9792	0.2891	0.5371	0.2190	0.1949	0.7095	0.2184
$\widehat{p}_1$	0.4091	0.5236	3.0722	2.4125	2.6767	0.2073	1.2001
$\widehat{p}_2$	1.1486	2.6012	2.2558	2.8297	2.9774	1.9104	
$\widehat{p}_3$	0.3454		0.5823				
LB$_{p_1}$	-0.2404	0.4839	0.8143	1.7961	1.9994	-0.1166	1.0258
LB$_{p_2}$	0.3768	2.1899	1.7933	2.0958	2.1489	1.6022	
LB$_{p_3}$	-0.0211		-0.1751				
UB$_{p_1}$	1.0586	0.5633	5.3301	3.0289	3.3541	0.5311	1.3743
UB$_{p_2}$	1.9205	3.0124	2.7182	3.5635	3.8060	2.2185	
UB$_{p_3}$	0.7119		1.3397				
SE$_{\widehat{p}_1}$	0.3314	0.0202	1.1520	0.3145	0.3456	0.1652	0.0889
SE$_{\widehat{p}_2}$	0.3938	0.2098	0.2359	0.3744	0.4228	0.1572	
SE$_{\widehat{p}_3}$	0.1870		0.3864				
Numerical Method	BFGS	BFGS	BFGS	CG	CG	BFGS	CG
Inital value for $\widehat{p}_1$	46.5515	91.4019	16.6926	10.4744	91.4145	61.8622	45.8126
Inital value for $\widehat{p}_2$	55.8020	24.2207	13.9139	16.9355	95.3848	47.9454	
Inital value for $\widehat{p}_3$	4.5296		22.7243				

6. Conclusions

In this paper, a new lifetime regression analysis with a newly introduced distribution is provided. The simulation study given in Subsection 4.1 indicates that proposed regression analysis can be used without any doubt.

Acknowledgement

The authors thank to Professor Necip Doğanaksoy and Professor Hassan Bakouch for constructive comments on our manuscript.

References

[1] Altun, E., Yousof, H.M. and Hamedani, G.G. (2018). A new log-location regression model with influence diagnostics and residual analysis. *International Journal of Applied Mathematics and Statistics*, 33(3), 417-449.

[2] Alzaatreh, A., Lee, C. and Famoye, F. (2013). A new method for generating families of continuous distributions. *Metron*, 71, 63-79.

[3] Brito, R.S. (2009). *Estudo de expansoes assintoticas, avaliacao numerica de momentos das distribuicoes beta generalizadas, aplicacoes em modelos de regressao e analise discriminante.* [Master's thesis, Universidade Federal Rural de Pernambuco].

[4] Cordeiro, G.M., Afify, A.Z., Yousof, H.M., Çakmakyapan, S. and Özel, G. (2018). The Lindley Weibull distribution: properties and applications. *Anais da Academia Brasileira de Ciências*, 90, 2579-2598.

[5] Cordeiro, G.M. and de Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, 81, 883-898.

[6] Cordeiro, G.M. and dos Santos Brito, R. (2012). The beta power distribution. *Brazilian Journal of Probability and Statistics*, 26, 88-112.

[7] Cordeiro, G.M., Ortega, E.M. and Nadarajah, S. (2010). The Kumaraswamy Weibull distribution with application to failure data. *Journal of the Franklin Institute*, 347, 1399-1429.

[8] Eugene, N., Lee, C. and Famoye, F. (2002). Beta-normal distribution and its applications. *Communications in Statistics-Theory and Methods*, 31, 497-512.

[9] Fraser, D.A.S. (1976). *Probability and Statistics: Theory and Applications.* Duxbury Press, North Scituate, Mass.

[10] Gradshteyn, I.S. and Ryzhik, I.M. (2014). *Table of integrals, Series, and Products.* Eighth Edition, Academic Press.

[11] Grassia, A. (1977). On a family of distributions with argument between 0 and 1 obtained by transformation of the gamma and derived compound distributions. *Australian Journal of Statistics*, 19, 108-114.

[12] Korkmaz, M.C. (2020). A new heavy-tailed distribution defined on the bounded interval: the logit slash distribution and its application. *Journal of Applied Statistics*, 47(12), 2097-2119.

[13] Korkmaz, M.C., Altun, E., Yousof, H.M. and Hamedani, G.G. (2019). The odd power Lindley generator of probability distributions: properties, characterizations and regression modeling. *International Journal of Statistics and Probability*, 8, 70-89.

[14] Mazucheli, J., Menezes, A.F.B. and Chakraborty, S. (2019). On the one parameter unit-Lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics*, 46, 700-714.

[15] Nadarajah, S. and Gupta, A.K. (2004). The beta Fréchet distribution. *Far East Journal of Theoretical Statistics*, 14, 15-24.

[16] Nadarajah, S. and Kotz, S. (2004). The beta Gumbel distribution. *Mathematical Problems in Engineering*, 4, 323-332.

[17] Nadarajah, S. and Kotz, S. (2006). The beta exponential distribution. *Reliability Engineering and System Safety*, 91, 689-697.

[18] Patil, G.P. and Rao, C.R. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics*, 34, 179-189.

[19] Pascoa, M.A., Ortega, E.M. and Cordeiro, G.M. (2011). The Kumaraswamy generalized gamma distribution with application in survival analysis. *Statistical Methodology*, 8, 411-433.

[20] Saraçoğlu, B. and Tanış, C. (2018). A new statistical distribution: cubic rank transmuted Kumaraswamy distribution and its properties. *Journal of the National Science Foundation of Sri Lanka*, 46, 505-518.

[21] Shaked, M. and Shanthikumar, J.G. (1994). *Stochastic Orders and Their Applications*. Academic Press, London.

[22] Tadikamalla, P.R., Johnson, N. L. (1982). Systems of frequency curves generated by transformations of logistic variables. *Biometrika*, 69, 461-465.

[23] Tanış, C. and Saraçoğlu, B. (2019). Comparisons of six different estimation methods for log-Kumaraswamy distribution. *Thermal Science*, 23, 344-344.

[24] Yousof, H.M., Altun, E., Rasekhi, M., Alizadeh, M., Hamedan, G.G. and Ali, M.M. (2019). A new lifetime model with regression models, characterizations and applications. *Communications in Statistics - Simulation and Computation*, 48, 264-286.

Appendix P*roof of Theorem 1*

For any $x > 0$, the ratio of the densities is given by

$$g\left(x\right)=\frac{\theta_1^2\left(1+\theta_2\right)\exp\left(-\theta_1\exp\left(\left(\frac{x}{\alpha}\right)^{\beta}\right)+\theta_1+2\left(\frac{x}{\alpha}\right)^{\beta}\right)}{\theta_2^2\left(1+\theta_1\right)\exp\left(-\theta_2\exp\left(\left(\frac{x}{\alpha}\right)^{\beta}\right)+\theta_2+2\left(\frac{x}{\alpha}\right)^{\beta}\right)}.$$

Consider the derivative of $\log\left(g\left(x\right)\right)$ in x

$$\frac{d\log\left(g\left(x\right)\right)}{dx}=\frac{\left(\theta_2-\theta_1\right)\beta\left(\frac{x}{\alpha}\right)^{\beta}\exp\left(\left(\frac{x}{\alpha}\right)^{\beta}\right)}{x}<0$$

for $\theta_1 > \theta_2$ and hence proof is completed.

İSTATİSTİK: JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION

Vol. 14, No. 2, December 2022, pp. 74–86

ISSN 1300-4077 | 22 | 2 | 74 | 86

GENERALIZED FIDUCIAL INFERENCE FOR THE CHEN DISTRIBUTION

Çağatay Çetinkaya*

Department of Accounting and Taxation,
Bingöl University,
12010, Bingöl, Turkey

Abstract: The fiducial inference idea was firstly proposed by Fisher [8] as a powerful method in statistical inference. Many authors such as Weeranhandi [24] and Hannig et. al. [12] improved this method from different points of view. Since the Bayesian method has some deficiencies such as assuming a prior distribution when there was little or no information about the parameters, the fiducial inference is used to overcome these adversities. This study deals with the generalized fiducial inference for the shape parameters of the Chen' s two-parameter lifetime distribution with bathtub shape or increasing failure rate [4]. The method based on the inverse of the structural equation which is proposed by Hannig et. al. [12] is used. We propose the generalized fiducial inferences of the parameters with their confidence intervals. Then, these estimations are compared with their maximum likelihood and Bayesian estimations. Simulation results show that the generalized fiducial inference is more applicable than the other methods in terms of the performances of estimators for the shape parameters of the Chen distribution. Finally, a real data example is used to illustrate the theoretical outcomes of these estimation procedures.

Key words: Bayesian inference, Generalized fiducial inference, Interval estimation, Chen distribution, Point estimation.

1. Introduction

The fiducial idea is firstly proposed by Fisher [8] as a powerful method in statistics. It is known that assuming a prior distribution in the case of insufficient information about the parameters causes adversities in Bayesian inference. The main idea of Fisher [8] with the fiducial method was to overcome this deficiency in the Bayesian framework. Then, some deficiencies in the fiducial inference and the philosophical concerns regarding the interpretation of fiducial probability were handled by various authors (See Zabell [28] for more details.). Thus, the idea of fiducial inference was improved by various authors. Recently, Hannig [11, 10] handled generalized fiducial inference. Then, Hannig et al. [12] defined the generalized fiducial inference method based on the inverse of the structural equation. The generalized fiducial inference is actually similar to the likelihood approach. It differs from the likelihood method by switching the role of the parameters and the observed data.

In statistics theory, there are several applications of fiducial inference. For example; Wandler and Hannig [21, 22] considered generalized fiducial inference on the largest mean of a multivariate normal distribution and also inference on the parameters and the extreme quantiles of the generalized Pareto distribution, respectively. Wang et al. [23] handled fiducial inference to construct prediction intervals for an arbitrary probability distribution. Further; O'Reilly and Rueda [17] studied the truncated exponential distribution, Li and Xu [15] studied inference of Birnbaum-Saunders distribution, Yan and Liu [27] studied generalized exponential distribution with the fiducial inference method.

* Corresponding author. E-mail address:ccetinkaya@bingol.edu.tr

On the other hand, Chen [4] proposed a two-parameter distribution with bathtub shaped or increasing hazard function. Chen [4] proposed using this model to analyze the lifetime datasets flexibly. It has the following probability density (pdf), distribution (cdf) and failure rate functions

$$f(x;\lambda,\beta)=\lambda\beta x^{\beta-1}e^{x^{\beta}}e^{\lambda(1-e^{x^{\beta}})},\ \ x>0,\ \ \lambda,\beta>0,$$

$$F(x;\lambda,\beta)=1-e^{\lambda(1-e^{x^{\beta}})},$$

and

$$h(x;\lambda,\beta)=\lambda\beta x^{\beta-1}e^{x^{\beta}}.$$

The Chen distribution has a bathtub shape failure rate when $\beta<1$ and also has an increasing failure rate function when $\beta\geq 1$ (see Figure 1).

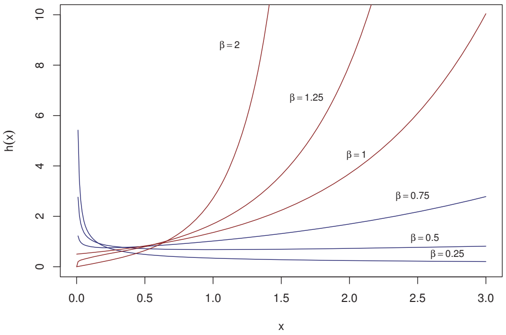


FIGURE 1. Failure rate functions when $\lambda=0.5$

Hence it provides an appropriate conceptual model for some electronic and mechanical products as well as the lifetime of humans. In addition to its positively skewed shape, it also has some other flexible properties (Chen, [4]). Further;

- It leads to the exponential power distribution when $\lambda=1$.
- If $X\sim CH(\lambda,\beta)$, then $Y=(e^{X^{\beta}}-1)\sim Exp(\lambda)$ and $Y=(e^{X^{\beta}}-1)^{\frac{1}{\theta}}\sim Weibull(\lambda,\theta)$.
- It leads to the $Gompertz(1,\lambda)$ distribution when $\beta=1$.

Many authors handled the Chen distribution in terms of statistical inference. For instance; Wu et al. [25] studied the estimation of its shape parameter. Then, Wu [26] studied its parameter estimations under progressive censoring. Sarhan et al. [19] obtained estimations of its parameters. Rastogi and Tripathi [18] handled parameter estimations under hybrid censored data. Ahmed [2] obtained Bayesian estimations and compared them with non-Bayesian estimations under progressive Type-II censoring scheme. Kayal et. al. [14] handled Chen distribution under progressive censoring and Kayal et al. [13] studied inference of its parameters under progressive first-failure censoring.

It should be noted that all cited references are based on classical and Bayesian estimation methods for both complete and censored data sets. The generalized fiducial inference method has never been considered in comparative inference studies based on the Chen distribution. It is known that Newton-Raphson (NR) method can provide unsatisfactory performances explained by the fact that it does not converge in some cases. Since the MLE of β needs some iterative methods such as NR, alternative inference methods for the parameters of the Chen distribution are needed to evaluate. On the other hand, determining a prior distribution in the case of insufficient prior information about the parameters affects the Bayesian inference performance. Consequently, the generalized fiducial method can be worthwhile to overcome these adversities.

In this study, we consider the generalized fiducial inference (GFI) method based on the inverse of the structural equation which is proposed by Hannig et. al. [12]. We obtain the estimation of the unknown parameters of the Chen distribution with the GFI method as an alternative to the maximum likelihood estimation (MLE) and Bayesian estimation methods. We also provide the MLE and Bayesian estimation methods to compare the performances of the estimates and their corresponding confidence intervals. All theoretical outcomes are illustrated with simulation studies and a real-data example.

2. Maximum likelihood estimations (MLE)

The likelihood function of the observed sample from the Chen distribution is given as

$$L\left(\boldsymbol{x},\lambda,\beta\right)=\lambda^{n}\beta^{n}e^{(\beta-1)\sum_{i=1}^{n}\log(x_i)}e^{\sum_{i=1}^{n}x_i^{\beta}}e^{\lambda\sum_{i=1}^{n}\left(1-e^{x_i^{\beta}}\right)}$$

and the corresponding log-likelihood function is given as

$$\ell(\boldsymbol{x},\lambda,\beta)=n\log(\lambda)+n\log(\beta)+(\beta-1)\sum_{i=1}^{n}\log(x_i)+\sum_{i=1}^{n}x_i^{\beta}+\lambda\sum_{i=1}^{n}\left(1-e^{x_i^{\beta}}\right).$$

To obtain the MLEs of the parameters, denoted by $\hat{\lambda}$ and $\hat{\beta}$ we should equate the partial derivates of $\ell(\boldsymbol{x},\lambda,\beta)$ to zero with respect to λ and β respectively. Then we obtain the MLE of λ as given in the following

$$\hat{\lambda}=\frac{n}{\sum_{i=1}^{n}\left(e^{x_i^{\hat{\beta}}}-1\right)},$$

where $\hat{\beta}$ is the solution of the following non-linear equation

$$\xi(\beta)=\frac{n}{\beta}+\sum_{i=1}^{n}\log(x_i)+\sum_{i=1}^{n}x_i^{\beta}\log(x_i)-\frac{n\sum_{i=1}^{n}\left(e^{x_i^{\beta}}x_i^{\beta}\log(x_i)\right)}{\sum_{i=1}^{n}\left(e^{x_i^{\beta}}-1\right)}.$$

Since $\hat{\beta}$ is a fixed point solution of the nonlinear equation $\xi(\beta)$, it can be obtained by using numerical methods such as the Newton-Raphson algorithm. Further, the confidence interval for λ and β can be obtained by the following asymptotic normality when the MLE and its large sample theory exist. That is

$$(\hat{\lambda},\hat{\beta})^{T}\longrightarrow N\left((\lambda,\beta)^{T},\mathbf{I}_n^{-1}\right),$$

where

$$\mathbf{I}_n^{-1}=-\begin{pmatrix}\frac{\partial^2\ell}{\partial\lambda^2} & \frac{\partial^2\ell}{\partial\lambda\partial\beta}\\ \frac{\partial^2\ell}{\partial\beta\partial\lambda} & \frac{\partial^2\ell}{\partial\beta^2}\end{pmatrix}^{-1}=\begin{pmatrix}\mathrm{Var}(\hat{\lambda}) & \mathrm{Cov}(\hat{\lambda},\hat{\beta})\\ & \mathrm{Var}(\hat{\beta})\end{pmatrix}.$$

The inverse of the expected Fisher information matrix can be obtained by ***mle.tools*** package [16] in ***R*** [6] software. The presentation of the derivatives is skipped for the sake of simplicity.

Thus, the asymptotic $100(1-\alpha)\%$ confidence intervals (ACI) for λ and β are

$$I^{\lambda}:\hat{\lambda}\pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\lambda})}\quad\text{and}\quad I^{\beta}:\hat{\beta}\pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\beta})},$$

where z_δ denotes $100\delta\%$ percentile of the standard normal distribution.

3. Generalized fiducial inference (GIF)

The main structure of the generalized fiducial inference (GFI) is similar to the likelihood method. It differs from the likelihood method by switching the roles of the data $\boldsymbol{x}$ and the model parameters $\boldsymbol{\theta}$. Let suppose that the data generating equation be

$$\boldsymbol{x} = \boldsymbol{G}(\boldsymbol{U}, \boldsymbol{\theta}),$$

where $\boldsymbol{x}$ is the data, $\boldsymbol{\theta}$ is the parameters, $\boldsymbol{U}$ is a complete known random vector and $\boldsymbol{G}$ is called the structural equation. It is seen from Eq. (3), the distribution of $\boldsymbol{x}$ is determined by using the parameters $\boldsymbol{\theta}$ and random vector $\boldsymbol{U}$. Under some differentiability conditions, Hannig et al. [12] showed that the generalized fiducial distribution for $\boldsymbol{\theta}$ is absolutely continuous with the density

$$f_F(\theta) = \frac{L(x \mid \theta) J(x;\theta)}{\int L\left(x \mid \theta'\right) J\left(x;\theta'\right) \mathrm{d}\theta'},$$

where $L(x|\theta)$ denotes the joint likelihood function of observed data and

$$J(x,\theta) = \textstyle\sum_{(i_1,\dots,i_p)} \left|\det\left(\left(\frac{\partial}{\partial u}\boldsymbol{G}(u,\boldsymbol{\theta})\right)^{-1} \frac{\partial}{\partial\theta}\boldsymbol{G}(u,\theta)\right)\right|_{u=G^{-1}(x,\theta)} \tag{3.1}$$

where the above sums go $\binom{n}{p}$ over of p-tuples of indexes $i = 1 \leq i_1 < \dots < i_p \leq n$, $\partial G(u,\theta)/\partial\theta$ and $\partial G(u,\theta)/\partial u$ are respectively $n \times p$ and $n \times n$ Jacobian matrices. For the Chen distribution, we have

$$U_i = F\left(x_i;\lambda,\beta\right), \quad i = 1,\dots,n, \tag{3.2}$$

where $F(x_i;\lambda,\beta) \equiv 1 - e^{\lambda\left(1-e^{x^\beta}\right)}$ is the distribution function of the Chen model and U_i denotes the sample from uniform distribution on the range $(0,1)$. Further, the data generating equation, $x = G(U,\theta)$, can be obtained from Eq. (3.2) and the ith component $x_i = G(U_i,\lambda,\beta)$ can be obtained as

$$x_i = \left[\ln\left(1 - (1/\lambda)\ln(1-u)\right)\right]^{1/\beta},$$

and we have

$$\left.\frac{\partial G_i}{\partial\lambda}\right|_{u_i=1-e^{\lambda\left(1-e^{x^\beta}\right)}} = \frac{1}{\lambda\beta}\left(e^{-x_i^\beta}-1\right)x_i^{1-\beta} \qquad \text{and} \qquad \left.\frac{\partial G_i}{\partial\beta}\right|_{u_i=1-e^{\lambda\left(1-e^{x^\beta}\right)}} = -\frac{x_i\ln(x_i)}{\beta}. \tag{3.3}$$

Then, by replacing (3.3) in (3.1) we obtain

$$J(x;\lambda,\beta) = \frac{1}{\lambda\beta^2}\sum_{1\leq i<j\leq n}\left|g(x_i,x_j,\lambda)\right|,$$

where

$$g(x_i,x_j,\beta) = x_j^{1-\beta}\left(e^{-x_j^\beta}-1\right)x_i\ln(x_i) - x_i^{1-\beta}\left(e^{-x_i^\beta}-1\right)x_j\ln(x_j)$$

in that $J(x;\lambda,\beta)$ plays a similar role with the prior distribution in the Bayesian inference and it reveals like a data dependent prior. The joint likelihood function of the observed data is obtained as

$$f_F(\lambda,\beta) \propto \lambda^{n-1}\beta^{n-2}e^{(\beta-1)\sum_{i=1}^n \log(x_i)}e^{\sum_{i=1}^n x_i^\beta}e^{\lambda\sum_{i=1}^n\left(1-e^{x_i^\beta}\right)}\sum_{1\leq i<j\leq n}\left|g(x_i,x_j,\beta)\right|.$$

Thus, the conditional fiducial density function of λ can be obtained in form of the gamma density as

$$f_F(\lambda|\beta,x) = \lambda^{n-1}e^{-\lambda\sum_{i=1}^n\left(e^{x_i^\beta}-1\right)}\mathrm{GA}\left(n,\sum_{i=1}^n\left(e^{x_i^\beta}-1\right)\right).$$

Then, the conditional fiducial density function of β is

$$f_F(\beta|\lambda,x) = \beta^{n-2} e^{(\beta-1)\sum_{i=1}^n \log(x_i) + \sum_{i=1}^n x_i^\beta + \lambda\sum_{i=1}^n \left(1-e^{x_i^\beta}\right)} \sum_{1\le i<j\le n} |g(x_i,x_j,\beta)|\,.$$

It is clearly seen that estimates of the parameters can be obtained by using the Gibbs algorithm since their conditional densities are obtained. The conditional estimates of λ can be easily generated from the gamma density. However, the density of β in $f_F(\beta|\lambda,x)$ can not be reduced analytically to well known distributions and therefore it is not possible to sample directly by standard methods. The conditional posterior density of β is observed that, it is likely to be the Gaussian distribution.

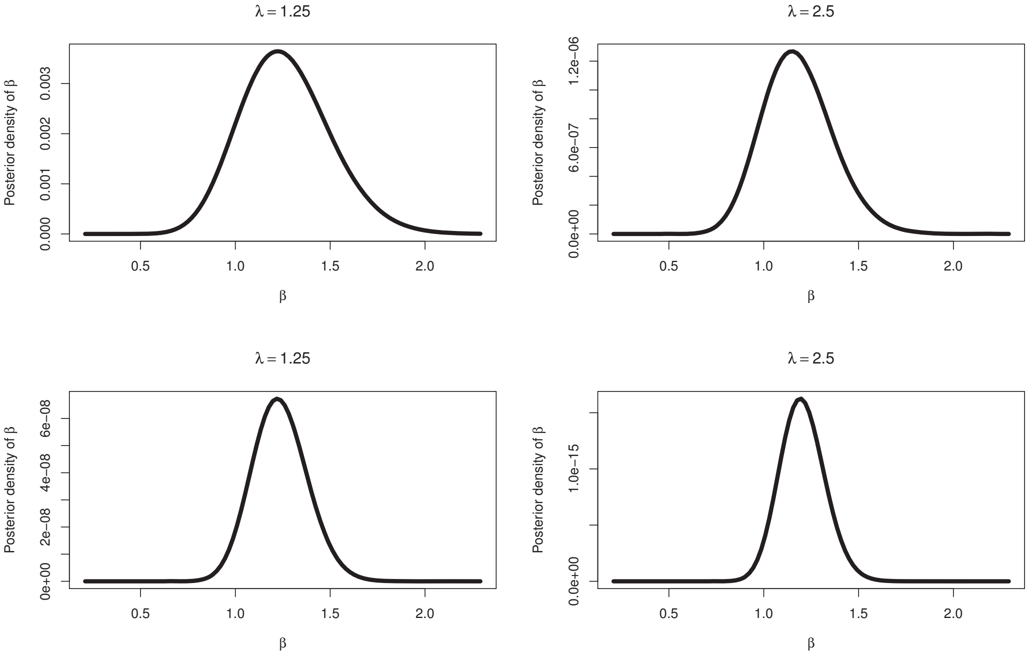


FIGURE 2. The posterior fiducial density of β for $n=20$ (up) and $n=50$ (down)

In this case, we propose to use the Metropolis-Hasting (M-H) sampling in Gibbs algorithm with normal proposal distribution (see Figure 2) as suggested by Tierney [20]. The Gibbs algorithm with the M-H sampling for the fiducial inference of the Chen distribution can be given as follows:

- **Step 1:** Start by using the initial values of $\lambda^{(0)}$ and $\beta^{(0)}$.
- **Step 2:** Set $t=1$.
- **Step 3:** Generate $\lambda^{(t)}$ from GA$\left(n, \sum_{i=1}^n \left(e^{x_i^\beta}-1\right)\right)$.
- **Step 4:** Draw a candidate $\beta^{(t)}$ from $f_F(\beta|\lambda,x)$ by using Metropolis-Hastings methods with normal proposal.
- **Step 5:** Repeat 2-4, M times.

It is known that a Markov chain algorithm naturally generates autocorrelated samples (Zhang, [29]) and so we should use a thinning operation to reduce the produced autocorrelation. We discard the first B_0 values as burn-in period and take every $L-$th (as an integer) observation from the remaining $(M-B_0)$ variates as an independent and identically distributed (i.i.d.) observation in thinning procedure. Thus, we obtain $M^{'}=(M-B_0)/L$ i.i.d. observations. If we denote the thinning procedure applied observations as $\alpha_i^{(t^{'})}$ and $\lambda_i^{(t^{'})}$ for $i=(B_0+L, B_0+2L, B_0+3L, \cdots, M)$, we obtain the fiducial inferences of the parameters as

$$\hat{\lambda}_F = \frac{1}{M'}\sum_{i=B_0+L}^{M}\lambda_i^{(t^{'})} \qquad \text{and} \qquad \hat{\beta}_F = \frac{1}{M'}\sum_{i=B_0+L}^{M}\beta_i^{(t^{'})}.$$

Then, the $100(1-\alpha)\%$ the fiducial cofidence intervals are

$$I_F^{\hat{\lambda}} \cong \left[\hat{\lambda}_{\alpha/2}, \hat{\lambda}_{1-\alpha/2}\right] \quad \text{and} \quad I_F^{\hat{\beta}} \cong \left[\hat{\beta}_{\alpha/2}, \hat{\beta}_{1-\alpha/2}\right],$$

where $\hat{\lambda}_\alpha$ and $\hat{\beta}_\alpha$ are the $100\alpha\%$th quantile of the $\lambda_i^{(t^{'})}$ and $\beta_i^{(t^{'})}$.

4. Bayesian inference

This section deals with the Bayesian inference to provide comparative estimates for the fiducial and maximum likelihood inference of the parameters. Since $J(x;\lambda,\beta)$ plays a similar role with the prior distribution in the Bayesian context and similarity of the mathematical structure of the fiducial inference process, the Bayesian inference method is handled as an alternative inference procedure. For this purpose, we first assume that the unknown parameters λ and β follow independent gamma priors such that $\pi(\lambda)\sim \mathrm{GA}(a_1,b_1)$ and $\pi(\beta)\sim \mathrm{GA}(a_2,b_2)$ with density functions are given as in the following

$$\pi(\lambda)\propto \lambda^{a_1-1}e^{-\lambda b_1} \qquad \text{and} \qquad \pi(\beta)\propto \beta^{a_2-1}e^{-\beta b_2},$$

where hyper parameters a_i and b_i, $(i=1,2)$ are assumed as non-negative and known.

The joint posterior density function of data, λ and β can be obtained by using the observed sample and the prior distributions for the parameters as in the following

$$\mathcal{L}(X,\lambda,\beta)=\mathcal{L}(X|\lambda,\beta)\pi(\lambda)\pi(\beta)$$

and the joint posterior density of λ and β given data is obtained by

$$\mathcal{L}(\lambda,\beta|X)=\frac{\mathcal{L}(X|\lambda,\beta)\pi(\lambda)\pi(\beta)}{\int_0^\infty\int_0^\infty \mathcal{L}(X|\lambda,\beta)\pi(\lambda)\pi(\beta)d\lambda d\beta}$$

and for the Chen distribution we have the following joint posterior density of the parameters

$$\mathcal{L}(\lambda,\beta|X)\propto \lambda^{n+a_1-1}\beta^{n+a_2-1}e^{-\beta\left(b_2-\sum_{i=1}^n \log(x_i)\right)}e^{\sum_{i=1}^n x_i^\beta}e^{-\lambda\left(b_1+\sum_{i=1}^n\left(e^{x_i^\beta}-1\right)\right)}.$$

Then, it is easily seen that the conditional posterior density functions of λ and β, denoted by $f_B(\lambda|\beta,x)$ and $f_B(\beta|\lambda,x)$, are obtained as

$$f_B(\lambda|\beta,x)\propto \lambda^{n+a_1-1}e^{-\lambda\left(b_1+\sum_{i=1}^n\left(e^{x_i^\beta}-1\right)\right)}\propto \mathrm{GA}(n+a_1,\sum_{i=1}^n(e^{x_i^\beta}-1)+b_1)$$

and

$$f_B(\beta|\lambda,x)\propto \beta^{n+a_2-1}e^{\beta\left(\sum_{i=1}^n\log(x_i)-b_2\right)+\sum_{i=1}^n x_i^\beta+\lambda\sum_{i=1}^n\left(1-e^{x_i^\beta}\right)}.$$

As in the fiducial process, we can easily generate samples of λ from the gamma density and Metropolis-Hasting with normal proposal is needed distribution for β. Thus, the point estimates of the parameters, $\hat{\lambda}_B$ and $\hat{\beta}_B$, can be obtained using the Gibbs sampling algorithm which was described in the fiducial process. Finally, the highest posterior density (HPD) $100(1-\gamma)\%$ credible intervals for the Bayesian estimates proposed by Chen and Shao [3] can be constructed as

$$I_B^{\hat{\lambda}} \cong \left(\hat{\lambda}_{B[\frac{\gamma}{2}(M-B_0)]}, \hat{\lambda}_{B[(1-\frac{\gamma}{2})(M-B_0)]}\right) \quad \text{and} \quad I_B^{\hat{\beta}} \cong \left(\hat{\beta}_{B[\frac{\gamma}{2}(M-B_0)]}, \hat{\beta}_{B[(1-\frac{\gamma}{2})(M-B_0)]}\right)$$

where $[\frac{\gamma}{2}(M-B_0)]$ and $[(1-\frac{\gamma}{2})(M-B_0)]$ are the smallest integers less than or equal to $\frac{\gamma}{2}(M-B_0)$ and $(1-\frac{\gamma}{2})(M-B_0)$, respectively.

5. Simulation studies

In this section, we perform some simulation studies to evaluate the performances of the generalized fiducial (GFI), ML and Bayesian estimators for the shape parameters of the Chen distribution. We consider three different combinations of the parameters (λ, β) as $(1.25, 1.25)$, $(0.50, 0.75)$ and $(2.00, 1.00)$. Small, moderate and larger sample sizes are considered as 10, 25 and 50, respectively. We run Markov chain with 3500 iteration, the first 500 values are discarded as Burn-in period then every third observation are taken in thinning procedure to generate uncorrelated and independent Markov chains. We replicate each chain 1000 times. The estimations are evaluated with their biases and mean squared errors (MSE). Further, we provide 95% approximate confidence intervals of the estimations and evaluated them according to their average lengths (AL) and coverage probabilities (CP). In the Bayesian estimations, we use the small and non-negative hyper-parameters as $a_1 = a_2 = b_1 = b_2 = 0.0001$ suggested by Congdon ([5], page 69) which are almost like Jeffrey's priors but they are proper, inversely. The biases and the MSEs of the estimates are reported in Table 1 and 2 then the corresponding credible intervals with their ALs and CPs are given in Table 3 and 4.

Table 1. The performances of estimations for β based on GFI, MLE and Bayesian methods

$\hat{\beta}$				Bias			MSE	
λ	β	n	GFI	MLE	BYS	GFI	MLE	BYS
1.25	1.25	10	0.02782	0.23687	0.08564	0.15144	0.22183	0.17905
1.25	1.25	25	0.01232	0.09554	0.04193	0.05587	0.06492	0.05984
1.25	1.25	50	0.00273	0.04103	0.01636	0.02553	0.02742	0.02650
0.50	0.75	10	0.02556	0.11549	0.02966	0.03893	0.05390	0.04403
0.50	0.75	25	0.01128	0.04732	0.01714	0.01466	0.01632	0.01521
0.50	0.75	50	0.00267	0.02032	0.00608	0.00666	0.00702	0.00682
2.00	1.00	10	0.02283	0.18994	0.06587	0.10344	0.14840	0.11809
2.00	1.00	25	0.00790	0.07682	0.03285	0.03770	0.04354	0.03988
2.00	1.00	50	0.00290	0.03294	0.01268	0.01755	0.01881	0.01809

TABLE 2. The performances of estimations for λ based on GFI, MLE and Bayesian methods

$\hat{\lambda}$				Bias			MSE	
λ	β	n	GFI	MLE	BYS	GFI	MLE	BYS
1.25	1.25	10	0.16140	0.33639	0.23991	2.14111	2.36967	2.26649
1.25	1.25	25	0.05823	0.10341	0.07452	1.10221	1.14054	1.11397
1.25	1.25	50	0.01813	0.03750	0.02475	0.73433	0.74718	0.73594
0.50	0.75	10	0.06494	0.03236	0.05080	0.75997	0.74430	0.74981
0.50	0.75	25	0.02384	0.00724	0.01654	0.45178	0.44779	0.44844
0.50	0.75	50	0.00997	0.00145	0.00619	0.31458	0.31361	0.31244
2.00	1.00	10	0.39813	0.85419	0.64314	4.67404	5.32950	5.25041
2.00	1.00	25	0.13142	0.25914	0.18812	2.16615	2.27428	2.21826
2.00	1.00	50	0.04035	0.09662	0.06515	1.38913	1.42470	1.39976

TABLE 3. The performances of aproximate confidence intervals for β based on GFI, MLE and Bayesian methods

$\hat{\beta}$				AL			CP	
λ	β	n	GFI	MLE	BYS	GFI	MLE	BYS
1.25	1.25	10	1.54294	1.62707	1.59565	94.60	96.30	95.20
1.25	1.25	25	0.90351	0.91553	0.90552	95.30	95.20	95.40
1.25	1.25	50	0.60660	0.61492	0.60934	93.70	94.40	94.00
0.50	0.75	10	0.79377	0.80036	0.79757	94.70	93.70	95.50
0.50	0.75	25	0.45399	0.45716	0.45258	95.10	94.10	95.10
0.50	0.75	50	0.30712	0.30958	0.30627	93.20	93.90	93.40
2.00	1.00	10	1.25740	1.33258	1.29795	94.50	95.90	96.00
2.00	1.00	25	0.74608	0.75799	0.74957	95.50	95.60	95.70
2.00	1.00	50	0.50429	0.51124	0.50675	93.40	94.20	93.80

TABLE 4. The performances of aproximate confidence intervals for λ based on GFI, MLE and Bayesian methods

$\hat{\lambda}$				AL			CP	
λ	β	n	GFI	MLE	BYS	GFI	MLE	BYS
1.25	1.25	10	2.14111	2.36967	2.26649	96.90	95.90	95.60
1.25	1.25	25	1.10221	1.14054	1.11397	96.20	96.20	95.50
1.25	1.25	50	0.73433	0.74718	0.73594	95.10	95.30	94.60
0.50	0.75	10	0.75997	0.74430	0.74981	96.30	92.10	95.20
0.50	0.75	25	0.45178	0.44779	0.44844	94.80	94.00	94.30
0.50	0.75	50	0.31458	0.31361	0.31244	94.90	94.00	95.50
2.00	1.00	10	4.67404	5.32950	5.25041	95.70	96.40	95.60
2.00	1.00	25	2.16615	2.27428	2.21826	95.80	97.50	95.60
2.00	1.00	50	1.38913	1.42470	1.39976	95.00	96.80	95.10

We observe satisfying consistency in the performances of the estimators. The biases, MSEs and the ALs of the confidence intervals decrease parallel to increasing sample sizes in all sets of the parameters. In whole cases, the GFI estimates have smaller biases, MSEs and ALs even in the small samples that are powerful side of the Bayesian estimation method. The differences between the performances of the proposed estimators are decreasing with the increasing sample sizes. In the whole case, the CPs are pretty close to their actual value 0.95. Many various values of the parameters are performed but only a few of them are reported here.

6. Numerical example

In this section, a real-life data is illustrated to compare different estimation procedures studied in this study. The data set which is given by Hand et al. [9] is handled. This data represents the survival period of 45 patients treated with chemotherapy. This data set is also fitted for the Chen distribution by Kayal et al. [13]. The data is given below as

1 63 105 129 182 216 250 262 301 301 342 354 356 358 380 383 383
388 394 408 460 489 499 523 524 535 562 569 675 676 748 778 786
797 955 968 1000 1245 1271 1420 1551 1694 2363 2754 2950

We divided data points by 3000 to simplify computations. Then, we fit this dataset with the Chen distribution by using the MLE, GFI and Bayesian inference methods.

We also evaluate the convergence of the Markov chains. We perform Markov chains 100 500 times and we discard the first 500 values as burn-in period and we count in every tenth variate as independent and uncorrelated samples. Thus, we obtain 10 000 uncorrelated and independent samples. In the Bayesian procedure, we use very small non-negative values of the hyper-parameters, i.e. $a_1 = a_2 = b_1 = b_2 = 0.0001$, as suggested by Congdon ([5], page 69) which are almost like Jeffrey's priors but they are proper, inversely.

The parameter estimates of the parameters and their corresponding confidence intervals are obtained as given in Tables 5-6, respectively.

Table 5. Estimations, K-S test values and p- values for the real data example

	MLE	GFI	Bayes
λ	3.1979	3.0759	3.1290
β	0.9804	0.9444	0.9571
K-S	0.1778 (0.4756)	0.1556 (0.6476)	0.2222 (0.2165)

Table 6. Confidence intervals with their lengths for the real data example

	ACI	FCI	BCI
λ	(1.9157,4.4803)	(1.9810,4.4835)	(1.9921, 4.5752)
	2.5646	2.5025	2.5831
β	(0.7348,1.2261)	(0.7108,1.2014)	(0.7212,1.2079)
	0.4914	0.4906	0.4867

The Kolmogorov–Smirnov (KS) test statistics and the associated p-values for all inference procedures are obtained as bigger than 0.05. Therefore, we can not reject the null hypothesis that this data set comes from the Chen distribution. Also, the estimated density and the emprical cdf plots support this observation as seen in Figure 3.

Further, the convergence of the Markov chains is evaluated with trace (Figures 4-5), density (Figures 6-7) and running mean (ergodic average) (Figures 8-9) plots. A trace plot is a plot of the parameter values in each iteration of the Markov chain against the iteration number. It is expected to observe that the Markov chain disperses around its center with a similar variation. In our example, trace plots of the Markov chains provide expectations and fluctuate around their centers with similar variations. Further, the posterior density plots of λ and β via GFI and Bayesian methods obtained almost symmetrical and in the shapes of unimodal. The trace, density and

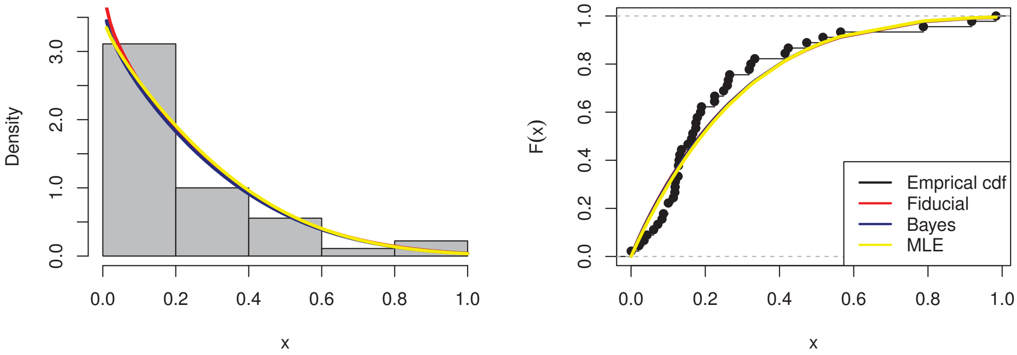


FIGURE 3. Estimated density and empirical cdf for real-data example fitted by the Chen distribution

running mean plots can be drawn using by the **traplot**, **denplot** and **rmeanplot** functions in library **mcmcplots** (Curtis et al., [7]) in **R** [6] software.

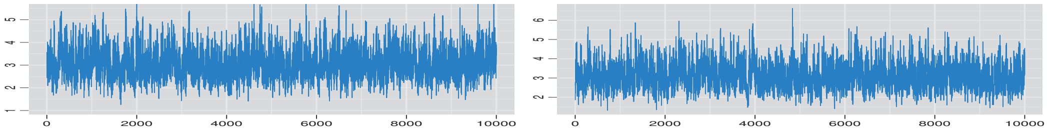
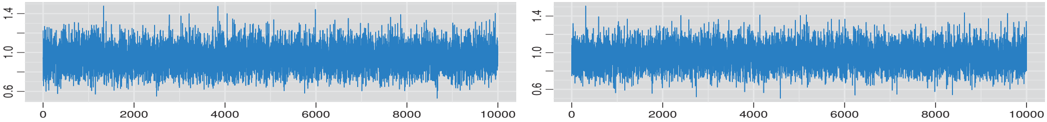


FIGURE 4. Trace plots for λ via GFI (on left) and the Bayesian (on right) methods



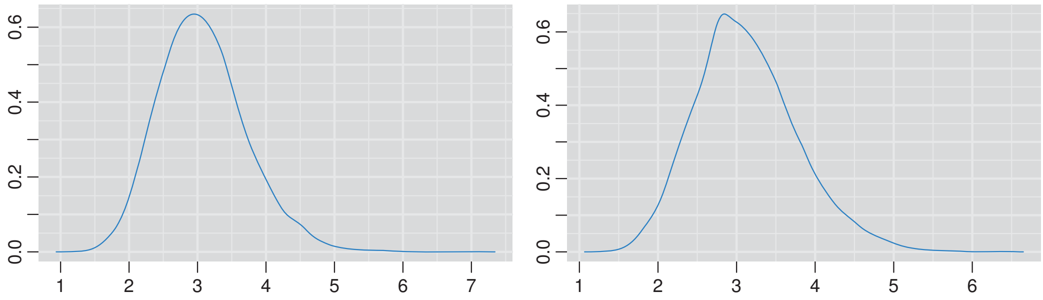


FIGURE 6. Density plot for the posterior distribution of λ via GFI (on left) and the Bayesian (on right) methods

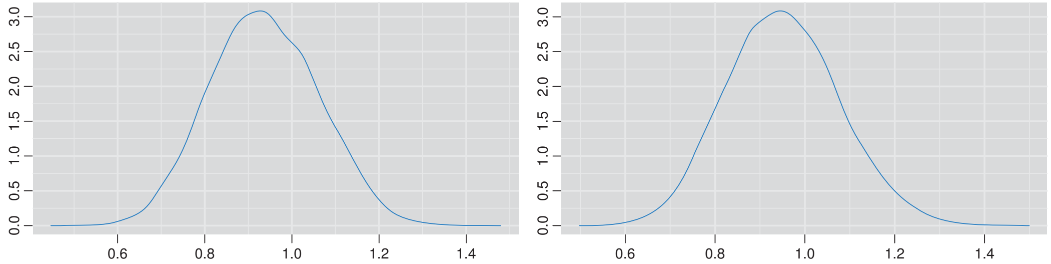


FIGURE 7. Density plot for the posterior distribution of β via GFI (on left) and the Bayesian (on right) methods

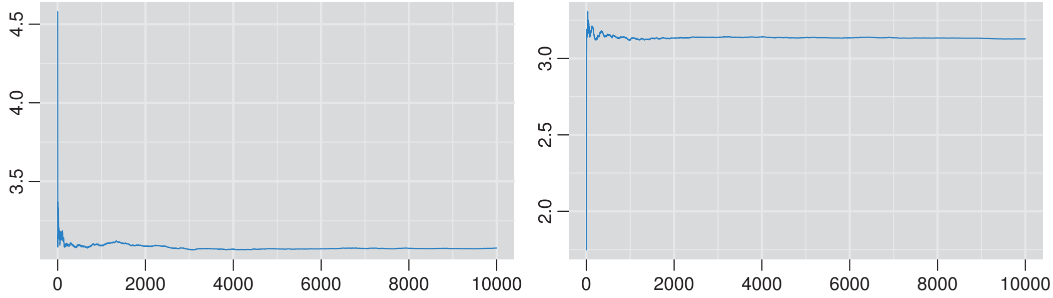


FIGURE 8. Running mean plot for λ via GFI (on left) and the Bayesian (on right) methods

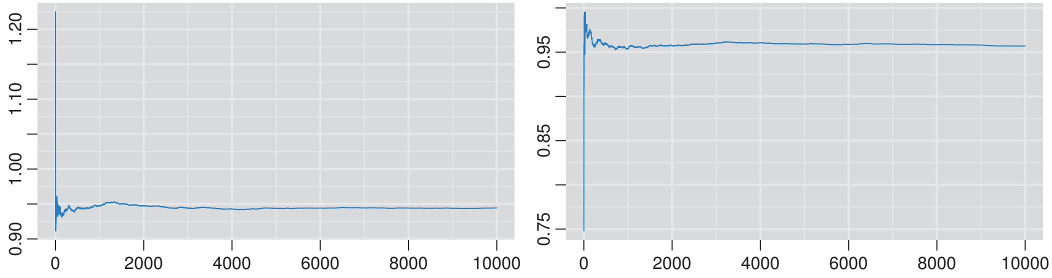


FIGURE 9. Running mean plot for β via GFI (on left) and the Bayesian (on right) methods

7. Conclusions

On the basis of this study, the generalized fiducial inference is considered for the parameter estimates of the Chen distribution. Further, the MLE and Bayesian procedures are handled as alternative methods to this inference method. The performances of the simulation schemes show that the GFI method has superiority in parameter estimations of the Chan distribution over the classical and Bayesian estimation methods. The GFI method provides better results than MLE and Bayesian methods in most cases even in the case of small, moderate or large sample sizes. The theoretical findings are also evaluated on a real data example. Additionally, the convergence of the Markov chains generated in the GFI and Bayesian procedures are provided. These observations are supported by the graphical methods. Consequently, the generalized fiducial inference method based on the inverse of the structural equation which is proposed by Hannig et. al. [12] should be proposed as a more efficient estimator for the parameter estimation of the Chen distribution.

References

[1] Abramowitz, M. and Stagun, I. (1964). *Handbook of Special Functions*. National Bureau of Standards, Dover Publications, New York.

[2] Ahmed, E.A. (2014). Bayesian estimation based on progressive Type-II censoring from two-parameter bathtub-shaped lifetime model: an Markov chain Monte Carlo approach. *Journal of Applied Statistics*, 41(4), 752-768.

[3] Chen, M.H. and Shao, Q.M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, 8(1), 69-92.

[4] Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics & Probability Letters*, 49(2), 155-161.

[5] Congdon, P. (2006). *Bayesian Statistical Modelling*. Second edition, John Wiley & Sons, England.

[6] Core Team, R. (2021). R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna. https://www.R-project.org

[7] Curtis, S.M., Goldin, I. and Evangelou, E. (2018). Package 'mcmcplots' [computer software]. R package version 0.4.3.

[8] Fisher, R.A. (1930). Inverse Probability. *Proceedings of the Cambridge Philosophical Society*, xxvi, 528-535.

[9] Hand, D.J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. (1993). *A handbook of small data sets*. First edition, CRC Press, Boca Raton.

[10] Hannig, J. (2013). Generalized fiducial inference via discretization. *Statistica Sinica*. 23(2), 489-514.

[11] Hannig, J. (2009). On generalized fiducial inference. *Statistica Sinica*, 19(2), 491-544.

[12] Hannig, J., Iyer, H., Lai, R.C. and Lee, T.C. (2016). Generalized fiducial inference: A review and new results. *Journal of the American Statistical Association*, 111(515), 1346-1361.

[13] Kayal, T., Tripathi, Y.M. and Wang, L. (2019). Inference for the Chen distribution under progressive first-failure censoring. *Journal of Statistical Theory and Practice*, 13(4), 1-27.

[14] Kayal, T., Tripathi, Y.M., Singh, D. P. and Rastogi, M.K. (2017). Estimation and prediction for Chen distribution with bathtub shape under progressive censoring. *Journal of Statistical Computation and Simulation*, 87(2), 348-366.

[15] Li, Y. and Xu, A. (2016). Fiducial inference for Birnbaum-Saunders distribution. *Journal of Statistical Computation and Simulation*, 86(9), 1673-1685.

[16] Mazucheli, J. and Mazucheli, M.J. (2017). Package 'mle. tools'.

[17] O'Reilly, F. and Rueda, R. (2007). Fiducial inferences for the truncated exponential distribution. *Communications in Statistics - Theory and Methods*, 36(12), 2207-2212.

[18] Rastogi, M.K. and Tripathi, Y.M. (2013). Estimation using hybrid censored data from a two-parameter distribution with bathtub shape. *Computational Statistics & Data Analysis*, 67, 268-281.

[19] Sarhan, A.M., Hamilton, D.C. and Smith, B. (2012). Parameter estimation for a two-parameter bathtub-shaped lifetime distribution. *Applied Mathematical Modelling*, 36(11), 5380-5392.

[20] Tierney, L. (1994) Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4), 1701-1728.

[21] Wandler, D.V. and Hannig, J. (2011). Fiducial inference on the largest mean of a multivariate normal distribution. *Journal of Multivariate Analysis*, 102(1), 87-104.

[22] Wandler, D.V. and Hannig, J. (2012). Generalized fiducial confidence intervals for extremes. *Extremes*, 15, 67-87.

[23] Wang, C.M., Hannig, J. and Iyer, H.K. (2012). Fiducial prediction intervals. *Journal of Statistical Planning and Inference*, 142(7), 1980-1990.

[24] Weeranhandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association*, 88, 899-905.

[25] Wu, J.W., Lu, H.L., Chen, C.H. and Wu, C.H. (2004). Statistical inference about the shape parameter of the new two-parameter bathtub-shaped lifetime distribution. *Quality and Reliability Engineering International*, 20(6), 607-616.

[26] Wu, S.J. (2008). Estimation of the two-parameter bathtub-shaped lifetime distribution with progressive censoring. *Journal of Applied Statistics*, 35(10), 1139-1150.

[27] Yan, L. and Liu, X. (2018). Generalized fiducial inference for generalized exponential distribution. *Journal of Statistical Computation and Simulation*, 88(7), 1369-1381.

[28] Zabell, S. L. (1992). R.A. Fisher and fiducial argument. *Statistical Science*, 369-387.

[29] Zhang, Z. (2014). WebBUGS: Conducting Bayesian statistical analysis online. *Journal of Statistical Software*, 61(1), 1-30.

İSTATİSTİK: JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION İSTATİSTİK
Vol. 14, No. 2, December 2022, pp. 87–96
ISSN 1300-4077 | 22 | 2 | 87 | 96

A NEW COMPUTATIONAL APPROACH BASED ON DENSITY CLUSTERING FOR OUTLIER PROBLEMS IN LINEAR MODELS

Fatma Yerlikaya-Özkurt*
Department of Industrial Engineering,
Atılım University,
06830, Ankara, Turkey

Abstract: Recently, collection of huge amount of data and analysis of that much data have vital importance for human activities in many different application areas. Advanced statistical methods play crucial role for modeling of such data when the data contains outliers. Although there are number of outlier detection methods for revealing outlier observations in data, most of them may not be reasonable and appropriate for prediction purposes due to structural and requirements of modeling. In this study, density based clustering algorithm named Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is considered in order to detect the location of outlier observations effectively with respect to form of the model for given data set. Based on obtained results, the Mean Shift Outlier Model (MSOM) is constructed as a robust linear model. This newly proposed computational approach based on DBSCAN uses power of data clustering and also minimize the impact of the outlier observations by MSOM. The numerical examples are also presented to reveal the performance of the proposed approach in this study.

Key words: Outlier problem, Mean shift outlier model, Density based clustering.

1. Introduction

Recently, collection of huge amount of data and analysis of that much data have vital importance for human activities in many different application areas. Most of the statistical applications involve regression models for doing estimation and prediction. Among regression models, Linear Regression Model (LRM) is one of the most used ones by many researchers who prefer well-established form, ease of application and interpretability of the model [11]. Generally, LRM is used to investigate the relationship between a response (dependent) variable and explanatory (predictor or independent) variable(s) through estimation parameter(s). Moreover, parameter estimation of LRM is mainly based on a least squares method which can be seriously hindered by the presence of outlier observation(s) [15, 16].

Outliers occur because of changes in system behavior, human or machine error, or natural deviations in observations. In fact, these observations reduce and affect the information that we may get from the source. For this reason, it is very important to identify the existing outlier observations in given dataset [2]. Although there are number of outlier detection methods for revealing outlier observations in the dataset, most of them may not be reasonable and appropriate for prediction or estimation [9]. Thus, advanced methods play crucial role for outlier identification.

In this study, outlier observations are considered as the data points that distorting model and reducing model performance. For the detection of such outlier observations, existing statistical methods can be categorized into two which are traditional and advanced approaches. The first approach, generally, provides good result for small or relatively medium size dataset, but they fail when the dataset is high dimensional. The second approach, on the other hand, gives good performance with very low computational time on any size of dataset but especially it provides

* Corresponding author. E-mail address:fatma.yerlikaya@atilim.edu.tr

very good results on high-dimensional datasets. Therefore, advanced methods play crucial role for outlier identification [3, 6, 12, 21, 22, 23]. In this study, the new approach is proposed for outlier identification with advanced data mining tool named clustering.

There are different types of clustering algorithms such as hierarchical clustering, partitioning clustering and density based clustering. Among them density based clustering is appropriate to find outliers since it captures the data structure well with respect to regional density [8]. The most popular density based clustering algorithm named Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is preferred in this study for outlier identification [7, 24]. Based on obtained results via DBSCAN, these observations are modeled with Mean Shift Outlier Model (MSOM) which is a robust linear model.

This paper is organized as follows: Section 2 briefly reviews linear models and then MSOM is presented in detail. Section 3 presents some outlier identification methods. A background on clustering and a new outlier detection approach based on DBSCAN algorithm are also provided in the same section. Section 4 includes applications and comparisons of the new approach against existing alternative in order to illustrate the efficacy of the proposed approach. Section 5 summarizes and concludes the paper.

2. Improvements on linear model with mean shift outlier model

There are various way of modeling to handle outlier observation(s) within a dataset of interest using Linear Models (LMs). The general form of the LRM with n observations and p independent variables is given by:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon, \tag{2.1}$$

where Y is the response variable and X_j $(j = 1, 2, ..., p)$ are the predictor variables. The vector of predictors is represented by $\boldsymbol{X} = (X_1, X_2, ..., X_p)^T$. The coefficient (unknown parameter) β_0 is the intercept and the rest of the unknown parameters β_j are the regression coefficients of the independent variables X_j $(j = 1, 2, ..., p)$, and ϵ is the random error term which is generally called noise [16]. If the response values (y_i) and predictor vectors $(\boldsymbol{x}_i$ $(i = 1, 2, \ldots, n))$ are inserted into the model in Eq. (2.1), the following linear system will be obtained:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}. \tag{2.2}$$

Here, $\boldsymbol{y}$ is an $(n \times 1)$-vector of the response variable, $\boldsymbol{x}_i$ $(i = 1, 2, ..., n)$ is a $(1 \times (p+1))$ row vectors of $\boldsymbol{X}$ matrix which is a *full rank* $(n \times (p+1))$-matrix of predictor variables and $\boldsymbol{\beta}$ is a $((p+1) \times 1)$-vector of coefficients. Moreover, $\boldsymbol{\epsilon}$ is an $(n \times 1)$-vector of independently, identically distributed random errors. The corresponding mean and standard deviation are given by $E(\boldsymbol{\epsilon} \mid \boldsymbol{X}) = 0$ and $\mathrm{Var}(\boldsymbol{\epsilon} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}$. Here, σ is an unknown parameter and $\boldsymbol{I}$ is the n dimensional identity matrix. Based on the least squares estimates, $\boldsymbol{\beta}$ and σ are given by $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $\sigma = \sqrt{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{y}/(n-p-1)}$, where $\boldsymbol{H} := \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ is called *hat operator* [17]. In order to handle outlier observation(s) for LMs, there are two main approaches called *Direct Approaches (DA)* and *Indirect Approaches (IA)*. These approaches are based on residuals from the robust regression. In general, the robust regression provides more stable results than LRM in the presence of outliers. There are three different types of outlier problems: Problems with outliers occurred in the vertical direction, problems with outliers occurred in the horizontal direction, and problems with outliers occurred at leverage points [1, 9, 18]. Figure 1 shows simple demonstration of the outliers in vertical direction (×), horizontal direction (+) and at leverage point (•). The mostly used robust regression methods to deal with outlier observation(s) in a dataset are M estimation [13], Least Trimmed Square estimation [18] and MSOM [5, 14]. In this study, the MSOM is employed in order to model the dataset consists of outliers which is describe next.

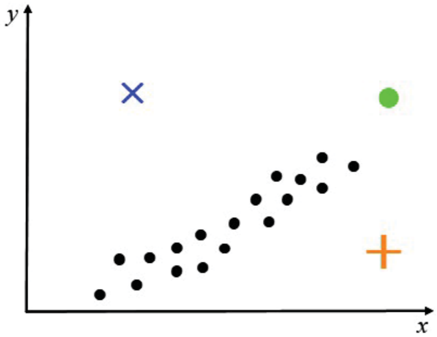


FIGURE 1. Graphical representation of outliers occurred in vertical direction, horizontal direction and at leverage point

2.1. Mean shift outlier model

The general form of the MSOM is given by:

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + \Theta\theta + \epsilon,$$

where $\Theta \in \{0,1\}$ is a constant term, and θ is the coefficient for outlier observation. In the absence of an outlier, $\Theta = 0$, and the contribution of an outlier is represented by the value θ. The linear system takes the following form after inserting all data values to the model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}_i\theta + \boldsymbol{\epsilon},$$

where $\boldsymbol{e}_i$ is the ith unit vector, i.e., $\boldsymbol{e}_i = (0,...,1,0,...,0)^T$ $(i=1,2,...,n)$. In this linear system, it is assumed that either y_i or $\boldsymbol{x}_i\boldsymbol{\beta}$ deviates systematically from the model $y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i$ by some value θ. Then, the ith observation $(y_i, \boldsymbol{x}_i\boldsymbol{\beta})$ would have a different intercept than the remaining observation, and $(y_i, \boldsymbol{x}_i\boldsymbol{\beta})$ would hence be an outlier [5, 14].

After detecting the m outliers $(m < n)$ in the dataset, the MSOM can be written as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ have same descriptions as in Eq. (2.2). On the other hand, $\boldsymbol{E}$ is an $(n \times m)$-matrix with m indicator variables, and $\boldsymbol{\theta}$ is an $(m \times 1)$-vector of the coefficients of the indicator variables. More compact form of the MSOM is rewritten as:

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\boldsymbol{X}^* = (\boldsymbol{X} \mid \boldsymbol{E})$ is an $(n \times (p+1+m))$ block matrix constructed by the matrices $\boldsymbol{X}$ and $\boldsymbol{E}$, and $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T\right)^T$ is an $((p+1+m) \times 1)$-vector constructed by the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

It should be noted that MSOM (as presented in Eq. (2.3)) gives the same residual sum of squares as the model fitted after omitting the outlier observations [20]. Therefore, MSOM is particularly convenient and preferred instead of the linear regression model in the presence of outliers.

3. Outlier identification methods

Identification of outliers is the key step before modeling with MSOM. In order to build MSOM with having good predictions, outliers should be carefully analyzed. Otherwise, the prediction model may give misleading results. Although there are various outlier detection methods, most of these methods are useless when modeling is taken into account. In this study, model based outlier identification methods are focused on. For this purposes, firstly traditional then advanced methods are introduced.

For a given dataset with n observations, the m outliers $(m < n)$ can be detected by direct approaches such as Likelihood-Ratio Test Statistic, Cooks Distance or Studentized Residuals which are described below [16].

Likelihood-Ratio Test Statistic (F_i):

$$F_i = \frac{(RSS_1 - RSS_2)/1}{RSS_2/(n-p-1)}.$$

Here, RSS_1 is the residual sum of squares obtained by using all the n observations in the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and RSS_2 is the residual sum of squares in the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}_i\theta + \boldsymbol{\epsilon}$

Cooks Distance (CD_i):

$$CD_{-i} = \frac{\left(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{-i}\right)^T \left(\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}_{-i}\right)}{p\hat{\sigma}^2},$$

where $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{y}}_{-i}$ represent the response vector and the estimated response vector after omission of the ith observation, respectively. And $\hat{\sigma}^2$ is obtained by sum of square error divided by $(n-p)$ [20].

Studentized Residuals (r_i):

$$r_i = \frac{\hat{\epsilon}_i}{\sigma_i\sqrt{(1-h_{ii})}} \quad (i = 1, 2, ..., n),$$

where σ_i is the standard deviation of the ith residual and $\hat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{H})\,\boldsymbol{y}$, $\hat{\epsilon}_i = \boldsymbol{e}_i^T\hat{\boldsymbol{\epsilon}}$, and $\boldsymbol{e}_i^T\boldsymbol{H}\,\boldsymbol{e}_i = h_{ii}$.

An observation is defined as an outlier if it has larger Cook's distance and Studentized residual values. In order to find all potential outliers, the following steps are applied to a given dataset and repeated until all of the outlier observations are defined.

1. The LRM is constructed to fit the data.
2. The fitted values and ordinary residuals are obtained to check the better prediction.
3. The direct approaches described above are calculated to extract potential outliers.
4. The potential outlier is removed from the dataset, and the first three steps are repeated until detecting all potential outliers.

These steps are computationally slow for analyzing and detecting outlier(s) in a dataset, especially, in case of large scale dataset. These steps also contain a high error rate. Thus, it is important to have an accurate, reliable, and fast computational method for the identification of the outliers. At this stage advanced data mining tools, especially, clustering techniques play crucial role.

3.1. A background on clustering

Clustering is a data mining technique to identify the group of unlabeled data points of a data set that are similar and dissimilar to each other. Clusters are formed by assigning most similar objects (data points, entities) to the same group and dissimilar ones to the separate groups as much as possible [8].

There are different types of clustering such as hierarchical clustering, partitioning clustering and density based clustering and each preferred for different purposes. Hierarchical clustering aims to construct clusters that have an ordering from bottom to top like a tree structure. As a result it produce the hierarchical relation between the created clusters. There are two kinds of hierarchical clustering named divisive and agglomerative. Divisive clustering is splitting the single all inclusive cluster into two until having only clusters with one data point. Whereas agglomerative clustering (bottom-up approaches) is starting from the single data point as an individual cluster and merging clusters at each iteration until getting a single all inclusive cluster [24].

Partitioning clustering is based on clustering of n unlabeled data points to k clusters in which each cluster contains at least one data points. The purpose of the partitioning clustering is to minimize the distances of data points in a cluster whereas to maximize the distances between the separated clusters. In partitioning clustering, after defining number of clusters (k), the next step is to assign k random initial centers. The cluster centers are updated based on the data points assigned to a given cluster. This procedure repeatedly continues until the assigned cluster points of a sample can not be updated [24].

On the other hand, density based clustering is a clustering method that identify the arbitrarily shaped clusters in data according to the idea of a cluster is being a region with high density and separated from the other such clusters by regions of low density. Although, this type of clustering algorithms have high complexity, they can easily identify outliers in the data set. Moreover, they can handle noise and can detect the clusters automatically since they can scan the data well [8]. The most popular density-based clustering algorithm that proposed in this study to identify outliers is Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

3.1.1. DBSCAN to identify outlier

DBSCAN is a density-based clustering algorithm based on the density of the data points or closeness of the data points [7]. The points outside the dense regions are extracted and treated as outliers. This property of the DBSCAN algorithm makes it a powerful method for outlier detection. The other clustering algorithms such as k-means clustering lack this property and are very sensitive to outliers since existence of outliers can easily influence the construction of the clusters [6, 12].

DBSCAN starts with the estimation of density over a dataset with n observations. It estimates the density around each observation using epsilon neighborhood concept (eps). DBSCAN depends on the eps and a threshold value (MinPts) to detect dense regions into dataset and to classify the observation as a core, a border, or an outlier. Illustration of the concept of DBSCAN algorithm is given in Figure 2 [10].

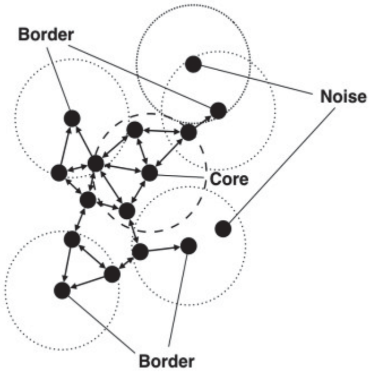


Figure 2. Illustration of the concept of DBSCAN algorithm given by Hahsler et. al. (2019)

The DBSCAN constructs all clusters by defining all core points which have high density and expanding each cluster to all reachable points by retrieving their epsilon neighborhood. The search continues until no more core points are found in the expanded neighborhood. End of the search, the cluster is constructed and the observations that are outside of the cluster are assigned as outliers [7, 10].

4. Applications and results

In this section, in order to apply the proposed approach, firstly, the datasets used in this study are introduced. Then, the well-known prediction performance measures are provided in the following subsections. Finally, the details of the applications and results and are presented in the last subsection. It should be noted that all computational parts of the DBSCAN, LRM and MSOM are conducted through R programming. Specifically, the R packages "dbscan" and "devtools" are installed for running DBSCAN algorithm while detecting outlier observations [10].

4.1. Data sets

4.1.1. Real world data set

The first dataset, a stack loss data, is selected from SAS Customer Support [19]. It is well-known and -studied for outlier analysis in LM. This dataset is about the operation of a plant for the oxidation of ammonia to nitric acid. It contains $n = 21$ observations, $p = 3$ explanatory variables which are the rate of operation (X_1), the cooling water inlet temperature (X_2), and the acid concentration (X_3). The response variable (Y) is the stack-loss. All variables' observations of this dataset are shown below:

X_1: 80 80 75 62 62 62 62 62 58 58 58 58 58 58 50 50 50 50 50 56 70

X_2: 27 27 25 24 22 23 24 24 23 18 18 17 18 19 18 18 19 19 20 20 20

X_3: 89 88 90 87 87 87 93 93 87 89 89 88 82 93 89 86 72 79 80 82 91

Y : 42 37 37 28 18 18 19 20 15 14 14 13 11 12 8 7 8 8 9 15 15

4.1.2. Simulation data set

For the simulation dataset, the data generation is based on the LM given in Eq. (2.1). The matrix of predictors is obtained from a multivariate normal distribution with zero mean and one constant ($N(0, 1)$). The random error vector is obtained from again normal distribution with $N(0,$ $1)$. For one dimensional vector of unknown parameter, the randomly generated value between zero and ten are preferred. In order to demonstrate the outlier identification ability of the proposed approach relatively large size dataset is chosen ($n = 1000$). After finalizing data generation, randomly 40 observations are defined and shifted with some values in order to convert them as outlier observations.

4.2. Prediction performance measures

The performance measures with their formulas used in this study are given as follows [15]:

Residual Sum of Squares (RSS)=: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$,

Mean Squared Error (MSE)=: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$,

Root Mean Square Error (RMSE)=: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$,

Multiple Coefficient of Determination (R^2)=: $1 - \left(\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\right)$,

Adjusted R^2 (AdjR^2)=: $1 - \left(\frac{(1-R^2)(1-n)}{n-p-1}\right)$,

Correlation Coefficient (r)=: $\sqrt{R^2}$, where y_i is ith observed response value, $\hat{y}_i$ is ith fitted response value, $\bar{y}$ represents mean response value.

4.3. Results and findings

For the first dataset, after applying carefully the outlier detection steps given in Section 3, the observations 1, 2, 3, 4 and 21 are defined as outliers. It take quite long time to identify these outliers since these steps require high human intervention and they are not conducted automatically.

The linear model and related performance results obtained in the presence of all observations are given below and in Table 1, respectively.

$$Y_{LM} = -42.1062 + 0.7124\,X_1 + 1.2625\,X_2 - 0.1159\,X_3.$$

The performance results obtained after each potential observation (observations 1, 2, 3, 4 and 21) is removed from the data one by one, are presented in Table 1. In addition, the performance results obtained after removing all potential observations from the dataset are presented in the same table.

Table 1. Performance results of LM and MSOM & performance results obtained after removing each or all of the potential outlier observation(s) from the dataset

Measures	**MSE**	**RMSE**	**R^2**	**AdjR^2**	**r**
LM	8.7338	2.9553	0.9114	0.8957	0.9547
Outlier(1)	8.3682	2.8928	0.9401	0.8837	0.8620
Outlier(2)	8.9394	2.9899	0.9450	0.8930	0.8730
Outlier(3)	7.9168	2.8137	0.9514	0.9052	0.8875
Outlier(4)	7.2903	2.7001	0.9620	0.9254	0.9114
Outlier(21)	5.3198	2.3065	0.9739	0.9484	0.9387
All Outliers	1.0059	1.0029	0.97050	0.9419	0.9274
MSOM	0.7664	0.8754	0.9922	0.9870	0.9961

However, if DBSCAN algorithm is applied to given data set with eps=8 and MinPts=4 parameters, exactly same outlier observations are detected in less than a minute. The graphical representation of the clusters and outliers is given in Figure 3. In this figure, the outliers are demonstrated by black points.

After detecting all outliers for the given dataset, MSOM is built and same performance measures are calculated and given in Table 1.

$$\begin{aligned} Y_{MSOM} \;=\; & -36.6978 + 0.6658\,X_1 + 0.5673\,X_2 - 0.0103\,X_3 \\ & +9.2070\,O_1 + 4.2172\,O_2 + 8.6601\,O_3 + 9.9131\,O_4 - 7.1848 O_{21}, \end{aligned}$$

where O_i for $i = 1,2,3,4,21$ represents the ith outlier observation. In addition, if the coefficients of the outlier observations are compared against the coefficients of independent variables, the contributions of outlier variables to the model are much more than independent variables. If the performance results of LM and MSOM are compared according to all performance measures, it is obvious that MSOM based on DBSCAN is much more better than LM and the LM that obtained even after removing all outlier observations.

Moreover, another application of the proposed approach is conducted by using simulated dataset which contains 40 outlier observations that randomly constructed. The DBSCAN algorithm is applied to this dataset with eps=0.3 and MinPts=10 parameters. All outliers are correctly defined and represented in Figure 4.

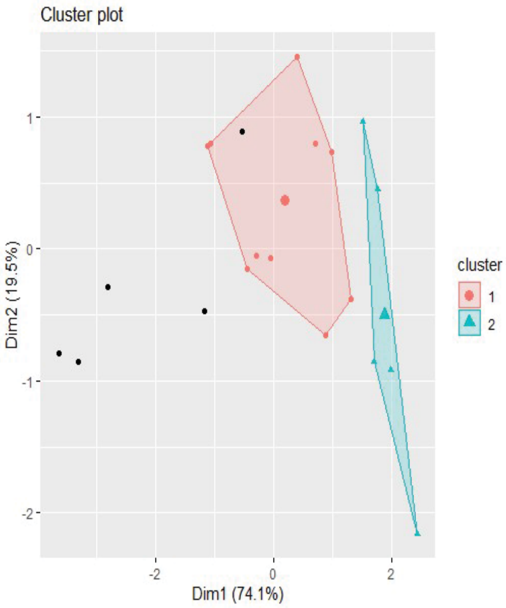


FIGURE 3. Graphical representation of outliers and clusters for stack loss dataset.

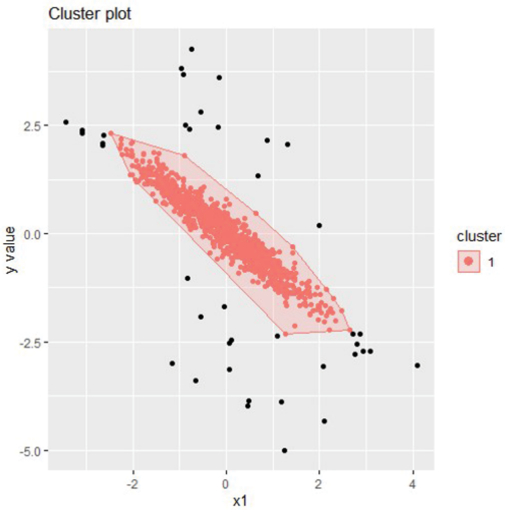


FIGURE 4. Graphical representation of outliers and clusters for simulated dataset

After detection of outliers for the simulated dataset, LM and MSOM are constructed. Performance results of the models are given in Table 2. The results in this table show that the proposed approach is still a much better than traditional approaches as the data size increases or the number of outliers in the dataset increases.

To summarize, for the computational process of outlier detection, we use DBSCAN algorithm. By using this clustering algorithm, the MSOM is improved in terms of CPU time and user effort.

TABLE 2. Performance results of LM and MSOM based on simulated dataset

Measures	**MSE**	**RMSE**	$\mathbf{R}^2$	$\mathbf{AdjR}^2$	**r**
LM	0.0473	0.2175	0.7278	0.7276	0.8531
MSOM	0.0112	0.1057	0.9357	0.9330	0.9973

5. Conclusion

Main goal of this study is proposing a new approach for a robust LM estimation within the existence of outliers. This new computational approach is based on DBSCAN and MSOM methods. DBSCAN is used for detecting the location of outlier observations effectively since it is fast, stable under perturbations on data and appropriate also for high dimensional dataset. On the other hand, MSOM is constructed as a robust linear model to overcome instability in modeling, and it also does not ignore outlier observations that are necessary to model the data adequately. The proposed method has been performed on real world and simulated datasets. It is observed that this approach performs quite well in terms of computational time and accurate detecting ability of the outlier observations than the traditional methods.

It is always possible to improve this new approach for future applications. Recommendations can be summarized as follows:

- In this study, MSOM is used as a robust model. In future, in order to capture nonlinear structure in the dataset, instead of independent variables, MSOM can be formed by using data based basis functions.
- In future applications, it is also possible to apply this new approach for classification type of datasets.
- This approach can also be effectively applied to high dimensional datasets in the existence of outliers.

References

[1] Bakar, Z.A., Mohemad, R., Ahmad, A. and Deris, M.M. (2006). A comparative study for outlier detection techniques in data mining. *In 2006 IEEE Conference on Cybernetics and Intelligent Systems IEEE*, 1-6.

[2] Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. Wiley, Great Britain.

[3] Campulova, M., Michalek, J. and Moucka, J. (2019). Generalised linear model-based algorithm for detection of outliers in environmental data and comparison with semi-parametric outlier detection methods. *Atmospheric Pollution Research*, 10(4), 1015-1023.

[4] Cook, R.D. (1979). Influential observations in linear regression, *Journal of the American Statistical Association*, 74, 1691-74.

[5] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, New York.

[6] Daneshgar, A., Javadi, R. and Razavi, S.S. (2013). Clustering and outlier detection using isoperimetric number of trees. *Pattern Recognition*, 46(12), 3371-3382.

[7] Ester, M., Kriegel, H.P., Sander, J. and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *In KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 96, 226–231.

[8] Gan, G., Ma, C. and Wu, J. (2020). *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia, PA, USA SIAM Press.

[9] Hadi, A.S. and Simonoff, J.S. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88, 1264-1272.

[10] Hahsler, M., Piekenbrock, M. and Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, 91(1), 1-30.

[11] Hastie, T., Tibshirani, R. and Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

[12] Huang, J., Zhu, Q., Yang, L., Cheng, D. and Wu, Q. (2017). A novel outlier cluster detection algorithm without top-n parameter. *Knowledge-Based Systems*, 121, 32-40.

[13] Huber, P.J. (1977). Robust covariances. *In Statistical Decision Theory and Related Topics*, 165-191.

[14] Kima, S.S., Parkb, S.H. and Krzanowskic, W.J. (1974). Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model. *Journal of Applied Statistics*, 35(3), 283–291.

[15] Montgomery, D.C. and Peck, E.A. (1992). *Introduction to Linear Regression Analysis*. John Wiley & Sons, New York.

[16] Rao, C.R., Toutenburg, H. and Fieger, A. (1999). *Linear Models: Least Squares and Alternatives*. Second edition, Springer.

[17] Rencher, A.C. (2000). *Linear Models in Statistics*. John Wiley & Sons, New York.

[18] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.

[19] SAS Customer Support. http://support.sas.com/

[20] Taylan, P., Yerlikaya-Özkurt, F. and Weber, G.W. (2014). An approach to the mean shift outlier model by Tikhonov regularization and conic programming. *Intelligent Data Analysis*, 18(1), 79-94.

[21] Wang, Y.F., Jiong, Y., Su, G.P. and Qian, Y.R. (2019). A new outlier detection method based on OPTICS. *Sustainable Cities and Society*, 45, 197-212.

[22] Xia, J., Gao, L., Kong, K., Zhao, Y., Chen, Y., Kui, X. and Liang, Y. (2018). Exploring linear projections for revealing clusters, outliers, and trends in subsets of multi-dimensional datasets. *Journal of Visual Languages and Computing*, 48, 52-60.

[23] Xu, X., Liu, H., Li, L. and Yao, M. (2018). A comparison of outlier detection techniques for high-dimensional data. *International Journal of Computational Intelligence Systems*, 11(1), 652-662.

[24] Xu, R. and Wunsch, D. (2008). *Clustering*. John Wiley & Sons, New Jersey.

VOLUME/CİLT 14 NUMBER/SAYI 2 DECEMBER/ARALIK 2022

İSTATİSTİK

JOURNAL OF THE TURKISH STATISTICAL ASSOCIATION

TÜRK İSTATİSTİK DERNEĞİ DERGİSİ

ISSN: 1300-4077