ISSN 1645-6726

Instituto Nacional de Estatística
Statistics Portugal

REVSTAT
Statistical Journal

Volume 19, No.1
January 2021

Catalogação Recomendada

REVSTAT. Lisboa, 2003-
Revstat : statistical journal / ed. Instituto Nacional
de Estatística. - Vol. 1, 2003- . - Lisboa I.N.E.,
2003- . - 30 cm
Trimestral. - Continuação de : Revista de Estatística =
ISSN 0873-4275. - edição exclusivamente em inglês
ISSN 1645-6726 ; e-ISSN 2183-0371

CREDITS

- **EDITOR-IN-CHIEF**
  - *Isabel Fraga Alves*
- **CO-EDITOR**
  - *Giovani L. Silva*
- **ASSOCIATE EDITORS**
  - *Marília Antunes*
  - *Barry Arnold*
  - *Narayanaswamy Balakrishnan*
  - *Jan Beirlant*
  - *Graciela Boente*
  - *Paula Brito*
  - *Vanda Inácio de Carvalho*
  - *Valérie Chavez-Demoulin*
  - *David Conesa*
  - *Charmaine Dean*
  - *Fernanda Figueiredo*
  - *Jorge Milhazes Freitas*
  - *Alan Gelfand*
  - *Stéphane Girard*
  - *Marie Kratz*
  - *Victor Leiva*
  - *Artur Lemonte*
  - *Shuangzhe Liu*
  - *Maria Nazaré Mendes-Lopes*
  - *Fernando Moura*
  - *John Nolan*
  - *Paulo Eduardo Oliveira*
  - *Pedro Oliveira*
  - *Carlos Daniel Paulino (2019-2021)*
  - *Arthur Pewsey*
  - *Gilbert Saporta*
  - *Alexandra M. Schmidt*
  - *Julio Singer*
  - *Manuel Scotto*
  - *Lisete Sousa*
  - *Milan Stehlík*
  - *María Dolores Ugarte*
- **FORMER EDITOR-IN-CHIEF**
  - *M. Ivette Gomes*
- **FORMER CO-EDITOR**
  - *M. Antónia Amaral Turkman*
- **EXECUTIVE EDITOR**
  - *José A. Pinto Martins*
- **FORMER EXECUTIVE EDITOR**
  - *Maria José Carrilho*
  - *Ferreira da Cunha*
- **SECRETARIAT**
  - *José Cordeiro*
  - *Olga Bessa Mendes*
- **PUBLISHER**
  - *Instituto Nacional de Estatística, I.P. (INE, I.P.) Web site: http://www.ine.pt*
- **COVER DESIGN**
  - *Mário Bouçadas, designed on the stain glass window at INE by the painter Abel Manta*
- **LAYOUT AND GRAPHIC DESIGN**
  - *Carlos Perpétuo*
- **PRINTING**
  - *Instituto Nacional de Estatística, I.P.*
- **EDITION**
  - *140 copies*
- **LEGAL DEPOSIT REGISTRATION**
  - *N.º 191915/03*
- **PRICE [VAT included]**
  - *€ 9,00*

INDEX

Abstracted/indexed in: *Current Index to Statistics, Google Scholar, Mathematical Reviews, Science Citation Index Expanded*®*, Zentralblatt für Mathematic, Scimago Journal & Country Rank* and *SCOPUS.*

REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 1–22

SKEWED PROBIT REGRESSION — IDENTIFIABILITY, CONTRACTION AND REFORMULATION

Authors: Janet van Niekerk
– CEMSE Division, King Abdullah University of Science and Technology,
Kingdom of Saudi Arabia
Janet.vanNiekerk@kaust.edu.sa

Håvard Rue
– CEMSE Division, King Abdullah University of Science and Technology,
Kingdom of Saudi Arabia
Haavard.Rue@kaust.edu.sa

Received: September 2020 Revised: October 2020 Accepted: November 2020

Abstract:

- Skewed probit regression is but one example of a statistical model that generalizes a simpler model, like probit regression. All skew-symmetric distributions and link functions arise from symmetric distributions by incorporating a skewness parameter through some skewing mechanism. In this work we address some fundamental issues in skewed probit regression, and more genreally skew-symmetric distributions or skew-symmetric link functions.
  We address the issue of identifiability of the skewed probit model parameters by reformulating the intercept from first principles. A new standardization of the skew link function is given to provide and anchored interpretation of the inference. Possible skewness parameters are investigated and the penalizing complexity priors of these are derived. This prior is invariant under reparameterization of the skewness parameter and quantifies the contraction of the skewed probit model to the probit model.
  The proposed results are available in the *R-INLA* package and we illustrate the use and effects of this work using simulated data, and well-known datasets using the link as well as the likelihood.

Keywords:

- *skew symmetric; probit; binary regression; penalizing complexity; INLA.*

AMS Subject Classification:

- 49A05, 78B26.

1. INTRODUCTION

Skew-symmetric distributions have acclaimed fame due to their ability to model skewed data, by introducing a skewness parameter to a symmetric distribution, through some skewing mechanism. In the preceding decades, an abundance of skewed distributions has been proposed from the basis of symmetric distributions, like the skew-normal [30, 3], skew-t [6] and more generally skew-elliptical distributions [21]. In each of these skew distributions, an additional parameter is introduced that indicates the direction of skewness or alternatively, symmetry.

With the introduction of the additional parameter, the inferential problem can become more challenging. The identifiability of the parameters and the existence of the maximum likelihood estimators (MLEs) are issues to keep in mind. In the Bayesian paradigm, the choice of a prior for the skewness parameter emerges. Either way, the inference of the skewness parameter is crucial in evaluating the appropriateness of the underlying (skewed) model.

A continuous random variable X, follows a skew-normal (SN) distribution with location, scale and skewness(shape) parameters ξ, ω and α, respectively, if the probability density function (pdf) is as follows:

$$g(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left[\alpha\left(\frac{x-\xi}{\omega}\right)\right], \tag{1.1}$$

where $\alpha \in \mathbb{R}$, $\omega > 0$, $\xi \in \mathbb{R}$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution function (CDF) of the standard Gaussian distribution, respectively. Denote by $G(x)$ the CDF of the skew-normal density.

The parameterisation in (1.1) poses difficulties since the mean and variance depends on α, as $E[X] = \xi + \omega\delta\sqrt{2/\pi}$ and $V[X] = \omega^2\left(1 - 2\delta^2/\pi\right)$, where $\delta = \alpha/\sqrt{1+\alpha^2}$. This implies that inference for α will also influence the inference for the mean and variance, since both are functions of α.

A similar challenge arises in the binary regression framework where the skew-normal link function is used as a generalization of probit regression, namely skewed probit regression. The need for asymmetric link functions have been noted by [14]. In binary regression, asymmetric link functions are essential in cases where the probability of a particular binary response approaches zero and one at different rates. In this case, a symmetric link function will result in substantially biased estimators with over(under)estimation of the mean probability of the binary response, due to the different rates of approaching zero and one (see [16] for more details on this issue). Skewed probit regression is an extension of probit regression, where covariates are transformed through the skew-normal CDF instead of the standard normal CDF.

Here, it might not be intuitive when the skewed link function is more appropriate than the symmetric link function. The estimate of the skewness parameter could provide some insights into this, only if the inference of the skewness parameter is reliable and interpretable.

Regarding the inference of the skewness parameter, α in (1.1), being it in the skewed probit regression or the skew-normal distribution as the underlying response model (which are conceptually the same estimation setup), various works have been contributed, most of them dedicated to the skew-normal response model framework. The identifiability of the parameters in the skew-normal response model was investigated by [22] (and skew-elliptical in general), [31] (for finite mixtures) and [13] (for extensions of the skew-normal distributions). For binary regression, identifiability of the parameters was considered by [25] where some issues concerning identifiability were raised. We address the identifiability problem from a first principles viewpoint, so that the parameters are identifiable, even with weak covariates, hence adding to [25].

In the skew-normal response model, the bias of the MLEs is a well-known fact (see [34] for more details). For small and moderate sample sizes, the MLE of the skewness parameter could be infinite with positive probability and the profile likelihood function has a singularity as the skewness parameter approaches zero, as noted early on by [3] (see also [26]). Some approaches to alleviate this feature of the skew-normal likelihood function have been proposed, including reparameterization of the model by [3] using the mean and variance (instead of location and scale parameters), or using a Bayesian framework by [27] (default priors) and [7] (proper priors). Also, [34] used the work of [19] to propose an adjusted (penalized) score function for frequentist estimation of the skewness parameter. A penalized MLE approach for all the parameters, including the skewness parameter, is presented by [5]. Bias-reduction regimes were proposed by [28].

From a Bayesian viewpoint, various priors for the skewness parameter have been proposed such as the Jeffrey's prior [27], truncated Gaussian prior [1], Student t prior and approximate Jeffery's prior [7], uniform prior [2], probability matching prior [11], informative Gaussian and unified skew-normal priors [12] and the beta-total variation prior [17]. All of these Bayesian approaches, with the exception of the latter, are based on somewhat arbitrary prior choices for mainly mathematical or computational convenience. These priors (as many others) are not invariant under reparameterization of the skewness parameter. The beta-total variation prior presented by [17] is based on the total variation from the symmetric Gaussian model to the skew-normal model, viewing the skewness parameter as a measure of perturbation. This prior is indeed invariant under one-to-one transformation of the skewness parameter.

Amongst the many works on the skew-normal response model, it seems that the genesis of the skew-normal model has been neglected. The skew-normal model was introduced by [3] as an (asymmetric) extension of the Gaussian model. The motivation for this extension is found in data. When data behaves like the Gaussian model, but the profile of the density is asymmetric, the skew-normal model might be appropriate. Conversely, we need an inferential framework wherein the skew-normal model would contract (or reduce) to the Gaussian model, in the absence of sufficient evidence of non-trivial skewness. The priors mentioned before do not provide a quantification framework with which the modeler can understand, and subsequently control this contraction. To achieve this, we need to consider the model (either skewed probit regression or the skew-normal response model) from an information theoretic perspective. Then we can construct a prior with which the quantification of contraction (or not) can be done, and used as a translation of prior information from the modeler to the model.

In this paper we address some issues (identifiability, standardizing, skewness parameters) prevalent in skewed-probit regression in Section 2 and construct the penalized complexity

(PC) prior for the skewness parameter of the link function (which is translatable to the skew-normal response model) in Section 4. This PC prior is implemented in the *R-INLA* [32] (see also [33], [29]) package for general use by others. We use a numerical study to illustrate the solutions proposed in Section 2 and apply the PC prior to simulated and real data in Sections 5 and Section 6. The paper is concluded by a discussion in Section 7 in which we sketch the wider applicability of this work and contributions to the wider skew-symmetric family.

2. SKEWED PROBIT REGRESSION AND ISSUES

We consider skewed probit regression as an extension of probit regression, where the link function is the skew-normal CDF instead of the standard normal CDF. We formulate skewed probit regression that can include random effects like spline functions of the covariates, spatial and/or temporal effects. For this paper, we assume the following structure. From a sample of size n, the responses $\boldsymbol{y}_{n\times 1}$ are counts of successful trials out of $N_{n\times 1}$ trials and hence we assume a Binomial distribution with success probability p. We gather all m covariates in $\boldsymbol{X}_{n\times m}$ and use these to build an additive linear predictor, defined as $\boldsymbol{\eta}_{n\times 1}$. So then,

$$\begin{aligned} y_i &\sim \text{Binomial}(N_i, p_i), \\ p_i &= G(\eta_i), \quad i = 1, ..., n, \end{aligned} \tag{2.1}$$

where $G(\cdot)$ is the CDF of the Skew-Normal that depends on (ξ, ω, α). The linear predictor η_i is an additive linear predictor defined as follows,

$$\eta_i = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \sum_{k=1}^{K} f^k(\boldsymbol{Z}_i), \tag{2.2}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the covariates for the fixed and random effects, respectively, the functions $\{f^k(.)\}$ are random effects like spatial, spline, temporal effects with hyperparameters $\boldsymbol{\theta}$.

2.1. Issue 1 – Standardizing the link function

With the aim of standardizing the link function, [25] assumed $\xi = 0, \omega = 1$, similar to [9] and many others. Initially, the idea behind this choice feels intuitive since the skew probit link is an extension of the probit link through the skewness parameter. However, the $(0, 1)$ parameter values of the probit link should not be naively copied to the skewed probit link. The choice, $\xi = 0, \omega = 1$ implicitly concedes that a skew-normal density (1.1) with mean

$$E[X] = \alpha\sqrt{\frac{2}{\pi(1+\alpha^2)}},$$

and variance

$$V[X] = 1 - \frac{2\alpha^2}{\pi(1+\alpha^2)},$$

is used to calculate the probability of success, for all α. This essentially implies that for different skewness parameter values, different means and variances are used. This way of standardizing is a parameter-based method, instead of the intended property-based method like in the probit link. We do not expect the assumption $\xi = 0, \omega = 1$ to work well since the mean and variance are not anchored and can attain many values based on different values of α.

We posit that the mean and the variance (properties of the link) should be fixed, like in the probit case, instead of the skew-normal location and scale parameters. This is analogous to the idea of the centered parametrization of the skew-normal density and mentioned by [8].

We propose the link function $F(y|\alpha)$ that is the CDF of the Skew-Normal density (1.1) scaled to have zero mean and unit variance for all values of α. That is,

$$F(y|\alpha) = \int_{-\infty}^{y} f(x|\alpha)\,dx$$

where

$$f(x|\alpha) = \frac{2}{\omega(\alpha)}\phi\left(\frac{x-\xi(\alpha)}{\omega(\alpha)}\right)\Phi\left[\alpha\left(\frac{x-\xi(\alpha)}{\omega(\alpha)}\right)\right], \tag{2.3}$$

$$\xi(\alpha) = -\omega(\alpha)\sqrt{\frac{2}{\pi(1+\alpha^2)}},$$

and

$$\omega(\alpha) = \sqrt{\left(1 - \frac{2\alpha^2}{\pi(1+\alpha^2)}\right)^{-1}}.$$

This provides an anchored link function with zero mean and unit variance, for all α. If this standardization is not used then an arbitrary unknown scale is introduced to the model, with no means of recovering it. By fixing the mean and variance, we have a better understanding of the properties of the link and we do approach the probit case in the neighborhood of $\alpha = 0$.

2.2. Issue 2 – The quantile intercept and identifiability of parameters

The identifiability of the parameters in skewed probit regression were first investigated by [25]. They showed that without the presence of a continuous covariate, the intercept β_0, and skewness parameters are not identifiable. This is expected due to the traditional definition of the skewed probit model (2.1) and (2.2). We rectify the formulation of the skewed probit regression intercept, by introducing the quantile intercept, and subsequently solve this issue of non-identifiability by returning to first principles.

In simple linear regression, the intercept is used to calculate the expected value of the linear predictor without any effect from covariates. In probit regression, the intercept contains information about the probability of the event, without the effects from covariates. The value of the intercept should not provide any information about the other parameters in the model.

However, when we introduce a skewness parameter to a symmetric family to formulate a skew-symmetric link then we are fundamentally changing the meaning of what is traditionally called the intercept of the linear predictor, i.e. β_0 in (2.2).

Consider probit regression with one centered covariate X,

$$p = \text{Prob}[Y = 1] = \Phi(\beta_0 + \beta_1 X).$$

Now if $\beta_1 X = 0$, then

$$q = \text{Prob}[Y = 1] = \Phi(\beta_0),$$

which implies that β_0 is the q^{th} quantile of the standard Gaussian distribution. There is thus a one-to-one relationship between q and β_0. When $\beta_1 \neq 0$, then $\text{Prob}[Y = 1]$ changes because of $\beta_1 X$, without affecting β_0, because Φ remains the same function. In this sense, β_0 is uninformative for β_1.

Conversely, consider skewed-probit regression from (2.1) and (2.3),

$$p = \text{Prob}[Y = 1] = F(\beta_0 + \beta_1 X|\alpha).$$

Here, β_0 should, in the same way, be uninformative for β_1. This does not hold because the dependence of α. We can ensure this, if

$$q = \text{Prob}[Y = 1] = F(\beta_0|\alpha)$$

is constant for varying α, which is the case if β_0 is defined as the q^{th} quantile of the distribution with CDF F. Therefore, we reformulate β_0 as

$$\beta_0(q, \alpha) = F^{-1}(q|\alpha), \tag{2.4}$$

so β_0 is the q^{th} quantile of $F(.|\alpha)$. The quantile level q is now the unknown intercept-parameter instead of β_0.

Note that there is (generally) not a one-to-one relationship between β_0 and q since the q^{th} quantile depends on α. In this new formulation, the intercept as defined implicitly by q, provides no information about β_1 and parameters of $F(\eta_i|\alpha)$ are identifiable. We return in 5.3 to a numerical study of this issue.

This formulation might seem surprising at first sight, but in the case of a symmetric link, the intercept is the quantile of a distribution with fixed (no) skewness. In the case of the probit or identity links for example, this formulation will reduce to the usual intercept parameter since in these cases there is a one-to-one relationship between β_0 and q.

In terms of implementation in *R-INLA*, the new formulation of the skew normal model in terms of q is available and subsequently, the prior distribution for q can be derived from a corresponding informative $N(\mu_0, \tau_0)$ prior for β_0 in the case where $\alpha = 0$. This will ensure that the probit and the skewed-probit models have comparable priors for their respective "intercept" parameters.

2.3. Issue 3 – Skewness-related parameters

It is well-known that the skew-normal likelihood has a (double) singularity in the neighbourhood $\alpha \simeq 0$ [3]. Various adaptations of maximum likelihood estimation and some Bayes estimators have been proposed as solutions to this singularity. [23] used the Fisher information to propose a reparameterization that uses α^3 as the skewness parameter since this solves the double singularity problem in the likelihood. In our venture to derive the PC prior for the skewness, we derived the Kullback-Leibler divergence (KLD) from the skew-normal link to the probit link and noticed the same feature as mentioned in [23]. This resemblance is expected since the Fisher information metric is the Hessian of the KLD.

From (2.3), the KLD for small $|\alpha|$ can be found to be

$$\begin{aligned}
\mathrm{KLD}(\alpha) &= \int f(x|\alpha)\log\frac{f(x|\alpha)}{f(x|\alpha=0)}dx \\
&= \frac{\pi^2+16-8\pi}{6\pi^3}\alpha^6 - \frac{144\pi+3\pi^3-38\pi^2-168}{6\pi^4}\alpha^8 \\
&\quad + \frac{-42240\pi-2560\pi^3+16176\pi^2+129\pi^4+39936}{120\pi^5}\alpha^{10} + \mathcal{O}(\alpha^{12}) \\
&\approx c_1\alpha^6 + c_2\alpha^8 + c_3\alpha^{10}. \qquad (2.5)
\end{aligned}$$

Interestingly, the behavior of α around $\alpha = 0$ does not have the usual asymptotics (consistency rate of $\sqrt{n}$) since the leading term is α^6. This implies that the estimator of α in the neighbourhood $\alpha \simeq 0$, has a consistency rate $n^{\frac{1}{6}}$ but a skewness parameter $\gamma = \alpha^3$, such that $\alpha = \mathrm{sign}(\gamma)\sqrt[3]{|\gamma|}$, will have the normal asymptotics in the sense that the estimator of γ will be $\sqrt{n}$ consistent.

Even though γ has the usual asymptotic behaviour, the estimate of it is hard to interperate since it does not relate easily to an interpretable property. We can instead focus on the more intepretable (standarised) skewness of the skew-normal distribution, γ_1, which is a monotone function of γ

$$\gamma_1 = \frac{(4-\pi)\left(\sqrt{\frac{2\delta^2}{\pi}}\right)^3}{2(1-\frac{2\delta^2}{\pi})^{\frac{3}{2}}}, \qquad (2.6)$$

where $\delta = \frac{\alpha}{\sqrt{1+\alpha^2}}$ (and $\gamma = \alpha^3$). The skewness takes values in the interval $-0.99527 < \gamma_1 < 0.99527$, which is correct up to five digits.

The question arises if we should formulate a prior for α, γ or the skewness γ_1. If priors are assigned more ad-hoc parameters, this question poses a challenge. The PC prior is invariant under reparameterizations [35], implying that this framework will produce equivalent priors for α, γ and γ_1. They are equivalent in the inferential sense, and will produce the same posterior inference.

3. SKEW-NORMAL MEAN REGRESSION

In this section we focus on skew-normal regression, although these issues also exist in more general skew-symmetric regression models.

In the preceeding section we mentioned the different parameters that can be used to capture the skewness in the skewed probit model, and the proposals pertain to the skew-normal regression model as well.

Most works on skew-normal regression propose a regression model for the location parameter, ξ, from (1.1). This generalization of Gaussian regression seems straightforward but when we keep in mind that the location parameter of the Gaussian is equal to the mean, then we can see that regressing through the location parameter of the skew-normal is not practical. In the spirit of generalizing Gaussian regression to skew-normal regression, we should formulate the regression model based on the mean. Hence for $y_i \sim SN(\xi, \omega, \alpha)$ from (1.1),

$$E[Y_i] = \eta_i, \tag{3.1}$$

with η_i from (2.2), instead of $\xi_i = \eta_i$. Note that here we do not reformulate the intercept as in Section 2.2 for skewed probit regression, since the identity link function is used. We illustrate the proposed skew-normal regression model in Section 6.

4. PENALIZING COMPLEXITY PRIOR FOR THE SKEWNESS PARAMETER

The work of [35] introduced the notion of penalizing complexity priors for parameters and provided the framework for deriving priors that quantify the contraction from a complex model to a simpler model. These PC priors are especially helpful and very needed in cases where priors have traditionally been chosen due to mathematical convenience, or convention (see [24] for more details on the performance of PC priors). PC priors have been used in various fields of research, for example [36] derived the PC priors for autoregressive models while [20] derived PC priors for Gaussian random fields.

In this section we derive the PC prior for α due to the invariance of the PC prior under reparameterization of the skewness parameter. The derivations of the PC prior for γ and γ_1 follows then directly from a change-of-variable exercise.

Using [35] and (2.5), define the uni-directional distance from the skew-normal to the Gaussian density as,

$$\begin{aligned} d(\alpha) &= \sqrt{2\text{KLD}(\alpha)} \\ &\approx \sqrt{2(c_1\alpha^6 + c_2\alpha^8 + c_3\alpha^{10})}\,. \end{aligned} \tag{4.1}$$

The penalizing complexity prior for the skewness parameter α is then formed by assigning an exponential prior with parameter θ to the distance. The parameter θ incorporates information from the user to control the tail behavior and thus the rate of contraction towards the probit link function. The penalizing complexity prior follows then directly, as

$$\begin{aligned} \pi(\alpha) &= \frac{1}{2}\theta \exp\left[-\theta d(\alpha)\right] \left|\frac{\partial d(\alpha)}{\partial \alpha}\right| \\ &\approx \frac{\theta}{2\sqrt{2(c_1\alpha^6 + c_2\alpha^8 + c_3\alpha^{10})}} \left|2(6c_1\alpha^5 + 8c_2\alpha^7 + 10c_3\alpha^9)\right| \\ &\quad \times \exp\left[-\theta|\alpha^3|\sqrt{2(c_1 + c_2\alpha^2 + c_3\alpha^4)}\right] \end{aligned} \tag{4.2}$$

for small values of $|\alpha|$. The user-defined parameter θ is used to govern the contraction towards probit regression, e.g., for small $p_U > 0$,

$$\text{Prob}(d(\alpha) > U) \;=\; p_U = \exp(-\theta U)$$

which gives $\theta = -\log p_U / U$. There is no explicit expression for the penalizing complexity prior of α in general, but the prior can be computed numerically. The prior for γ_1 is available in the *R-INLA* package [32] with `prior = "pc.sn"` and parameter `param =` θ. We use the γ_1 reparameterization, since γ_1 quantifies the skewness as a *property* with good interpretation.

The PC priors of α and γ_1 are illustrated in Figure 1 for $\theta = 5$, on the α and γ_1 scales. In Figure 2 various values for θ are considered to provide an intuition about the effect of θ. From this Figure it is clear that larger values of θ results in higher contraction rates with little mass away from 0. The posterior inference of the skewness is not sensitive to the value of θ for moderate and large samples. In the case of small samples, a very large value of θ will contract the Bayes estimator towards 0 at a fast rate.

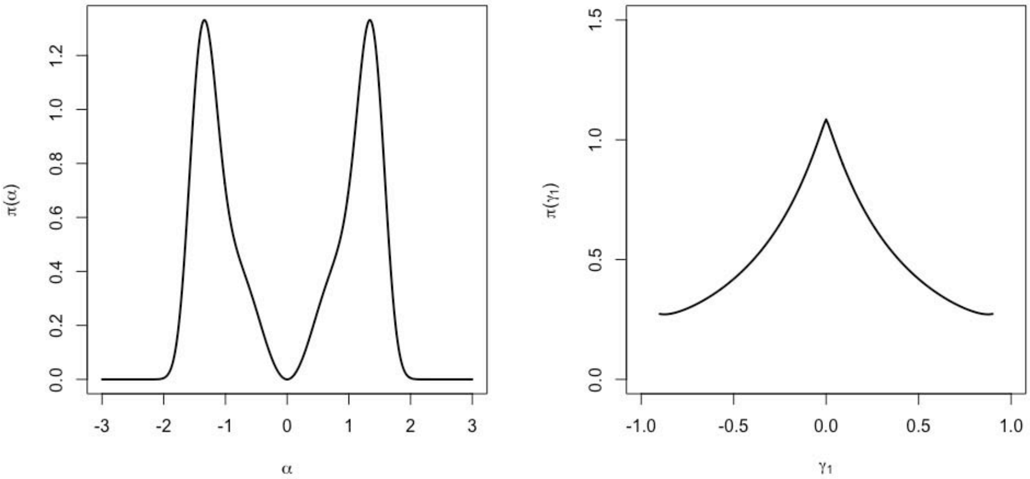


Figure 1: PC prior (4.2) for $\theta = 5$ on the α scale (left) and the γ_1 scale (right).

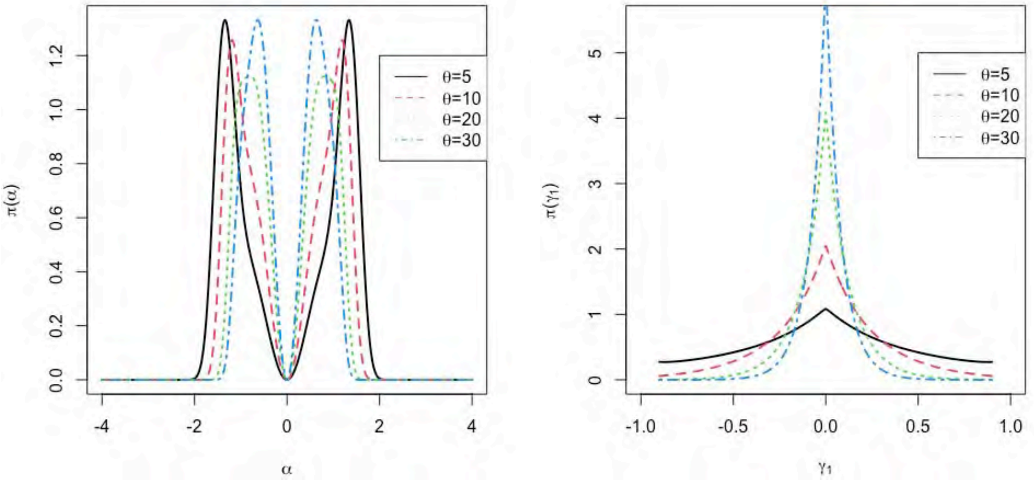


Figure 2: PC prior (4.2) for various θ's on the α scale (left) and the γ_1 scale (right).

From Figure 1 we can see the shape of the PC prior for α is quite peculiar, but has a clear interpretation in terms of a prior on the distance. It just shows that if we assign priors to parameters, like α, instead of to a property, like γ_1, it is highly improbable that we could think of a density function for the parameter that has good translatable properties. Another interesting note is that from the prior density of α around $\alpha = 0$, we can see that most priors of α proposed in literature actually results in underfitting, instead of the usual overfitting, since they assign too much density to the neighborhood around $\alpha = 0$. Conversely, the PC prior of γ_1 is as expected with a mode at the value for the probit link.

5. SIMULATION STUDY

In this section we present condensed results from a simulation study with the aim to show the results proposed in this work for experiments with a large and small number of trials. The setup is to simulate linear predictors $\eta_i = \beta_0(\alpha, q) + \beta_1 x_i$, where $x_i \sim \mathrm{N}(0, 0.5)$ for $i = 1, ..., n$. The success probabilities are then $p_i = F(\eta_i|\alpha)$ from (2.1) and subsequently the response variable y_i, wherere $y_i \sim \mathrm{Bin}(N_i, p_i)$. To investigate the performance of the PC prior for the skewness, we consider the PC prior as well as a weak Gaussian prior. Throughout this simulation study, we assume $\theta = 5$ for the PC prior and a weak Gaussian prior with parameters $(0, 10^2)$ for the skewness.

5.1. Large number of trials

For an experiment that consists of a large number of trials, we consider four simulation scenario's which can be summarized as:

1. $q = \frac{1}{3}, \beta_1 = 1, \gamma_1 = 0(\alpha = 0), N_i = 200;$

2. $q = 0.25, \beta_1 = -1, \gamma_1 = \frac{2}{3}(\alpha = 10), N_i = 200;$

3. $q = 0.30, \beta_1 = 1, \gamma_1 = \frac{1}{3}(\alpha = 2), N_i = 200;$

4. $q = 0.10, \beta_1 = -1, \gamma_1 = -\frac{1}{3}(\alpha = -2), N_i = 200.$

In each case we consider the PC prior as well as the Gaussian prior for the skewness γ_1, and weakly informative Gaussian priors for the fixed effects.

5.1.1. Results

The fixed effects were recovered well and here we focus on the skewness γ_1. From Table 1 it is clear that the PC prior (and the Gaussian prior) performs as expected since the sample size and number of trials are large. In Figure 3 the posterior results for the skewness are summarised with coverage probability and median length of the credible interval. The results for other scenarios are similar and omitted here. From this (and many other) simulation studies, we conclude that for a large number of trials the skewed-probit link works well and the inference is accurate. It is clear that the PC prior does not contract towards the probit model when the data presents strong support for the skewed probit model (scenarios 2, 3 and 4).

Table 1: Coverage probability (CP) and median length of the credible interval (MLCI) for the skewness γ_1 under the PC and Gaussian (G) priors, for large N_i.

Scenario	**PC prior**		**Gaussian prior**	
	CP	**MLCI**	**CP**	**MLCI**
1	95	0.28	94	0.35
2	96	0.28	97	0.34
3	95	0.31	95	0.34
4	95	0.32	95	0.35

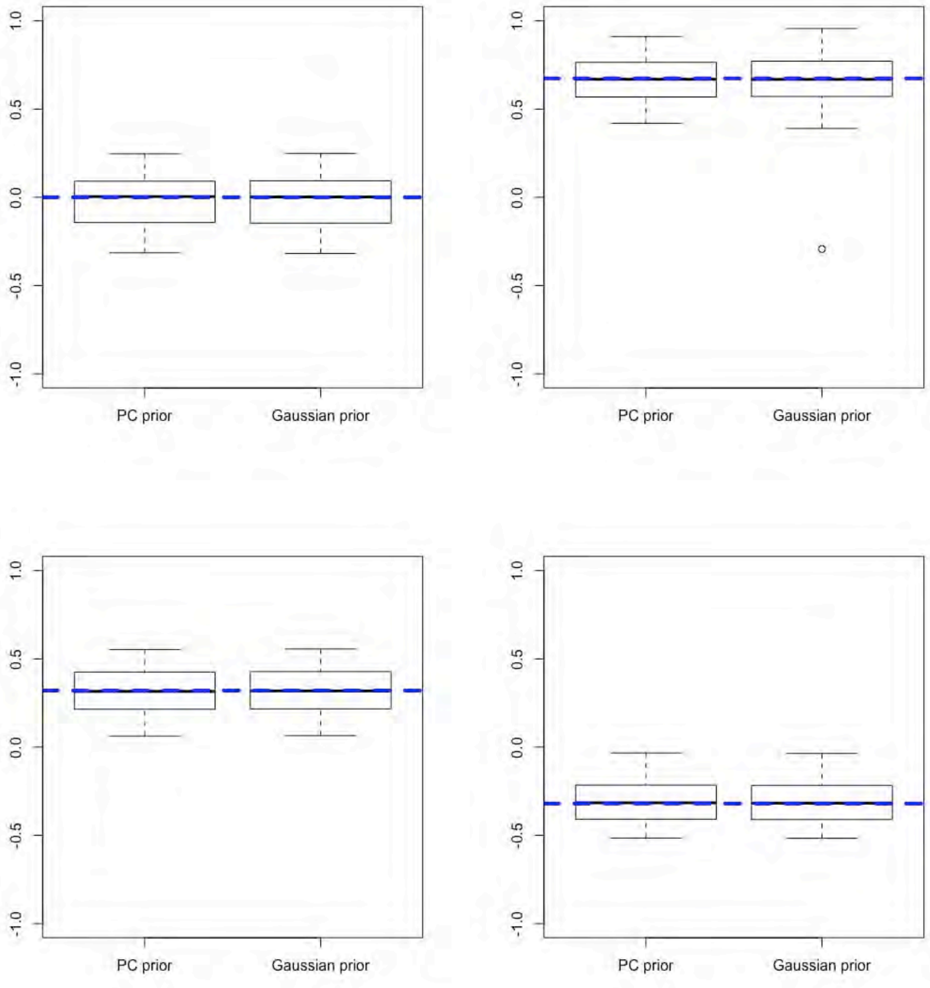


Figure 3: Median of 95% credible intervals for the different scenario's with the true skewness (dashed line): Scenario 1, 2 (top left to tight), 3 and 4 (bottom left to right).

5.2. Small number of trials

Here we focus our attention on samples of size 200 of binary trials, and the scenario's we consider are:

1. $q = \frac{1}{2}, \beta_1 = 1, \gamma_1 = -\frac{2}{3}(\alpha = -10), N_i = 1;$
2. $q = \frac{1}{2}, \beta_1 = 1, \gamma_1 = 0(\alpha = 0), N_i = 1.$

We consider the PC prior as well as the Gaussian prior for the skewness parameter, and weakly informative Gaussian priors for the fixed effects.

5.2.1. Results

From Table 2 it is clear that the skewness is not recovered well for a small number of trials. In the case of the PC prior, the coverage is poor but the credible intervals are still relatively narrow. For the Gaussian prior, the coverage is high mainly due to the very wide credible intervals. For a small number of trials or binary trials, the skewness is hard to capture.

Table 2: Coverage probability (CP) and median length of the credible interval (MLCI) for the skewness γ_1 under the PC and Gaussian (G) priors, for small N_i.

Scenario	**PC prior**		**Gaussian prior**	
	CP	**MLCI**	**CP**	**MLCI**
1	65	0.41	90	1.24
2	95	0.33	90	1.45

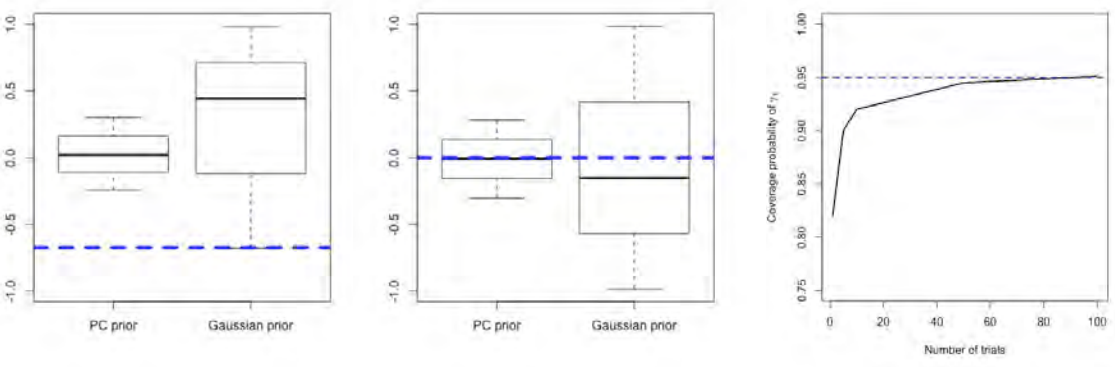


Figure 4: 95% credible intervals for γ_1 with $n_i = 1$ and $\gamma_1 = -\frac{2}{3}$ (left) or $\gamma_1 = 0$ (middle). Coverage probabilities for γ_1 under scenario 1 as N_i increases (right).

Even though the nominal coverage for the Gaussian prior is still high from Table 2, the median length of the credible interval implies that the credible intervals span most of the support of γ_1. However, the PC prior contracts to zero with relatively narrow credible intervals and exhibits poor coverage for $\gamma_1 \neq 0$. It is evident that the skewness is hard to estimate with a small number of trials. This is not unexpected since in binary data, we only observe a success or failure for each subject and subsequently the data does not provide sufficient information about the skewness. We need repetitions in the data to learn more about the skewness. We can see in Figure 4 that the PC prior contracts to zero if there is not enough evidence for the skewed link, but the Gaussian prior proposes an arbitrary value for the skewness from most of the range of γ_1 (possibly with the wrong sign as in Figure 4). In this case, using the skewed-probit link for binary data might not be useful.

5.3. Confounding and the effect of the quantile intercept

In this section we look at the effect of not using the new quantile intercept. We used a simulated dataset, similar to the preceeding section, with $q = 0.4, \beta_1 = 0.1, \gamma_1 = -\frac{2}{3}$. In this setup the linear predictor is close to zero, for a centered covariate, the confounding between the classical intercept and the skewness parameter is clear. In Figure 5 the median of the 95% credible intervals of the skewness (for 500 repetitions) as well as the true value of the skewness are presented. On the left we have the case of the quantile intercept and on the right, the classical intercept. By using the classical intercept, as in the case of GLM, the skewness is not estimated correctly in the sense that the direction is not even recovered. It is clear that the quantile intercept solves the confounding of the intercept of the linear predictor, with the skewness of the link.

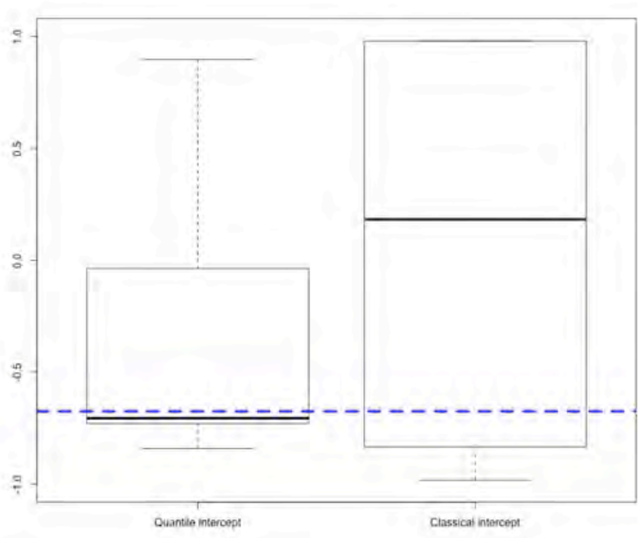


Figure 5: Median credible intervals for the skewness γ_1 using the quantile intercept vs the classical intercept.

6. APPLICATIONS

In this section we illustrate the use of skewed probit regression with the PC prior using two well-known datasets, the beetle mortality data [10] (binomial response with multiple trails) and the UCI Cleveland heart disease data [18] (Bernoulli response). We also present the analysis of the Wines data to illustrate the use of this work in the skew-normal likelihood.

6.1. Beetle mortality data

In this well-known dataset from [15] the number of adult flour beetles killed by differing dosages of poison is modelled based on the centered dosage value. We use the proposed skewed probit model with the PC prior and the quantile intercept. We also fit a probit model and compare the fitted values of both with the observed data. These, together with the 95% credible intervals are presented in Figure 6. We note that the skewed probit model seem to fit the observed data better than the probit model, and the 95% credible interval for the skewness of the skewed probit model from Table 3 does not include 0. The marginal log-likelihood for the skewed probit model is -21.75 versus -23.93 from the probit model. The difference between the marginal log-likelihoods does not provide a convincing argument in favor of the skewed probit model, as opposed to the probit model.

Table 3: Posterior estimates for the beetle mortality data.

Effect	**Estimate**	**95% credible interval**
Quantile of the intercept (q)	0.643	$(0.572; 0.703)$
Dosage	19.132	$(16.074; 22.316)$
Skewness (γ_1)	-0.456	$(-0.848; -0.053)$

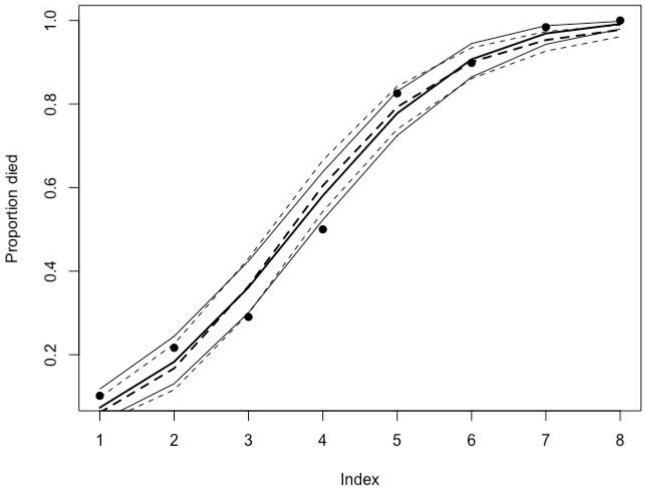


Figure 6: Fitted and observed proportions (– Skewed Probit, -- Probit) with 95% credible intervals.

6.2. Heart disease data

We will use the Cleveland data obtained by Robert Detrano from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

The response is a binary observation indicating the occurrence of a $> 50\%$ diameter narrowing in an angiography. Various covariates are available in this data and we will use a subset of these namely, gender (male/female), type of chest pain (1 – typical angina, 2 – atypical angina, 3 – non-anginal pain, 4 – asymptomatic), resting blood pressure, the slope of the peak exercise ST segment (1 – upsloping, 2 – flat, 3 – down sloping), the number of colored vessels by fluoroscopy and the results from the thallium heart scan (3 – normal, 6 – fixed defect, 7 – reversable defect). We centered the two continuous covariates, resting heart rate and the number of colored vessels by fluoroscopy. Further details can be found in [18].

There are 297 subjects with complete information in the dataset of which 137 experienced the event of $> 50\%$ diameter narrowing in an angiography. We fit a skewed-probit regression model to explain the probability of the event based on the values of the covariates similar to [25]. In [25] divergent results were obtained based on different estimation frameworks, namely maximum likelihood estimation, bootstrap bias correction, Jeffrey's prior, generalized information matrix prior and Cauchy prior penalized frameworks. The inconsistent results could be attributed to the issues we mentioned in this paper, since all these estimation methods were developed for the skewed-probit regression model without the good standardization, based on the skewness parameter α and defined using the classical intercept.

Also, there is a lack of information on the skewness in binary data. The consequence is thus that various values of the skewness could be supported. This case is a prime example that illustrates the need for the PC prior of the skewness, so that we prefer zero skewness a priori (probit regression) and use the data to advocate for non-trivial skewness (skewed probit regression).

Here, we can use the PC prior (4.2) for the skewness and the quantile intercept from Section 2.2. All quantitative covariates are centered. The results are given in Table 4.

Table 4: Results for the Cleveland heart disease data.

	Posterior mean	**95% credible interval**
Quantile Intercept (q)	0.045	$(0.006; 0.184)$
Gender (male)	1.025	$(0.605; 1.461)$
Type of chest pain (2)	0.198	$(-0.538; 0.942)$
Type of chest pain (3)	-0.074	$(-0.732; 0.590)$
Type of chest pain (4)	1.288	$(0.673; 1.920)$
Resting heart rate	0.016	$(0.005; 0.027)$
Slope of the peak exercise (2)	1.027	$(0.637; 1.452)$
Slope of the peak exercise (3)	0.791	$(0.059; 1.540)$
Number of colored vessels	0.704	$(0.477; 0.945)$
Skewness (γ_1)	0.02	$(-0.214; 0.235)$

From the estimate of γ_1 in Table 4 we deduce that the skewness is not supported by the data and a probit regression model could be sufficient. We did the analysis using probit regression and the inference is very similar. This result of zero skewness coincides with the skewness estimates in [25] using the MLE, bootstrap correction, generalized information matrix and cauchy prior penalization approaches. The posterior densities (and prior densities in dashed) of the skewness, γ_1, and quantile intercept, q, are presented in Figure 7.

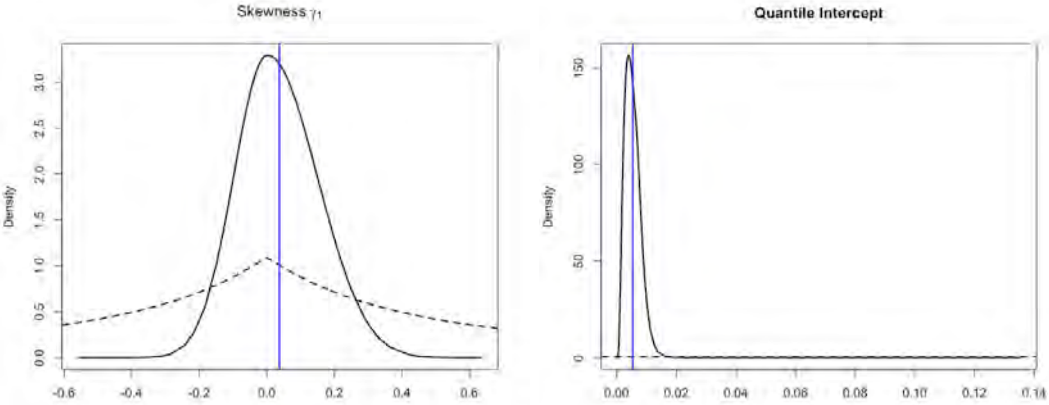


Figure 7: Posterior (prior – dashed) density of the skewness γ_1 (left) and quantile intercept q (right) with the corresponding point estimates (vertical line).

We also see that being a male, having asymptomatic chest pain, higher resting heart rate, a flat or downwards slope of the peak exercise ST segment and more colored vessels by fluoroscopy, all contribute to a higher probability of the event under investigation, i.e. $> 50\%$ diameter narrowing in an angiography.

The posterior densities (and prior densities in dashed) of the fixed effects are presented in Figure 8.

We calculated the marginal log-likelihoods for the probit and skewed-probit models to be -150.62 and -158.41, respectively, indicating that the probit model is preferred by the data. Both models achieved a correct classification percentage of 84.55%, on a 50% holdout sample.

6.3. Wines data

This section illustrates the new results when the response variable is continuous and assumed to follow a skew-normal distribution. As mentioned in Section 3, the results derived in this paper hold for skewed-probit models, as well as skew-normal regression models. We use the wines dataset from [4], where the acidity of the wine is assumed to follow a skew-normal distribution as illustrated in Figure 9, where we see the tail behaviour is correctly captured by the fitted Gaussian density, but not the skewness. The mean acidity (not the location parameter) is modelled using the type of wine, sugar content and pH level as covariates (after backwards elimination). We assign PC priors for the precision [35] as well as skewness (4.2).

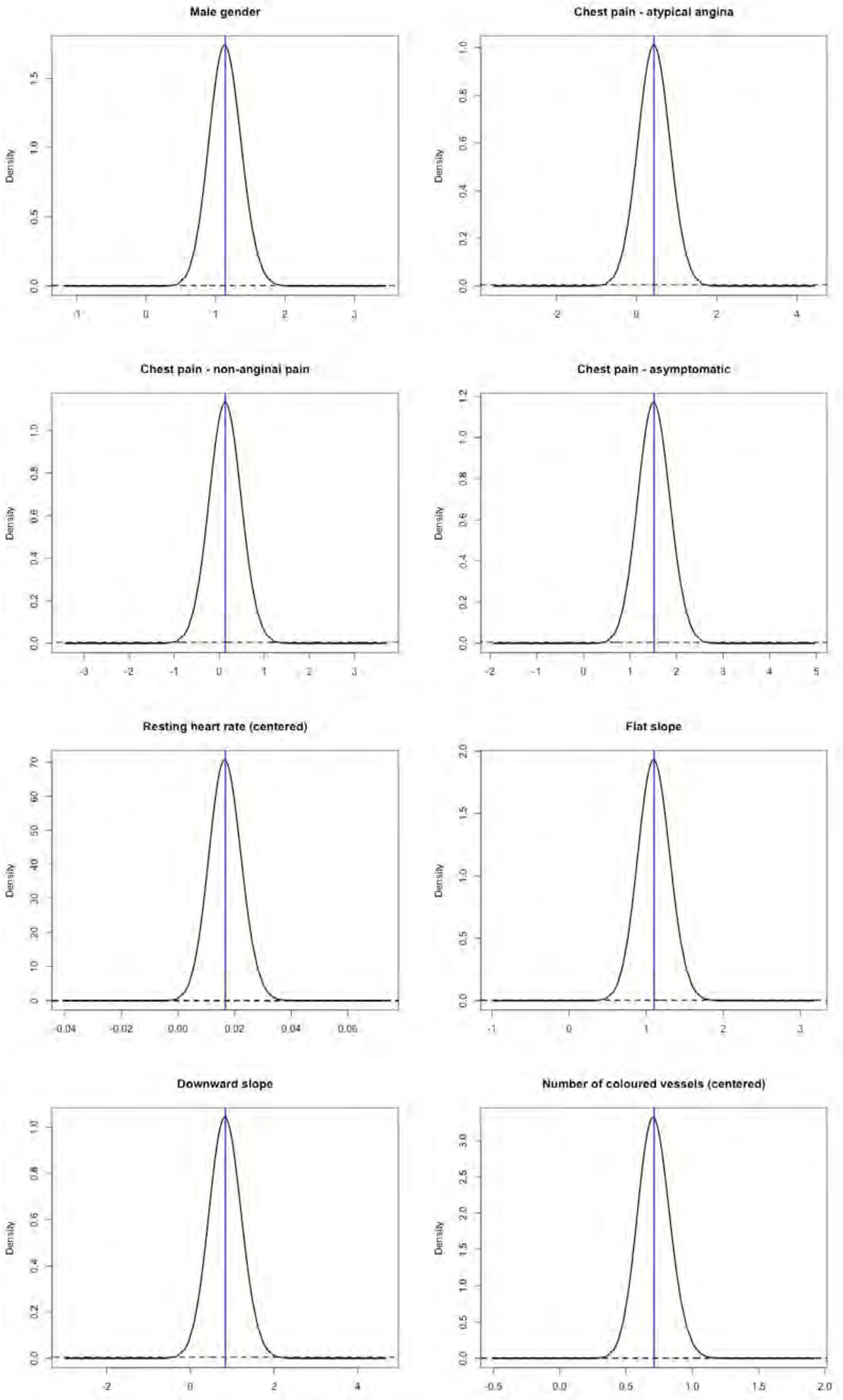


Figure 8: Posterior (prior – dashed) densities of the fixed effects with the corresponding point estimate (vertical line).

The results are given in Table 5. The marginal log-likelihood for the skew-normal model is -722.21 and for the Gaussian model it is -724.59.

Table 5: Results for the wines data.

	Posterior mean	**95% credible interval**
Intercept	77.053	$(73.824; 80.252)$
Wine (Grignolino)	5.088	$(0.478; 9.693)$
Wine (Barbera)	23.613	$(19.003; 28.280)$
Sugar	3.118	$(1.150; 5.080)$
pH	-8.350	$(-10.122; -6.574)$
Skewness (γ_1)	0.439	$(0.128; 0.702)$
Precision for the data	0.008	$(0.006; 0.009)$

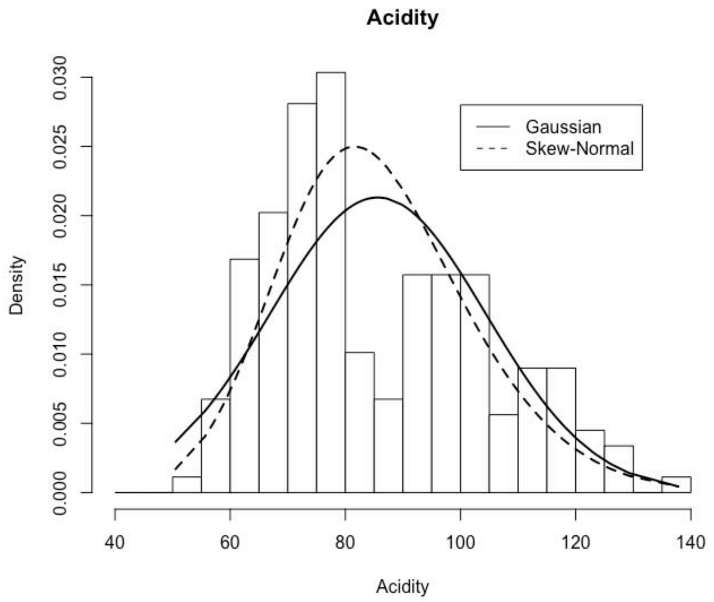


Figure 9: Histogram with model-based Gaussian curve and skew-normal curve.

7. DISCUSSION

The use of skew-symmetric distributions or links is popular due to the perceived flexibility inherited through the extra parameter that controls the skewness. The skew normal skewness parameter in particular, poses various challenges in the inference thereof. As we set out with the initial aim to derive the penalizing complexity prior for the skewness, we realized that there are various other issues that we could not found addressed in the literature. It is apparent that with the generalizing to skew-symmetric distributions and links from the symmetric counterparts, various fundamental concepts have gone amiss.

Here we rectify the formulation of the intercept in the linear predictor of all skew-symmetric links, firstly to ensure that it behaves as an intercept and secondly due to the confounding with the skewness parameter and fixed effects. We also show that the popular method of standardizing the skewed link function by inheriting the parameter values of the symmetric link, fundamentally changes the way the link function maps the data to the linear predictor, and we provide an anchored standardization approach. We believe that many of the contradicting works in this area can be attributed to the inappropriate use of the classical intercept and parameter-based standardization, instead of property-based standardization. In skew-symmetric regression models, we formulate the regression model based on the mean, instead of the location parameter.

After the fundamental corrections to the formulation of the skewed-probit link, the penalizing complexity prior for the skewness was derived. One particular advantage of this prior is that it is invariant to reparameterizations of the skewness parameter. In light of this, we implemented the PC prior for the skewness in *R-INLA* [32] for use by others. We noted, expectedly, that binary data (or with few trials) does not provide information about the skewness, and we thus advise against the use of the skewed-probit link for data with a small number of trials. We advocate the use of the PC prior even more feverently because of this feature, since the PC prior will contract to the simpler probit link instead of providing an incorrect unreliable estimate of the skewness. Other inferential frameworks might not be able to ensure this contraction in the absence of convincing evidence from the data about the necessary skewness, and could lead to unfounded complicated models.

We hope that the issues raised and addressed here will improve the inference of the skewed probit model (and more broadly the skew-symmetric links and likelihoods) and provide insights into the fundamental considerations necessary when distributions or links are generalized.

A. APPENDIX

We give here a small example for how to do skew probit regression in *R-INLA*. In the code below, the unusual statement is `remove.names="(Intercept)"` which remove the intercept in the formula *after* doing the expansion of factors in the model. We need this as we replace the traditional intercept with the quantile intercept in the link, and the expansion of factors depends on the presence or not, of an intercept in the model.

```
library(INLA)
n = 200
Ntrials = 200
x = rnorm(n, sd = 0.5)
eta = x
skew <- 0.5
prob = inla.link.invsn(eta, skew = skew, intercept = 0.75)
y = rbinom(n, size = Ntrials, prob = prob)
r = inla(y ~ 1 + x,
family = "binomial",
data = data.frame(y, x),
Ntrials = Ntrials,
control.fixed = list(remove.names = "(Intercept)",
 prec = 1),
control.family = list(
control.link = list(model = "sn",
hyper = list(
skew = list(param = 10)))))
summary(r)
```

REFERENCES

[1] Arellano-Valle, R.B.; Bolfarine, H. and Lachos, V.H. (2007). Bayesian inference for skew-normal linear mixed models, *Journal of Applied Statistics*, **34**(6), 663–682.

[2] Azevedo, C.L.N.; Bolfarine, H. and Andrade, D.F. (2011). Bayesian inference for a skew-normal IRT model under the centered parameterization, *Computational Statistics & Data Analysis*, **55**(1), 141–163.

[3] Azzalini, A. (1985). A class of distributions which includes the normal ones, *Scandinavian Journal of Statistics*, 171–178.

[4] Azzalini, A. (2013). *The Skew-Normal and Related Families*, 3rd ed., Cambridge University Press.

[5] Azzalini, A. and Arellano-Valle, R.B. (2013). Maximum penalized likelihood estimation for skew-normal and skew-t distributions, *Journal of Statistical Planning and Inference*, **143**(2), 419–433.

[6] Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(2), 367–389.

[7] Bayes, C.L. and Branco, M.D. (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution, *Brazilian Journal of Probability and Statistics*, 141–163.

[8] Bazán, J.L.; Bolfarine, H. and Branco, M.D. (2010). A framework for skew-probit links in binary regression, *Communications in Statistics-Theory and Methods*, **39**(4), 678–6972.

[9] Bazán, J.L.; Branco, M.D.; Bolfarine, H. and others (2006). A skew item response model, *Bayesian Analysis*, **1**(4), 861–892.

[10] Bliss, C.I. (1935). The calculation of the dosage-mortality curve, *Annals of Applied Biology*, **22**(1), 134–167.

[11] Cabras, S.; Racugno, W.; Castellanos, M.E. and Ventura, L. (2012). A matching prior for the shape parameter of the skew-normal distribution, *Scandinavian Journal of Statistics*, **39**(2), 236–247.

[12] Canale, A.; Kenne, P.; Euloge, C. and Scarpa, B. (2016). Bayesian modeling of university first-year students' grades after placement test, *Journal of Applied Statistics*, **43**(16), 3015–3029.

[13] Castro, L.M.; Martín, E.S. and Arellano-Valle, R.B. (2013). A note on the parameterization of multivariate skewed-normal distributions, *Brazilian Journal of Probability and Statistics*, 110–115.

[14] Chen, M.; Dey, D.K. and Shao, Q. (1999). A new skewed link model for dichotomous quantal response data, *Journal of the American Statistical Association*, **94**(448), 1172–1186.

[15] Collet, D. (2003). *Modelling Binary Data*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL.

[16] Czado, C. and Santner, T.J. (1992). The effect of link misspecification on binary regression inference, *Journal of Statistical Planning and Inference*, **33**(2), 213–231.

[17] Dette, H.; Ley, C. and Rubio, F. (2018). Natural (non-) informative priors for skew-symmetric distributions, *Scandinavian Journal of Statistics*, **45**(2), 405–420.

[18] Dua, D. and Graff, C. (2017). *UCI Machine Learning Repository*, University of California, Irvine, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml.

[19] Firth, D. (1993). Bias reduction of maximum likelihood estimates, *Biometrika*, **80**(1), 27–38.

[20] Fuglstad, G.A.; Simpson, D.; Lindgren, F. and Rue, H. (2018). Constructing priors that penalize the complexity of Gaussian random fields, *Journal of the American Statistical Association*, 1–8.

[21] Genton, M.G. (2004). *Skew-elliptical Distributions and their Applications: A Journey beyond Normality*, CRC Press.

[22] Genton, M.G. and Zhang, H. (2012). Identifiability problems in some non-Gaussian spatial random fields, *Chilean Journal of Statistics*, **3**(2), 171–179.

[23] Hallin, M.; Ley, C. and others (2014). Skew-symmetric distributions and Fisher information: the double sin of the skew-normal, *Bernoulli*, **20**(3), 1432–1453.

[24] Klein, N. and Kneib, T. (2016). Scale-dependent priors for variance parameters in structured additive distributional regression, *Bayesian Analysis*, **11**(4), 1071–1106.

[25] Lee, D. and Sinha, S. (2019). Identifiability and bias reduction in the skew-probit model for a binary response, *Journal of Statistical Computation and Simulation*, **89**(9), 1621–1648.

[26] Liseo, B. (1990). The skew-normal class of densities: inferential aspects from a Bayesian viewpoint, *Biometrika*, **50**, 59–70.

[27] Liseo, B. and Loperfido, N. (2006). A note on reference priors for the scalar skew-normal distribution, *Journal of Statistical Planning and Inference*, **136**(2), 373–389.

[28] Maghami, M.M.; Bahrami, M. and Sajadi, F.A. (2020). On bias reduction estimators of skew-normal and skew-t distributions, *Journal of Applied Statistics*, 1–23.

[29] Martins, T.G.; Simpson, D.; Lindgren, F. and Rue, H. (2013). Bayesian computing with INLA: new features, *Computational Statistics and Data Analysis*, **67**, 68–83.

[30] O'Hagan, A. and Leonard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints, *Biometrika*, **63**(1), 201–203.

[31] Otiniano, C.E.G.; Rathie, P.N. and Ozelim, L.C.S.M. (2015). On the identifiability of finite mixture of skew-normal and skew-t distributions, *Statistics & Probability Letters*, **106**, 103–108.

[32] Rue, H.; Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(2), 319–392.

[33] Rue, H.; Riebler, A.; Sørbye, S.H.; Illian, J.B.; Simpson, D. and Lindgren, F. (2017). Bayesian computing with INLA: a review, *Annual Reviews of Statistics and Its Applications*, **4**, 395–421.

[34] Sartori, N. (2006). Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions, *Journal of Statistical Planning and Inference*, **136**(12), 4259–4275.

[35] Simpson, D.; Rue, H.; Riebler, A.; Martins, T.G.; Sørbye, S.H. and others (2017). Penalizing model component complexity: a principled, practical approach to constructing priors, *Statistical Science*, **32**(1), 1–28.

[36] Sørbye, S.H. and Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes, *Journal of Time Series Analysis*, **38**(6), 1467–1492.

REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 23–34

JACKKNIFE EMPIRICAL LIKELIHOOD INFERENCE FOR THE VARIANCE RESIDUAL LIFE FUNCTION

Author: Vali Zardasht
– Department of Statistics, Faculty of Sciences, University of Mohaghegh Ardabili, Ardabil, Iran
zardasht@uma.ac.ir

Received: November 2017 Revised: August 2017 Accepted: August 2018

Abstract:

- In life testing situations, the residual life time of a component which has survived t units of time is $X_t = X - t|X > t$. In this paper, we give a central limit theorem result for the estimator of $\mathrm{Var}(X_t)$, the variance residual life(VRL) function. The result is used to construct normal approximation based confidence interval for the VRL. Furthermore, we use the jackknife empirical likelihood ratio procedure to obtain confidence interval for the VRL function. These intervals are compared through simulation study in terms of the average length and coverage probability. Finally, a numerical example illustrating the theory is also given.

Keywords:

- *confidence interval; coverage probability; jackknife empirical likelihood; U-statistic.*

AMS Subject Classification:

- 62N02, 62N05.

1. INTRODUCTION

Let X be a lifetime random variable with distribution function F and survival function $\bar{F} = 1 - F$ such that $E(X) < \infty$. The residual life random variable at age t, denoted by $X_t = X - t|X > t$, is simply the remaining lifetime beyond that age. The mean residual life (MRL, also known as the mean remaining life) function is defined formally as $\mu(t) = E(X - t|X > t)$. In industrial reliability studies of repair and replacement strategies, the MRL function may prove to be more relevant than the failure (hazard) rate function. The former summarizes the entire residual life distribution, whereas the latter relates only to the risk of immediate failure. In studies of human populations, demographers often refer the MRL under the names of life expectancy or expectation of life. Obviously, the MRL is of vital importance to actuarial work relating to life insurance policies. For a comprehensive literature review about the MRL see Lai and Xie [21].

Another function which has also generated some interest in the recent years is the variance residual life function defined as $\sigma^2(t) = \mathrm{Var}(X - t|X > t)$, see for example, Launer [22] and Gupta *et al.* [14]. An alternative expression for the residual variance in above is given by

$$\sigma^2(t) = E[(X_t - \mu(t))^2] = \frac{1}{\bar{F}(t)} \int_t^\infty (x - t - \mu(t))^2 dF(x) = \frac{2}{\bar{F}(t)} \int_t^\infty \bar{F}(x)\mu(x)dx - \mu^2(t),$$

where $\mu(t)$ is the mean residual life function.

Numerous research works reveal the importance of the VRL function as a reliability function useful in inference procedures and characterizations, and as a means to classify lifetime distribution using its mathematical behaviour. $\sigma^2(t)$ appears in the formula for $\mathrm{Var}(\hat{\mu}_n(t))$, where $\hat{\mu}_n(t)$ is an estimator of the MRL function, see Hall and Wellner [15]. It also appears in the expression of weights assigned for censored observations, see Schmee and Hahn [29]. Launer [22] used $\sigma^2(t)$ to define certain new classes of life distributions and to provide bounds for the reliability function for certain specified class of distributions. Gupta *et al.* [14] shew that the bihaviour of the VRL function is intimately connected to the behaviour of the mean residual life function of the equilibrium distribution. Lynn and Singpurwalla [25] viewed the burn-in concept as a process of reduction of uncertainty of the lifetime of a component. One approach to this is to minimize the VRL. Combining this with maximizing the MRL leads Block *et al.* [5] to consider balancing mean and variance residual life through minimizing the residual coefficient of variation (CV). Characterizations of distributions using the VRL function can be found in Huang and Su [16] and references therein.

The role and properties of the variance residual life and the residual coefficient of variation in reliability have been discussed considerably for continuous lifetime random variables by various authors such as Gupta and Kirmani [11], [12], [13], El-Arishi [8], Al-Zahrani and Stoyanov [4] and Abu-Youssef [1], [2], [3]. Gupta [9], [10] studied the VRL, its monotonicity and the associated aging classes of lifetime distributions. Karlin [19] has studied the monotonic behaviour of $\sigma^2(t)$ when the density is log-convex(log-concave). Kanwar and Madhu [18] gave a test for the VRL. Khorashadizadeh, *et al.* [20] studied properties of the VRL in discrete case. Some stochastic orders have also been defined based on the VRL function (cf. Lai and Xie, [21], p. 61).

Empirical Likelihood (EL) method was originally introduced by Thomas and Grunkkemeier [31] and Owen [26] as a method for constructing nonparametric confidence intervals. During the past decades, the EL method has developed as a very competitive nonparametric test procedure for quite general settings, including the test of a parameter defined by $\int g(t)dF(t)$ with censored survival data (see, e.g., Owen, [27]; Zhao and Qin, [33]; Zhou and Jeong, [34] and the references therein). Inference based on EL has many attractive properties: typically, it does not require estimation of any variance, the range of the parameter space is automatically respected, confidence regions have greater accuracy than those based on the normal approximation approach, furthermore, it inherits all the good properties of the likelihood ratio test and can handle more general types of censored data.

Empirical likelihood has been widely utilized in many settings. However, there exist a lot of computational difficulties when applied to complicated nonlinear functional. To overcome the computational difficulties, a modified EL method was proposed by Jing *et al.* [17], which was called jackknife empirical likelihood (JEL). The main idea of the JEL is to "turn the statistic of interest into a sample mean based on jackknife pseudo-values" (see Quenouille, [28]). The goal of this paper is to develop the jackknife empirical likelihood (JEL) method for interval estimation of the VRL function.

The rest of the paper is organized as follows. A U-statistic based estimator of the VRL, the asymptotic normality of the estimator and the corresponding confidence interval/band are given in Section 2. In this Section, we also propose a jackknife empirical likelihood, an adjusted jackknife empirical likelihood for the VRL function, finding better interval estimators of the VRL function. In Section 3, performance of the jackknife empirical likelihood ratio confidence intervals is compared with the normal approximation based ones in terms of coverage probability and average length through a simulation study. Section 4 looks at a real data example illustrating the methods and finally, some concluding remarks are given in Section 5.

2. INFERENCE METHODS

In this section we give the normal approximation based interval for the VRL function. We also develop new interval estimator using jackknife EL methods. In order to overcome the potential undercoverage problem that the JEL methods may encounter as observed in Jing *et al.* [17], we further propose the adjusted jackknife empirical likelihood by adding one more pseudo-value.

2.1. Normal approximation method

First, note that $\sigma^2(t)$ can be rewritten as

$$\sigma^2(t) = \frac{1}{\bar{F}^2(t)}\left[\bar{F}(t)\int_t^\infty x^2 dF(x) - \left(\int_t^\infty x dF(x)\right)^2\right].$$

Then, given a random sample $X_1, ..., X_n$ from the population of X with distribution function F, the VRL function can be estimated as a ratio of two U-statistics

$$U_n^{(1)} = \frac{1}{\binom{n}{2}} \sum_{1\le i<j\le n} \phi_t^{(1)}(X_i, X_j)$$

and

$$U_n^{(2)} = \frac{1}{\binom{n}{2}} \sum_{1\le i<j\le n} \phi_t^{(2)}(X_i, X_j)$$

with the symmetric kernels $\phi_t^{(1)}(X_1, X_2) = [0.5(X_1^2 + X_2^2) - X_1X_2]I(X_1 > t)I(X_2 > t)$ and $\phi_t^{(2)}(X_1, X_2) = I(X_1 > t)I(X_2 > t)$, that is

$$\hat{\sigma}_n^2(t) = \frac{U_n^{(1)}}{U_n^{(2)}},$$

where $I(\cdot)$ is the indicator function. The following theorem gives the asymptotic distribution of $\hat{\sigma}_n^2(t)$.

Theorem 2.1. *Assume that $E(X^4) < \infty$. Then*

$$\sqrt{n}(\hat{\sigma}_n^2(t) - \sigma^2(t)) \xrightarrow{d} N(0, \upsilon^2(t)),$$

($\xrightarrow{d}$ represents convergence in distribution). $N(0, \upsilon^2(t))$ represents the normal random variable with mean 0 and variance

$$\upsilon^2(t) = 4\left[\frac{\mu_4(t)}{4\bar{F}^2(t)} + \frac{2\mu_1^2(t)\mu_2(t)}{\bar{F}^4(t)} - \frac{\mu_1^4(t)}{\bar{F}^5(t)} - \frac{\mu_1(t)\mu_3(t)}{\bar{F}^3(t)} - \frac{\mu_2^2(t)}{4\bar{F}^3(t)}\right],$$

where $\mu_i(t) = \int_t^\infty x^i dF(x)$, $i = 1, 2, 3, 4$.

Proof: The result immediately follows from Theorem 6.1.6 in Lehmann ([24], p. 376) and the standard delta method. □

It is obvious that $\upsilon^2(t)$ can be consistently estimated by its empirical counterpart,

$$\hat{\upsilon}_n^2(t) = 4\left[\frac{\hat{\mu}_4(t)}{4\bar{F}_n^2(t)} + \frac{2\hat{\mu}_1^2(t)\hat{\mu}_2(t)}{\bar{F}_n^4(t)} - \frac{\hat{\mu}_1^4(t)}{\bar{F}_n^5(t)} - \frac{\hat{\mu}_1(t)\hat{\mu}_3(t)}{\bar{F}_n^3(t)} - \frac{\hat{\mu}_2^2(t)}{4\bar{F}_n^3(t)}\right] I(X_{(n)} > t),$$

where $F_n(t) = \frac{1}{n}\sum_{i=1}^n I(X_i \le t)$ is the empirical distribution function, $\bar{F}_n = 1 - F_n$,

$$\hat{\mu}_i(t) = \int_t^\infty x^i dF_n(x) = \frac{1}{n}\sum_{j=1}^n X_j^i I(X_j > t), \quad i = 1, 2, 3, 4,$$

and $X_{(n)} = \max\{X_1, ..., X_n\}$. Thus, an asymptotic $100(1-\alpha)\%$ confidence interval for $\sigma^2(t)$ at fixed time t based on the above normal approximation can be given by

$$\left\{\sigma^2(t) : n(\hat{\sigma}_n^2(t) - \sigma^2(t))^2 \le \hat{\upsilon}^2(t)\chi_{1-\alpha}^2(1)\right\},$$

where $\chi_{1-\alpha}^2(1)$ is the $100(1-\alpha)$-percentile of the chi-square distribution with one degree of freedom.

The following theorem gives the weak convergence of the stochastic process based on $\hat{\sigma}_n^2(t)$ which can be used to construct a simultaneous confidence band for $\sigma^2(t)$. Let $b < \infty$ and $b \in [0, \tau]$, where $\tau = \inf\{t : F(t) = 1\}$ and denote

$$\begin{aligned} \rho(s,t) &= E\Big[(X - s - \mu(s))^2 (X - t - \mu(t))^2 I(X > t)\Big], \\ \nu(s,t) &= \int_t^\infty (x - s - \mu(s))^2 dF(x). \end{aligned}$$

Theorem 2.2. *Suppose that $E(X^4) < \infty$. Then the process $\sqrt{n}(\hat{\sigma}_n^2(t) - \sigma^2(t))$ for $t \in [0, b]$ converges in distribution to a Gaussian process $U(t)$ with mean zero and covariance function*

$$\begin{aligned} \Gamma(s,t) &= \frac{1}{\bar{F}(s)\bar{F}(t)} \Big[\rho(b,b) - \rho(t,b) - \rho(s,b) + \rho(s,t) - \bar{F}^4(b)\sigma^4(b) \\ &\quad + \bar{F}(s)\bar{F}(b)\sigma^2(s)\sigma^2(b) + \bar{F}(t)\bar{F}(b)\sigma^2(t)\sigma^2(b) - \sigma^2(t)\nu(s,t)\Big], \end{aligned}$$

where $0 \le s \le t \le b$.

Proof: First note that the estimator $\hat{\sigma}_n^2(t)$ can also be given by

$$\begin{aligned} \hat{\sigma}_n^2(t) &= \frac{1}{n\bar{F}_n(t)} \sum_{i=1}^n (X_i - t - \mu_n(t))^2 I(X_i > t) \\ &= \frac{1}{n\bar{F}_n(t)} \sum_{i=1}^n (X_i - t - \mu(t))^2 I(X_i > t) - [\mu_n(t) - \mu(t)]^2, \end{aligned}$$

where $\mu_n(t) = \frac{1}{\bar{F}_n(t)} \int_t^\infty \bar{F}_n(x)dx$ is the empirical estimator of the mean residual life function. Then

$$\sqrt{n}(\hat{\sigma}_n^2(t) - \sigma^2(t)) = \frac{1}{\bar{F}_n(t)} \Big\{V_n(t) - \sigma^2(t)\sqrt{n}[\bar{F}_n(t) - \bar{F}(t)]\Big\} - \sqrt{n}[\mu_n(t) - \mu(t)]^2,$$

where

$$V_n(t) = n^{-\frac{1}{2}} \sum_{i=1}^n \Big[(X_i - t - \mu_n(t))^2 I(X_i > t) - \sigma^2(t)\bar{F}(t)\Big].$$

Applying the same procedure of proof of Lemma 3 in Yang [32] follows that $V_n(t)$ weakly converges to a Gaussian process $V(t)$ with $E[V(t)] = 0$ and

$$\begin{aligned} E[V(s)V(t)] &= \rho(b,b) - \rho(t,b) - \rho(s,b) + \rho(s,t) - \bar{F}^4(b)\sigma^4(b) \\ &\quad + \bar{F}(s)\bar{F}(b)\sigma^2(s)\sigma^2(b) + \bar{F}(t)\bar{F}(b)\sigma^2(t)\sigma^2(b) - \sigma^2(s)\sigma^2(t)\bar{F}(s)\bar{F}(t), \end{aligned}$$

where $0 \le s \le t \le b$. On the other hand, Theorem 1 in Yang [32] implies that $\sqrt{n}[\mu_n(t) - \mu(t)]^2 = o_p(1)$, uniformly in $t \in [0, b]$. The result now follows from the fact that $\sqrt{n}[\bar{F}_n(t) - \bar{F}(t)]$ converges to a Brownian bridge and $\bar{F}_n^{-1}(t) \to \bar{F}^{-1}(t)$ uniformly in $t \in [0, b]$ with probability one. □

Theorem 2.2 can be used to obtain the following confidence band for $\sigma^2(t)$. By the continuous mapping theorem we have

$$\sup_{0 \le t \le b} \Big\{\sqrt{n}(\hat{\sigma}_n^2(t) - \sigma^2(t))\Big\} \xrightarrow{d} \sup_{0 \le t \le b} U(t).$$

Now, we can define the asymptotic $100(1-\alpha)\%$ simultaneous confidence band for $\sigma^2(t)$ in $t \in [0,b]$ as follows:

$$\Big\{\sigma^2(t) : \sqrt{n}(\hat{\sigma}_n^2(t) - \sigma^2(t)) \leq c_\alpha\Big\},$$

where c_α is the upper α-percentile of the distribution of $\sup_{0\leq t\leq b} U(t)$.

2.2. Jackknife empirical likelihood method

In this subsection, we construct a confidence interval for the true $\sigma^2(t)$ via jackknife empirical likelihood (JEL). Let $X_1,...,X_n (n \geq 2)$ be a random sample from a distribution function F. We define a one-sample U-statistic of degree 2

$$U_n(\sigma^2(t)) = \frac{1}{\binom{n}{2}} \sum_{1\leq i<j\leq n} \phi_t(X_i, X_j; \sigma^2(t)),$$

with symmetric kernel

$$\phi_t(X_1, X_2; \sigma^2(t)) = \Big[\sigma^2(t) + X_1X_2 - 0.5(X_1^2 + X_2^2)\Big] I(X_1 > t) I(X_2 > t).$$

It is easy to check that $E[U_n(\sigma^2(t))] = 0$, for the true $\sigma^2(t)$. To apply the JEL, we define our jackknife pseudo-values by

$$\hat{V}_i(\sigma^2(t)) = nU_n(\sigma^2(t)) - (n-1)U_{n-1}^{(-i)}(\sigma^2(t)),$$

where $U_{n-1}^{(-i)}$ is the U-statistic after deleting the ith observation X_i. It can be easily shown that $E[\hat{V}_i] = 0$ and

$$U_n(\sigma^2(t)) = \frac{1}{n}\sum_{i=1}^{n} \hat{V}_i(\sigma^2(t)).$$

Then, one can apply the standard EL method to $\hat{V}_i$. Let $\mathbf{p} = (p_1, ..., p_n)$ be the probability vector over $\hat{V}_i$. The jackknife empirical likelihood ratio at true value $\sigma^2(t)$ is defined by

$$R(\sigma^2(t)) = \max\left\{\prod_{i=1}^{n} np_i : p_i \geq 0, i = 1, ..., n, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i\hat{V}_i(\sigma^2(t)) = 0\right\}.$$

By using the standard Lagrange multiplier method, we know that $R(\sigma^2(t))$ is maximized when

$$p_i = \frac{1}{n}\Big\{1 + \lambda\hat{V}_i(\sigma^2(t))\Big\}^{-1}, \qquad i = 1, ..., n,$$

where $\lambda = \lambda(\sigma^2(t))$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{V}_i(\sigma^2(t))}{1 + \lambda\hat{V}_i(\sigma^2(t))} = 0.$$

Let $g(x) = E[\phi_t(x, X_2; \sigma^2(t))]$ and $\sigma_g^2 = \mathrm{Var}(g(X_1))$. Now we have Wilks' theorem for the JEL as follows.

Theorem 2.3. *Assume that $E(X^4) < \infty$ and $\sigma_g^2 > 0$. Then, as $n \to \infty$*

$$-2\log R(\sigma^2(t)) \xrightarrow{d} \chi_1^2,$$

where χ_1^2 is a chi-distribution with one degree of freedom.

Theorem 2.3 is a special case of Theorem 1 in Jing *et al.* [17] with $m = 2$. Instead of the regularity condition $E[\phi_t^2(X_1, X_2; \sigma^2(t))]$ required by Theorem 1 in Jing *et al.* [17], Theorem 2.3 requires existence of the forth moment because of the specific form of the VRL function.

Following this, an asymptotic $100(1-\alpha)\%$ confidence interval for $\sigma^2(t)$ at time t can be given by

$$\Big\{\tilde{\sigma}^2(t) : -2\log R(\tilde{\sigma}^2(t)) \le \chi^2_{1-\alpha}(1)\Big\},$$

where $\chi^2_{1-\alpha}(1)$ is the is $100(1-\alpha)$-percentile of the chi-square distribution with one degree of freedom.

From practical point of view, the function `el.cen.EM2` inside the package `emplik`, which is an extension package to be used with the R software, carries out calculating the above confidence interval.

Remark 2.1. Using the same procedure as the proof of Theorem 2.2 of Zhao and Qin [33] and following Theorem 2.1 of Jing *et al.* [17], the above Theorem 2.2 implies that

$$-2\log R(\sigma^2(t)) \xrightarrow{d} \frac{W(t)}{4\sigma_g^2},$$

where $W(t)$ is a Gaussian process with mean zero and covariance function

$$\mathrm{Cov}(W(s), W(t)) = \bar{F}(s)\bar{F}(t)\Gamma(s,t).$$

Thus, an JEL-based asymptotic $100(1-\alpha)\%$ simultaneous confidence band for $\sigma^2(t)$ in $t \in [0, b]$ can be given by

$$\Big\{\tilde{\sigma}^2(t) : -2\log R(\tilde{\sigma}^2(t)) \le k_\alpha\Big\},$$

where k_α is the upper α-percentile of the distribution of $\sup_{0\le t\le b} \frac{W(t)}{4\sigma_g^2}$.

2.3. Adjusted jackknife empirical likelihood method

Chen *et al.* [7] developed an adjusted empirical likelihood method, which significantly improves the performance of the empirical likelihood method in terms of coverage probability when the sample size is not large. We adapt their approach to the JEL for $\sigma^2(t)$ by adding one more jackknife pseudo-value

$$\hat{V}_{n+1}(\sigma^2(t)) = -\frac{a_n}{n}\sum_{i=1}^{n} \hat{V}_i(\sigma^2(t)),$$

for constant $a_n = \max\{1, \frac{1}{2}\log(n)\}$. The adjusted jackknife empirical likelihood (AJEL) ratio of $\sigma^2(t)$ is given by

$$R^{ad}(\sigma^2(t)) = \max\Bigg\{\prod_{i=1}^{n+1}(n+1)p_i : p_i \ge 0, i = 1, ..., n+1, \sum_{i=1}^{n+1} p_i = 1, \sum_{i=1}^{n+1} p_i \hat{V}_i(\sigma^2(t)) = 0\Bigg\}.$$

With the same conditions given by Jing *et al.* [17], Wilk's theorem of the AJEL has been established by Chen and Ning [6]. Thus, as a special case, the following theorem holds for the above AJEL ratio. For the proof, we refer the reader to Chen and Ning [6].

Theorem 2.4. *Assume that $E(X^4) < \infty$ and $\sigma_g^2 > 0$. Then, as $n \to \infty$*

$$-2 \log R^{ad}(\sigma^2(t)) \xrightarrow{d} \chi_1^2.$$

A $100(1-\alpha)\%$ confidence interval for $\sigma^2(t)$ by the adjusted JEL method can be developed similarly.

3. SIMULATION STUDY

Simulation exercises were undertaken to assess the performance of the normal approximation (NA) based confidence interval, comparing with the jackknife empirical likelihood (JEL) and adjusted jackknife empirical likelihood (AJEL) confidence intervals in terms of the average length and coverage probability. In the simulation, we considered the following two models for the underling lifetime distribution of X:

- **(i)** X is uniformly distributed on $(0,1)$,
- **(ii)** X has a Weibull distribution with survival function $\bar{F}(x) = e^{-\frac{1}{2}x^2}$.

One can readily show that in case (i)

$$\sigma^2(t) = \frac{1}{3(1-t)}(1 - 3t + 3t^2 - t^3) - \frac{1}{4}(1-t)^2,$$

and in case (ii)

$$\sigma^2(t) = 2\left[1 - \frac{t\bar{\Phi}(t)}{\phi(t)}\right] - 2\pi e^{t^2}\bar{\Phi}^2(t),$$

where $\phi(t)$ and $\bar{\Phi}(t)$ refer to the standard normal density and survival function, respectively. In each case, we ran 2000 simulation trials of different sample sizes $n = 50, 100$ and 150 to obtain confidence intervals with nominal confidence level of 0.95. We compute the average length of intervals and coverage probabilities, i.e. the proportion of intervals which cover the true value $\sigma^2(t)$ for different values of t.

Table 1 – Table 2 summarize the results of the 2000 simulation trials for both models. From the tables, as the sample size n increases, all methods improve in terms of coverage probabilities. It is also evident from the tables that, specially in Weibull model, the coverage probability of the NA confidence interval is not satisfied when the sample size is small and moderate. However, JEL and AJEL produce slightly better coverage probabilities for the same sample size. When the sample size is large, NA, JEL and AJEL methods have similar performance in terms of coverage probability. We can see coverage probability for AJEL is very close to nominal level 0.95, and AJEL has better performance than JEL for the small sample size. Though, for large values of t, the coverage probability of all the methods is slightly far from the nominal level.

For all the methods, the length of confidence interval becomes shorter when the sample size becomes larger. When the sample size increases from moderate to large, the length of confidence interval for all the methods are very close. It seems that, for large values of t, the length of the NA confidence intervals is slightly shorter than JEL and AJEL confidence intervals.

Table 1: Empirical coverage probabilities (average length) for $\sigma^2(t)$, uniform model.

n	Method	$t = 0$	$t = 0.2$	$t = 0.4$	$t = 0.6$	$t = 0.8$
50	NA	0.930 (0.041)	0.935 (0.029)	0.923 (0.019)	0.906 (0.010)	0.822 (0.003)
	JEL	0.929 (0.040)	0.935 (0.029)	0.924 (0.018)	0.888 (0.010)	0.908 (0.068)
	AJEL	0.940 (0.042)	0.946 (0.030)	0.937 (0.019)	0.908 (0.011)	0.946 (0.089)
100	NA	0.939 (0.029)	0.936 (0.021)	0.930 (0.013)	0.928 (0.007)	0.900 (0.002)
	JEL	0.935 (0.028)	0.930 (0.020)	0.930 (0.013)	0.922 (0.007)	0.919 (0.003)
	AJEL	0.941 (0.029)	0.938 (0.021)	0.935 (0.013)	0.931 (0.007)	0.932 (0.003)
150	NA	0.941 (0.024)	0.936 (0.017)	0.944 (0.011)	0.938 (0.006)	0.920 (0.002)
	JEL	0.937 (0.022)	0.932 (0.016)	0.943 (0.011)	0.940 (0.006)	0.947 (0.002)
	AJEL	0.941 (0.022)	0.937 (0.017)	0.947 (0.011)	0.942 (0.006)	0.950 (0.002)

Table 2: Empirical coverage probabilities (average length) for $\sigma^2(t)$, Weibull model.

n	Method	$t = 0$	$t = 0.25$	$t = 0.5$	$t = 1$	$t = 1.7$
50	NA	0.887 (0.328)	0.878 (0.316)	0.864 (0.300)	0.834 (0.277)	0.679 (0.250)
	JEL	0.909 (0.304)	0.903 (0.306)	0.894 (0.302)	0.862 (0.296)	0.870 (0.426)
	AJEL	0.924 (0.315)	0.910 (0.318)	0.908 (0.314)	0.873 (0.310)	0.896 (0.480)
100	NA	0.913 (0.239)	0.924 (0.236)	0.908 (0.223)	0.886 (0.214)	0.778 (0.207)
	JEL	0.925 (0.240)	0.929 (0.240)	0.919 (0.167)	0.917 (0.154)	0.834 (0.241)
	AJEL	0.935 (0.245)	0.935 (0.246)	0.928 (0.170)	0.923 (0.157)	0.845 (0.251)
150	NA	0.927 (0.200)	0.926 (0.193)	0.923 (0.185)	0.909 (0.179)	0.811 (0.187)
	JEL	0.929 (0.200)	0.938 (0.144)	0.929 (0.132)	0.926 (0.184)	0.859 (0.201)
	AJEL	0.934 (0.203)	0.940 (0.146)	0.934 (0.134)	0.930 (0.187)	0.865 (0.205)

4. REAL DATA ANALYSIS

In this section, we use a real data coming from reliability engineering to illustrate applications of the NA-based and JEL-based confidence intervals for the VRL function. Since the variance estimator $\hat{v}^2(t)$ is unstable, the NA-based confidence interval for the VRL contains negative values. In the following computation results, the values outside of the positive range of the VRL are removed and the negative lower bounds of the confidence intervals are replaced with zero.

Lawless [23] used the breaking strengths of single carbon fibers of different to fit a parametric regression model. We use the data set consisting of breaking strengths of 57 single carbon fibers with unit length taken from Lawless [23] to estimate $\sigma^2(t)$. Table 3 gives the estimated VRL function and corresponding 95% lower bound (LB), upper bound (UB) and length based on the NA, JEL and AJEL methods at different time points t. We can see from the table that the lengths of confidence intervals for the NA is longer than one for the JEL and AJEL methods. Also, there is no big difference among the lengths of the JEL and AJEL confidence intervals.

Table 3: Estimated variance residual lifetimes, 95% confidence intervals and lengths, carbon fiber data.

t	VRL	NA			JEL			AJEL		
		LB	UB	Length	LB	UB	Length	LB	UB	Length
0.5	0.697	0	2.503	3.612	0.477	0.998	0.521	0.472	1.015	0.543
2.5	0.637	0	2.232	3.189	0.443	0.901	0.458	0.433	0.911	0.477
3.5	0.432	0	1.627	2.389	0.291	0.643	0.352	0.284	0.654	0.370
4.5	0.250	0	1.057	1.615	0.145	0.413	0.268	0.138	0.423	0.285
5.0	0.130	0	0.620	0.979	0.001	1.129	1.128	0.328	1.129	0.801

5. CONCLUSION

In this paper, we have considered an estimator of the VRL function. The estimator was shown to converge in distribution to a normal random variable. Furthermore, a confidence interval for the VRL function at time t was constructed by using the normal approximation (NA) method. As alternative methods, we have also considered constructing confidence interval/band for the VRL function using the jackknife empirical likelihood (JEL) and adjusted jackknife empirical likelihood (AJEL) approaches. A major advantage of the EL-based method is no need for nonparametric estimation of any kind of variance for statistical inference. A simulation exercise was undertaken to compare between the performance of the NA-based and El-based confidence intervals in terms of coverage probabilities and the average lengths. As shown from the simulation study, the coverage probability for the NA method is far away from our expectation when the sample size is small. However, the coverage probability of confidence intervals for JEL and AJEL methods is very close to nominal level.

The length of confidence interval for all the methods is very close when the sample size increases from moderate to large. Finally, using a numerical example, the application of the methods for constructing confidence intervals was illustrated.

ACKNOWLEDGMENTS

The author is grateful for the suggestions by the referee and the Associate Editor which improved the presentation of the article.

REFERENCES

[1] Abu-Youssef, S.E. (2004). Nonparametric test for monotone variance residual life class of life distributions with hypothesis testing applications, *Applied Mathematics and Computations*, **158**, 817–826.

[2] Abu-Youssef, S.E. (2007). Testing decreasing (increasing) variance residual class of life distributions using kernel method, *Applied Mathematical Sciences*, **1**, 1915–1927.

[3] Abu-Youssef, S.E. (2009). A goodness of fit approach to monotone variance residual life class of life distributions, *Applied Mathematical Sciences*, **3**(15), 715–724.

[4] Al-Zahrani, B. and Stoyanov, J. (2008). On some properties of life distributions with increasing elasticity and log-concavity, *Applied Mathematical Sciences*, **2**(48), 2349–2361.

[5] Block, H.W.; Savits, T.H. and Singh, H. (2002). A criterion for burn-in that balances mean residual life and residual variance, *Operations Research*, **50**, 290–296.

[6] Chen, Y.J. and Ning, W. (2016). Adjusted jackknife empirical likelihood, *arXiv:1603.04093v1*.

[7] Chen, J.; Variyath, A.M. and Abraham, B. (2008). Adjusted empirical likelihood and its properties, *Journal of Computational and Graphical Statistics*, **17**, 426–443.

[8] El-Arishi, S. (2005). A conditional variance characterization of some discrete probability distributions, *Statistical Papers*, **46**, 31–45.

[9] Gupta, R.C. (1987). On the monotonic properties of the residual variance and their application in reliability, *Journal of Statistical Planning and Inference*, **16**, 329–335.

[10] Gupta, R.C. (2006). Variance residual life function in reliability studies, *Metron-International Journal of Statistics*, **LXIV**(3), 343–355.

[11] Gupta, R.C. and Kirmani, S.N.U.A. (1998). On the proportional mean residual life model and its implications, *Statistics*, **32**, 175–187.

[12] Gupta, R.C. and Kirmani, S.N.U.A. (2000). Residual coefficient of variation and some characterization results, *Journal of Planning and Statistical Inference*, **91**, 23–31.

[13] Gupta, R.C. and Kirmani, S.N.U.A. (2004). Moments of residual life and some characterization results, *Journal of Applied Statistical Science*, **13**(2), 155–167.

[14] Gupta, R.C.; Kirmani, S.N.U.A. and Launer, R.L. (1987). On life distributions having monotone residual variance, *Probability in the Engineering and Informational Sciences*, **1**, 299–307.

[15] Hall, W.J. and Wellner, J.A. (1981). *Mean residual life.* In "Proceedings of the International Symposium on Statistics and Related Topics" (M. Csorgo, D.A. Dawson, J.N.K. Rao and A.K.Md.E. Saleh, Eds.), North-Holland Publishing Co., Amsterdam, 169–184.

[16] Huang, W.J. and Su, N.C. (2012). Characterizations of distributions based on moments of residual life, *Communications in Statistics – Theory and Methods*, **41**(15), 2750–2761.

[17] Jing, B.; Yuan, Q. and Zhou, W. (2009). Jackknife empirical likelihood, *Journal of the American Statistical Association*, **104**, 1224–1232.

[18] Kanwar, S. and Madhu, B.J. (1991). A test for the variance residual life, *Communications in Statistics – Theory and Methods*, **20**(1), 327–331.

[19] Karlin, S. (1982). *Some results on optimal partinioning of variance and monotonicity with truncation level.* In "Statistics and Probability: Essays in honor of C.R. Rao" (G. Kallianpur, P.R. Krishnaiah and J.K. Ghosh, Eds.), North Holland Publishing Co., Amsterdam, 375–382.

[20] Khorashadizadeh, M.; Rezaei Roknabadi, A.H. and Mohtashami Borzadran, G.R. (2010). Variance residual life function in discrete random ageing, *Metron-International Journal of Statistics*, **LXVIII**(1), 67–75.

[21] Lai, C. and Xie, M. (2006). *Stochastic Ageing and Dependence for Reliability*, Springer, New York.

[22] Launer, R.L. (1984). Inequalities for NBUE and NWUE life distributions, *Operations Research*, **32**, 660–667.

[23] Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*, John Wiley and Sons, New Jersey.

[24] Lehmann, E.L. (1999). *Elements of Large-Sample Theory*, Springer-Verlag, New York.

[25] Lynn, N.J. and Singpurwalla, N.D. (1997). Comment: "Burn-in" makes us feel good, *Statistical Science*, **12**, 13–19.

[26] Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, **75**, 237–249.

[27] Owen, A. (2001). *Empirical Likelihood*, Chapman and Hall, London.

[28] Quenouille, M. (1956). Notes on bias in estimation, *Biometrika*, **10**, 353–360.

[29] Schmee, J. and Hahn, G.J. (1979). A simple method for regression analysis with censored data, *Technometrics*, **21**(4), 417–432.

[30] Sen, K. and Jain, M.B. (1991). A test for the variance residual life, *Communications in Statistics – Theory and Methods*, **20**(1), 327–331.

[31] Thomas, D.R. and Grunkemeier, G.L. (1975). Confidence interval estimation of survival probabilities for censored data, *Journal of the American Statistical Association*, **70**, 865–871.

[32] Yang, G.L. (1978). Estimating of a biometric function, *Annals of Statistics*, **6**(1), 112–116.

[33] Zhao, Y. and Qin, G. (2006). Inference for the mean residual life function via empirical likelihood, *Communications in Statistics: Theory and Methods*, **35**, 1025–1036.

[34] Zhou, M. and Jeong, J.H. (2011). Empirical likelihood ratio test for median and mean residual lifetime, *Statistics in Medicine*, **30**(2), 152–159.

REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 35–46

FINITE MIXTURES OF MULTIVARIATE SKEW LAPLACE DISTRIBUTIONS

Authors: Fatma Zehra Doğru
– Department of Statistics, Faculty of Arts and Science, Giresun University,
Giresun, Turkey
fatma.dogru@giresun.edu.tr

Y. Murat Bulut
– Department of Statistics, Faculty of Science and Letters, Eskisehir Osmangazi University,
Eskişehir, Turkey
ymbulut@ogu.edu.tr

Olcay Arslan
– Department of Statistics, Faculty of Science, Ankara University,
Ankara, Turkey
oarslan@ankara.edu.tr

Received: February 2017 Revised: September 2018 Accepted: September 2018

Abstract:

- This paper proposes finite mixtures of multivariate skew Laplace distributions in order to model both skewness and heavy-tailedness in heterogeneous data sets. Maximum likelihood estimators for the parameters of interest are obtained using the EM algorithm. The paper offers a small simulation study and a real data example to illustrate the performance of the proposed mixture model.

Keywords:

- *EM algorithm; ML estimation; multivariate mixture model; MSL.*

AMS Subject Classification:

- 62H12, 65C60.

1. INTRODUCTION

Finite mixture models are used to model heterogeneous data sets thanks to their flexibility. These models are commonly applied in fields such as classification, cluster and latent class analysis, density estimation, data mining, image analysis, genetics, medicine, pattern recognition and suchlike; for more detail see [7, 12, 20, 21, 27].

In general, the distribution of mixture model components is assumed to be normal because of its tractability and wide applicability. In practice, however, the data sets may be asymmetric and/or heavy-tailed. For instance, there have been a number of studies focusing on multivariate mixture modeling using asymmetric and/or heavy-tailed distributions: [21] propose finite mixtures of multivariate t distributions as a robust extension of the multivariate normal mixture model ([20]); [16] introduces multivariate skew normal mixture models; [24] and [17] examine finite mixtures of restricted and unrestricted variants of the multivariate skew t distributions of [25]; [8] explore multivariate mixture modeling based on skew-normal independent distributions; and [18] introduce flexible mixture modeling based on skew-t-normal distribution.

In multivariate analysis, the multivariate skew normal (MSN) distribution, [5], [14] and [2], is proposed as an alternative to the multivariate normal (MN) distribution in order to deal with skewness in the data. However, certain alternative heavy-tailed skew distributions are required to model skewness and heavy-tailedness because MSN distribution is not heavy-tailed. One such example of heavy-tailed skew distribution is the multivariate skew t (MST) distribution, which is defined by [4] and [13]. [3] also proposes another heavy-tailed skew distribution called the multivariate skew Laplace (MSL) distribution, using a variance-mean mixture of the normal distribution. One advantage of the MSL distribution is that it has a smaller number of parameters than the MST distribution and has the same number of parameters as the MSN distribution. Regarding finite mixtures of the multivariate skew distributions, finite mixtures of MSN distributions were proposed by [16] to model heterogeneous data sets as they may not be able to modeled by mixtures of MN distributions due to the skew feature of data. On the other hand, data sets may not only have a skewness problem, but may also have a heavy-tailedness problem to be dealt with. For this reason, in this study, finite mixtures of MSL distributions as an alternative to finite mixtures of MSN distributions are explored in order to deal with both skewness and heavy-tailedness in heterogeneous data sets.

The rest of the paper is organized as follows: Section 2 summarizes certain properties of the MSL distribution; see [3] for further details of the MSL distribution. Section 3 presents mixtures of MSL distributions and gives the Expectation-Maximization (EM) algorithm to obtain maximum likelihood (ML) estimators for the parameters of the proposed mixture model. Section 4 offers the empirical information matrix of MSL distribution to compute standard errors of proposed estimators. Section 5 provides a small simulation study and a real data example to illustrate the performance of the proposed mixture model. Finally, Section 6 is devoted to conclusions.

2. MULTIVARIATE SKEW LAPLACE DISTRIBUTION

Let $\boldsymbol{Y} \in R^p$ be a p-dimensional random vector which has the MSL distribution ($\boldsymbol{Y} \sim \mathrm{MSL}_p(\boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$) proposed by [3]. The probability density function (pdf) of this distribution is given below:

$$\begin{aligned} f_{\mathrm{MSL}}(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma}) &= \frac{|\,\Sigma\,|^{-\frac{1}{2}}}{2^p \pi^{\frac{p-1}{2}} \alpha \Gamma\left(\frac{p+1}{2}\right)} \\ &\quad \times \exp\left\{-\alpha\sqrt{(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}} \Sigma^{-1} (\boldsymbol{y}-\boldsymbol{\mu})} + (\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}} \Sigma^{-1} \boldsymbol{\gamma}\right\}, \end{aligned} \tag{2.1}$$

where $\alpha = \sqrt{1+\boldsymbol{\gamma}^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\gamma}}$, $\boldsymbol{\mu} \in R^p$ is the location parameter, $\boldsymbol{\gamma} \in R^p$ is the skewness parameter, Σ is the positive definite scatter matrix and $\Gamma(\cdot)$ represents the complete gamma function.

Proposition 2.1. *The characteristic function of* $\mathrm{MSL}_p(\boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ *is*

$$\Phi_{\boldsymbol{Y}}(t) = e^{it^{\mathsf{T}}\boldsymbol{\mu}} \left[1 + t^{\mathsf{T}}\Sigma t - 2it^{\mathsf{T}}\boldsymbol{\gamma}\right]^{-\frac{p+1}{2}}, \quad t \in R^p.$$

See [3] for proof of this proposition.

If $\boldsymbol{Y} \sim \mathrm{MSL}_p(\boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ then the expectation and variance of $\boldsymbol{Y}$ are:

$$\begin{aligned} E(\boldsymbol{Y}) &= \boldsymbol{\mu} + (p+1)\boldsymbol{\gamma}, \\ \mathrm{Var}(\boldsymbol{Y}) &= (p+1)\left(\Sigma + 2\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathsf{T}}\right). \end{aligned}$$

The MSL distribution can be obtained as a variance-mean mixture of MN distribution and inverse gamma (IG) distribution. The variance-mean mixture representation is given as follows:

$$\boldsymbol{Y} = \boldsymbol{\mu} + V^{-1}\boldsymbol{\gamma} + \sqrt{V^{-1}}\Sigma^{\frac{1}{2}}\boldsymbol{X}, \tag{2.2}$$

where $\boldsymbol{X} \sim N_p(\boldsymbol{0}, I_p)$ and $V \sim IG\left(\frac{p+1}{2}, \frac{1}{2}\right)$. Note that if $\boldsymbol{\gamma} = \boldsymbol{0}$, the density function of $\boldsymbol{Y}$ reduces to the density function of symmetric multivariate Laplace distribution given by [22]. In addition, the conditional distribution of $\boldsymbol{Y}$ given $V = v$ will be:

$$\boldsymbol{Y}|v \sim N_p\left(\boldsymbol{\mu} + v^{-1}\boldsymbol{\gamma}, v^{-1}\Sigma\right).$$

The joint density function of $\boldsymbol{Y}$ and V is:

$$\begin{aligned} f(\boldsymbol{y}, v) &= \frac{|\,\Sigma\,|^{-\frac{1}{2}}\, e^{(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\gamma}}}{2^p \pi^{\frac{p-1}{2}} \alpha \Gamma\left(\frac{p+1}{2}\right)} \\ &\quad \times \left\{v^{-\frac{3}{2}} e^{-\frac{1}{2}\left\{(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})v + \left(1+\boldsymbol{\gamma}^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\gamma}\right)v^{-1}\right\}}\right\}. \end{aligned}$$

Then, we have the following conditional density function of V given $\boldsymbol{Y}$:

$$\begin{aligned} f(v|\boldsymbol{y}) &= \frac{\alpha}{\sqrt{2\pi}} e^{\alpha\sqrt{(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})}} \\ &\quad \times v^{-\frac{3}{2}} e^{-\frac{1}{2}\left\{(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})v + \alpha^2 v^{-1}\right\}}, \quad v > 0. \end{aligned} \tag{2.3}$$

Using the conditional density function given in (2.3), the conditional expectations can be obtained as follows:

$$E\left(V|\boldsymbol{y}\right) = \frac{\sqrt{1+\boldsymbol{\gamma}^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\gamma}}}{\sqrt{\left(\boldsymbol{y}-\boldsymbol{\mu}\right)^{\mathsf{T}}\Sigma^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)}}, \tag{2.4}$$

$$E\left(V^{-1}|\boldsymbol{y}\right) = \frac{1+\sqrt{\left(1+\boldsymbol{\gamma}^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\gamma}\right)\left(\boldsymbol{y}-\boldsymbol{\mu}\right)^{\mathsf{T}}\Sigma^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)}}{1+\boldsymbol{\gamma}^{\mathsf{T}}\Sigma^{-1}\boldsymbol{\gamma}}. \tag{2.5}$$

Note that these conditional expectations will be used in the EM algorithm given in subsection 3.1; see [3] for further details of the MSL distribution.

3. FINITE MIXTURES OF MSL DISTRIBUTIONS

Let $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n$ be p-dimensional random sample which comes from a g-component mixtures of MSL distributions. The pdf of a g-component finite mixtures of MSL distributions is given by:

$$f\left(\boldsymbol{y}|\boldsymbol{\Theta}\right) = \sum_{i=1}^{g} \pi_i f\left(\boldsymbol{y};\boldsymbol{\mu}_i,\Sigma_i,\boldsymbol{\gamma}_i\right), \tag{3.1}$$

where π_i denotes the mixing probability with $\sum_{i=1}^{g}\pi_i = 1,\ 0 \le \pi_i \le 1,\ f\left(\boldsymbol{y};\boldsymbol{\mu}_i,\Sigma_i,\boldsymbol{\gamma}_i\right)$ represents the pdf of the i-th component (pdf of the MSL distribution) given in (2.1) and $\boldsymbol{\Theta} = \left(\pi_1, ..., \pi_g, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_g, \Sigma_1, ..., \Sigma_g, \boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_g\right)^{\mathsf{T}}$ is the unknown parameter vector.

3.1. ML estimation

The ML estimator of $\boldsymbol{\Theta}$ can be found by maximizing the following log-likelihood function:

$$\ell\left(\boldsymbol{\Theta}\right) = \sum_{j=1}^{n}\log\left(\sum_{i=1}^{g}\pi_i f\left(\boldsymbol{y}_j;\boldsymbol{\mu}_i,\Sigma_i,\boldsymbol{\gamma}_i\right)\right). \tag{3.2}$$

However, there is not an explicit maximizer of (3.2). Therefore, in general, the EM algorithm ([9]) is used to obtain the ML estimator of $\boldsymbol{\Theta}$. Here, we will use the following EM algorithm:

Let $\boldsymbol{Z}_j = \left(Z_{1j}, ..., Z_{gj}\right)^{\mathsf{T}}$ be the latent variables with

$$Z_{ij} = \begin{cases} 1, & \text{if } j^{\text{th}} \text{ observation belongs to } i^{\text{th}} \text{ component,} \\ 0, & \text{otherwise,} \end{cases} \tag{3.3}$$

where $j = 1, ..., n$ and $i = 1, ..., g$. To implement the steps of the EM algorithm, we will use the stochastic representation of the MSL distribution given in (2.2). If we do so, the hierarchical representation for the mixtures of MSL distributions will be:

$$\begin{aligned} \mathbf{Y}_j|v_j, z_{ij} = 1 &\sim N\left(\boldsymbol{\mu} + v_j^{-1}\boldsymbol{\gamma}, v_j^{-1}\Sigma\right), \\ V_j|z_{ij} = 1 &\sim IG\left(\frac{p+1}{2}, \frac{1}{2}\right). \end{aligned} \tag{3.4}$$

Let $(\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{z})$ be the complete data, where $\boldsymbol{y} = \left(\boldsymbol{y}_1^{\mathsf{T}}, ..., \boldsymbol{y}_n^{\mathsf{T}}\right)^{\mathsf{T}}$, $\boldsymbol{v} = (v_1, ..., v_n)$ and $\boldsymbol{z} = (z_1, ..., z_n)^{\mathsf{T}}$. Using the hierarchical representation given above and ignoring the constants, the complete data log-likelihood function can be written as:

$$\begin{aligned} \ell_c(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{v}, \boldsymbol{z}) &= \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\left\{\log \pi_i - \frac{1}{2}\log \mid \Sigma_i \mid + (\boldsymbol{y}_j - \boldsymbol{\mu}_i)^{\mathsf{T}} \Sigma_i^{-1}\boldsymbol{\gamma}_i \right. \\ &\left. - \frac{1}{2}v_j (\boldsymbol{y}_j - \boldsymbol{\mu}_i)^{\mathsf{T}} \Sigma_i^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_i) - \frac{1}{2}\boldsymbol{\gamma}_i^{\mathsf{T}}\Sigma_i^{-1}\boldsymbol{\gamma}_i v_j^{-1} - \frac{1}{2}\left(3\log v_j + v_j^{-1}\right)\right\}. \end{aligned} \tag{3.5}$$

To overcome the latency of the latent variables given in (3.5), we have to take the conditional expectation of the complete data log-likelihood function given the observed data $\boldsymbol{y}_j$

$$\begin{aligned} E\left(\ell_c(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{v}, \boldsymbol{z}|\boldsymbol{y}_j)\right) &= \sum_{j=1}^{n}\sum_{i=1}^{g} E(Z_{ij}|\boldsymbol{y}_j)\left\{\log \pi_i - \frac{1}{2}\log \mid \Sigma_i \mid - (\boldsymbol{y}_j - \boldsymbol{\mu}_i)^{\mathsf{T}} \Sigma_i^{-1}\boldsymbol{\gamma}_i \right. \\ &\left. - \frac{1}{2}E(V_j|\boldsymbol{y}_j)(\boldsymbol{y}_j - \boldsymbol{\mu}_i)^{\mathsf{T}} \Sigma_i^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_i) - \frac{1}{2}\boldsymbol{\gamma}_i^{\mathsf{T}}\Sigma_i^{-1}\boldsymbol{\gamma}_i E\left(V_j^{-1}|\boldsymbol{y}_j\right)\right\}. \end{aligned} \tag{3.6}$$

Since the last part of the complete data log-likelihood function does not include the parameters of interest it is omitted and the conditional expectation of the other terms are taken. The conditional expectations $E(V_j|\boldsymbol{y}_j)$ and $E\left(V_j^{-1}|\boldsymbol{y}_j\right)$ can be calculated using the conditional expectations given in (2.4) and (2.5), and the conditional expectation $E(Z_{ij}|\boldsymbol{y}_j)$ can be computed using the classical theory of mixture modeling. Next, the steps of the EM algorithm can be formed as follows:

EM algorithm:

1. Set initial parameter estimate $\boldsymbol{\Theta}^{(0)}$ and a stopping rule Δ.

2. E-Step: Compute the following conditional expectations for $k = 0, 1, 2, ...$ iteration

$$\widehat{z}_{ij}^{(k)} = E\left(Z_{ij}|\boldsymbol{y}_j, \widehat{\boldsymbol{\Theta}}^{(k)}\right) = \frac{\widehat{\pi}_i^{(k)} f\left(\boldsymbol{y}_j; \widehat{\boldsymbol{\mu}}_i^{(k)}, \widehat{\Sigma}_i^{(k)}, \widehat{\boldsymbol{\gamma}}_i^{(k)}\right)}{f\left(\boldsymbol{y}_j; \widehat{\boldsymbol{\Theta}}^{(k)}\right)}, \tag{3.7}$$

$$\widehat{v}_{1ij}^{(k)} = E\left(V_j|\boldsymbol{y}_j, \widehat{\boldsymbol{\Theta}}^{(k)}\right) = \frac{\sqrt{1 + \widehat{\boldsymbol{\gamma}}_i^{(k)^{\mathsf{T}}} \widehat{\Sigma}_i^{(k)^{-1}} \widehat{\boldsymbol{\gamma}}_i^{(k)}}}{\sqrt{\left(\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_i^{(k)}\right)^{\mathsf{T}} \widehat{\Sigma}_i^{(k)^{-1}} \left(\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_i^{(k)}\right)}}, \tag{3.8}$$

$$\begin{aligned} \widehat{v}_{2ij}^{(k)} &= E\left(V_j^{-1}|\boldsymbol{y}_j, \widehat{\boldsymbol{\Theta}}^{(k)}\right) \\ &= \frac{1 + \sqrt{\left(1 + \widehat{\boldsymbol{\gamma}}_i^{(k)^{\mathsf{T}}} \widehat{\Sigma}_i^{(k)^{-1}} \widehat{\boldsymbol{\gamma}}_i^{(k)}\right)\left(\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_i^{(k)}\right)^{\mathsf{T}} \widehat{\Sigma}_i^{(k)^{-1}} \left(\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_i^{(k)}\right)}}{1 + \widehat{\boldsymbol{\gamma}}_i^{(k)^{\mathsf{T}}} \widehat{\Sigma}_i^{(k)^{-1}} \widehat{\boldsymbol{\gamma}}_i^{(k)}}. \end{aligned} \tag{3.9}$$

Using these conditional expectations, we form the following objective function:

$$\begin{aligned} Q\left(\boldsymbol{\Theta}; \widehat{\boldsymbol{\Theta}}^{(k)}\right) &= \sum_{j=1}^{n}\sum_{i=1}^{g} \widehat{z}_{ij}^{(k)}\left\{\log \pi_i - \frac{1}{2}\log \mid \Sigma_i \mid - (\boldsymbol{y}_j - \boldsymbol{\mu}_i)^{\mathsf{T}} \Sigma_i^{-1}\boldsymbol{\gamma}_i \right. \\ &\left. - \frac{1}{2}\widehat{v}_{1ij}^{(k)} (\boldsymbol{y}_j - \boldsymbol{\mu}_i)^{\mathsf{T}} \Sigma_i^{-1} (\boldsymbol{y}_j - \boldsymbol{\mu}_i) - \frac{1}{2}\widehat{v}_{2ij}^{(k)}\boldsymbol{\gamma}_i^{\mathsf{T}}\Sigma_i^{-1}\boldsymbol{\gamma}_i\right\}. \end{aligned} \tag{3.10}$$

3. M-Step: Maximize the $Q\left(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}}^{(k)}\right)$ with respect to $\boldsymbol{\Theta}$ to get the $(k+1)$-th parameter estimates for the parameters. This maximization yields the following updating equations:

$$\widehat{\pi}_i^{(k+1)} = \frac{\sum_{j=1}^n \widehat{z}_{ij}^{(k)}}{n}, \tag{3.11}$$

$$\widehat{\boldsymbol{\mu}}_i^{(k+1)} = \frac{\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\widehat{v}_{1ij}^{(k)}\boldsymbol{y}_j - \sum_{j=1}^n \widehat{z}_{ij}^{(k)}\widehat{\boldsymbol{\gamma}}_i^{(k)}}{\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\widehat{v}_{1ij}^{(k)}}, \tag{3.12}$$

$$\widehat{\boldsymbol{\gamma}}_i^{(k+1)} = \frac{\left(\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\widehat{v}_{1ij}^{(k)}\right)\left(\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\boldsymbol{y}_j\right) - \left(\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\right)\left(\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\widehat{v}_{1ij}^{(k)}\boldsymbol{y}_j\right)}{\left(\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\widehat{v}_{1ij}^{(k)}\right)\left(\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\widehat{v}_{2ij}^{(k)}\right) - \left(\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\right)^2}, \tag{3.13}$$

$$\widehat{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\widehat{v}_{1ij}^{(k)}\left(\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_i^{(k)}\right)\left(\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_i^{(k)}\right)^{\mathsf{T}} - \widehat{\boldsymbol{\gamma}}_i^{(k)}\widehat{\boldsymbol{\gamma}}_i^{(k)\mathsf{T}}\sum_{j=1}^n \widehat{z}_{ij}^{(k)}\widehat{v}_{2ij}^{(k)}}{\sum_{j=1}^n \widehat{z}_{ij}^{(k)}}. \tag{3.14}$$

4. Repeat E and M steps until the convergence rule $\| \widehat{\boldsymbol{\Theta}}^{(k+1)} - \widehat{\boldsymbol{\Theta}}^{(k)} \| < \Delta$ is obtained. Alternatively, the absolute difference of the actual log-likelihood $\left\|\ell\left(\widehat{\boldsymbol{\Theta}}^{(k+1)}\right) - \ell\left(\widehat{\boldsymbol{\Theta}}^{(k)}\right)\right\| < \Delta$ or $\left\|\ell\left(\widehat{\boldsymbol{\Theta}}^{(k+1)}\right)/\ell\left(\widehat{\boldsymbol{\Theta}}^{(k)}\right) - 1\right\| < \Delta$ can be used as a stopping rule ([10]).

3.2. Initial values

In order to determine the initial values for the EM algorithm, the following procedure given by [16] will be used:

i) Perform the K-means clustering algorithm ([15]).

ii) Initialize the component labels $\widehat{\boldsymbol{z}}_j^{(0)} = \{z_{ij}\}_{i=1}^g$ according to the K-means clustering results.

iii) The initial values of mixing probabilities, component locations and component scale variances can be set as:

$$\widehat{\pi}_i^{(0)} = \frac{\sum_{j=1}^n \widehat{z}_{ij}^{(0)}}{n},$$

$$\widehat{\boldsymbol{\mu}}_i^{(0)} = \frac{\sum_{j=1}^n \widehat{z}_{ij}^{(0)}\boldsymbol{y}_j}{\sum_{j=1}^n \widehat{z}_{ij}^{(0)}},$$

$$\widehat{\Sigma}_i^{(0)} = \frac{\sum_{j=1}^n \widehat{z}_{ij}^{(0)}\left(\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_i^{(0)}\right)\left(\boldsymbol{y}_j - \widehat{\boldsymbol{\mu}}_i^{(0)}\right)^{\mathsf{T}}}{\sum_{j=1}^n \widehat{z}_{ij}^{(0)}}.$$

iv) For the skewness parameters, use the skewness coefficient vector of each cluster.

4. THE EMPIRICAL INFORMATION MATRIX

We will compute the standard errors of ML estimators using the information based method given by [6]. At this point, we will use the inverse of the empirical information matrix in order to have an approximation to the asymptotic covariance matrix of the estimators.

This information matrix can be obtained as:

$$\widehat{I}_e = \sum_{j=1}^{n} \widehat{s}_j \widehat{s}_j^{\mathsf{T}}, \tag{4.1}$$

where $\widehat{s}_j = E_{\widehat{\boldsymbol{\Theta}}}\left(\frac{\partial \ell_{cj}(\boldsymbol{\Theta};\boldsymbol{y}_j,\boldsymbol{v}_j,\boldsymbol{z}_j)}{\partial \boldsymbol{\Theta}} | \boldsymbol{y}_j\right)$, $j=1,2,...,n$ are the individual scores and $\ell_{cj}\left(\boldsymbol{\Theta};\boldsymbol{y}_j,\boldsymbol{v}_j,\boldsymbol{z}_j\right)$ is the complete data log-likelihood function for the j-th observation. The components of the score vector $\widehat{\boldsymbol{s}}_j$ are $\left(\widehat{\boldsymbol{s}}_{j,\pi_1},...,\widehat{\boldsymbol{s}}_{j,\pi_{g-1}},\widehat{\boldsymbol{s}}_{j,\boldsymbol{\mu}_1},...,\widehat{\boldsymbol{s}}_{j,\boldsymbol{\mu}_g},\widehat{\boldsymbol{s}}_{j,\sigma_1},...,\widehat{\boldsymbol{s}}_{j,\sigma_g},\widehat{\boldsymbol{s}}_{j,\gamma_1},...,\widehat{\boldsymbol{s}}_{j,\gamma_g}\right)$. After straightforward algebra, we obtain these components as follows:

$$\widehat{\boldsymbol{s}}_{j,\pi_r} = \frac{\widehat{z}_{rj}}{\widehat{\pi}_r} - \frac{\widehat{z}_{gj}}{\widehat{\pi}_g}, \quad r=1,...,g-1, \tag{4.2}$$

$$\widehat{\boldsymbol{s}}_{j,\boldsymbol{\mu}_i} = \widehat{z}_{ij}\widehat{\Sigma}_i^{-1}\left(\widehat{v}_{1ij}\left(\boldsymbol{y}_j-\widehat{\boldsymbol{\mu}}_i\right)-\widehat{\boldsymbol{\gamma}}_i\right), \tag{4.3}$$

$$\begin{aligned}\widehat{\boldsymbol{s}}_{j,\sigma_i} = {} & \operatorname{vech}\Big(\widehat{z}_{ij}\Big\{-\Big(\widehat{\Sigma}_i^{-1}-\widehat{v}_{1ij}\widehat{\Sigma}_i^{-1}\left(\boldsymbol{y}_j-\widehat{\boldsymbol{\mu}}_i\right)\left(\boldsymbol{y}_j-\widehat{\boldsymbol{\mu}}_i\right)^{\mathsf{T}}\widehat{\Sigma}_i^{-1}-\widehat{v}_{2ij}\widehat{\Sigma}_i^{-1}\widehat{\boldsymbol{\gamma}}_i\widehat{\boldsymbol{\gamma}}_i^{\mathsf{T}}\widehat{\Sigma}_i^{-1}\Big)\\ & +\frac{1}{2}\operatorname{diag}\Big(\widehat{\Sigma}_i^{-1}-\widehat{v}_{1ij}\widehat{\Sigma}_i^{-1}\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)^{\mathsf{T}}\widehat{\Sigma}_i^{-1}-\widehat{v}_{2ij}\widehat{\Sigma}_i^{-1}\widehat{\boldsymbol{\gamma}}_i\widehat{\boldsymbol{\gamma}}_i^{\mathsf{T}}\Big)\Big\}\Big),\end{aligned} \tag{4.4}$$

$$\boldsymbol{s}_{j,\boldsymbol{\gamma}_i} = \widehat{z}_{ij}\widehat{\Sigma}_i^{-1}\left(\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i\right)-\widehat{v}_{2ij}\boldsymbol{\gamma}_i\right). \tag{4.5}$$

Therefore, using these equations we can form the information matrix I_e given in (4.1). After this, the standard errors of the ML estimators $\widehat{\boldsymbol{\Theta}}$ will be found using the square root of the matrix $\widehat{I}_e^{-1}$.

5. APPLICATIONS

This section will illustrate the performance of the proposed mixture model based on a small simulation study and a real data example. All computations for the simulation study and real data example are conducted using an MATLAB R2013a. For all computations, the stopping rule Δ is taken as 10^{-6}. The codes are available upon request.

5.1. Simulation study

In the simulation study, the data set is generated from the following two-component mixtures of MSL distributions:

$$f\left(\boldsymbol{y}_i|\boldsymbol{\Theta}\right) = \pi_1 f_p\left(\boldsymbol{y}_j;\boldsymbol{\mu}_1,\Sigma_1,\boldsymbol{\gamma}_1\right) + \left(1-\pi_1\right) f_p\left(\boldsymbol{y}_j;\boldsymbol{\mu}_2,\Sigma_2,\boldsymbol{\gamma}_2\right),$$

where

$$\boldsymbol{\mu}_i = \left(\mu_{i1},\mu_{i2}\right)^{\mathsf{T}}, \quad \Sigma_i = \begin{bmatrix}\sigma_{i,11} & \sigma_{i,12}\\ \sigma_{i,21} & \sigma_{i,22}\end{bmatrix}, \quad \boldsymbol{\gamma}_i = \left(\gamma_{i1},\gamma_{i2}\right)^{\mathsf{T}}, \quad i=1,2,$$

with the parameter values

$$\boldsymbol{\mu}_1 = (2,2)^{\mathsf{T}}, \quad \boldsymbol{\mu}_2 = (-2,-2)^{\mathsf{T}}, \quad \Sigma_1 = \Sigma_2 = \begin{bmatrix}1.5 & 0\\ 0 & 1.5\end{bmatrix},$$

$$\boldsymbol{\gamma}_1 = (1,1)^{\mathsf{T}}, \quad \boldsymbol{\gamma}_2 = (-1,-1)^{\mathsf{T}}, \quad \pi_1 = 0.6.$$

The sample sizes are set as 500, 1000 and 2000 and the number of replicates (N) is taken as 500. The table contains the bias, standard errors (SEs) and the mean Euclidean distance values of the estimates. The formula of bias is given below:

$$\widehat{\text{bias}}\left(\widehat{\theta}\right) = \bar{\theta} - \theta,$$

where θ is the true parameter value, $\bar{\theta} = \frac{1}{N}\sum_{j=1}^{N}\widehat{\theta}_j$ and $\widehat{\theta}_j$ is the estimate of θ for the j-th simulated data. The mean Euclidean distances of the estimators are computed using the average of the Euclidian norm between the estimates and the true parameter values. For instance, for the mean Euclidean distance of $\widehat{\boldsymbol{\mu}}_i$ will be as follows:

$$|| \, \widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i \, || = \frac{1}{N}\left(\sum_{j=1}^{N}\left(\widehat{\mu}_{ij} - \mu_{ij}\right)^2\right)^{\frac{1}{2}}.$$

The other mean Euclidean distances of other estimates are also obtained in a similar way. The distance for $\widehat{\pi}_i$, on the other hand, will be the mean squared error (MSE). The formula of MSE is given as:

$$\widehat{\text{MSE}}\left(\widehat{\pi}\right) = \frac{1}{N}\sum_{j=1}^{N}\left(\widehat{\pi}_j - \pi\right)^2,$$

where π is the true parameter value, $\widehat{\pi}_j$ is the estimate of π for the j-th simulated data and $\bar{\pi} = \frac{1}{N}\sum_{j=1}^{N}\widehat{\pi}_j$. We calculate the SEs of estimates using the empirical information matrix of the finite mixture model based on the MSL distribution given in Section 4.

Table 1: Bias, SEs and mean Euclidean distance values of the estimates for $n = 500$, 1000 and 2000.

n	Parameter	Component 1				Component 2			
		True	Bias	SE	Distance	True	Bias	SE	Distance
500	π_1	0.6	0.001489	0.122734	0.000533				
	μ_{i1}	2	-0.001200	0.174686	0.201026	-2	-0.014079	0.210878	0.273981
	μ_{i2}	2	0.010775	0.176092		-2	0.031688	0.208902	
	$\sigma_{i,11}$	1.5	-0.018232	0.182958	0.198493	1.5	-0.005309	0.232425	0.254685
	$\sigma_{i,12}$	0	0.002903	0.156863		0	-0.006157	0.170339	
	$\sigma_{i,22}$	1.5	-0.004315	0.188067		1.5	-0.033095	0.224071	
	γ_{i1}	1	-0.001810	0.256792	0.106922	-1	-0.000255	0.291138	0.139768
	γ_{i2}	1	-0.005965	0.255197		-1	-0.016220	0.290819	
1000	π_1	0.6	0.001075	0.082913	0.000236				
	μ_{i1}	2	-0.007428	0.122541	0.147008	-2	-0.001296	0.146090	0.180637
	μ_{i2}	2	-0.006677	0.122956		-2	-0.003134	0.146122	
	$\sigma_{i,11}$	1.5	-0.011271	0.127618	0.136552	1.5	-0.002545	0.155037	0.169090
	$\sigma_{i,12}$	0	-0.008758	0.106796		0	-0.000529	0.116473	
	$\sigma_{i,22}$	1.5	-0.000720	0.128693		1.5	-0.009886	0.155416	
	γ_{i1}	1	0.005704	0.174375	0.078061	-1	-0.001849	0.194632	0.093384
	γ_{i2}	1	0.006670	0.174589		-1	-0.002547	0.192690	
2000	π_1	0.6	0.000593	0.057421	0.000119				
	μ_{i1}	2	-0.002704	0.086564	0.103007	-2	0.004608	0.102633	0.126149
	μ_{i2}	2	0.000165	0.086560		-2	-0.003348	0.103218	
	$\sigma_{i,11}$	1.5	-0.006294	0.088880	0.098289	1.5	-0.008891	0.106630	0.116648
	$\sigma_{i,12}$	0	-0.003700	0.074673		0	-0.000537	0.080978	
	$\sigma_{i,22}$	1.5	0.001142	0.089714		1.5	0.000997	0.108406	
	γ_{i1}	1	0.002448	0.121117	0.056854	-1	-0.003021	0.133069	0.067364
	γ_{i2}	1	0.001445	0.120912		-1	0.000548	0.133948	

Table 1 shows the simulation results for the sample sizes 500, 1000 and 2000. We give the bias, SEs and mean Euclidean distance values of estimates and true parameter values. It can be seen from the table that the proposed model works accurately to obtain the estimates for all the parameters. Furthermore, the mean Euclidian distances get smaller when the sample sizes increase. We observe similar results for the SEs of the estimates. These values decrease as the sample sizes increase. All these findings confirm that the proposed finite mixture model will be an alternative finite mixture model for modelling heterogeneous data with skew and heavy-tail components.

5.2. Real data example

This real data example will investigate the bank data set, which was given in Tables 1.1 and 1.2 by [11] and examined by [19], to model through a skew-symmetric distribution. Concerning this data set, there are six measurements made on 100 genuine and 100 counterfeit old Swiss 1000 franc bills. This data set was also analyzed by [16] to model mixtures of MSN distributions. He used the variables X_1, the width of the right edge, and X_2, the length of the image diagonal, that reveal a bimodal distribution with asymmetric components. Following this, the current study uses Swiss bank data to illustrate the applicability of the finite mixtures of multivariate skew Laplace distributions (FM-MSL) and compares the results with the finite mixtures of multivariate skew normal distributions (FM-MSN). The estimation results are displayed in Table 2 for FM-MSN and FM-MSL. The table contains the ML estimates, standard errors of the estimates for all components, the log-likelihood, the values of the Akaike information criterion (AIC) ([1]) and the Bayesian information criterion (BIC) ([26]).

Table 2: ML estimation results of the Swiss bank data set for FM-MSN and FM-MSL.

	FM-MSN				FM-MSL			
	1		2		1		2	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
π_1	0.504	0.036	—	—	0.521	0.163	—	—
μ_{i1}	130.38	0.122	129.32	0.062	130.20	0.118	129.65	0.076
μ_{i2}	140.06	0.064	141.39	1.125	139.50	0.152	141.76	0.201
$\sigma_{i,11}$	0.068	0.023	0.037	0.016	0.067	0.054	0.104	0.030
$\sigma_{i,12}$	0.051	0.015	−0.012	0.015	0.001	0.037	−0.023	0.043
$\sigma_{i,22}$	0.056	0.027	0.154	0.032	0.371	0.100	0.194	0.218
γ_{i1}	−0.230	0.043	0.494	0.077	−0.017	0.108	0.034	0.060
γ_{i2}	−0.800	0.067	0.177	1.433	0.054	0.154	−0.148	0.198
$\ell(\widehat{\Theta})$			−310.07				**−152.30**	
AIC			650.14				**334.60**	
BIC			699.61				**384.08**	

Additionally, we give results and criterion values for FM-MSN which was computed by [16]. According to information criterion values, the FM-MSL has better fit than the FM-MSN.

Figure 1 displays a scatter plot of the data together with contour plots of the fitted two-component FM-MSL model. From this plot, it can be seen that the proposed mixture model of MSL distributions captures bimodality and asymmetry and provides a satisfactory fit to the data.

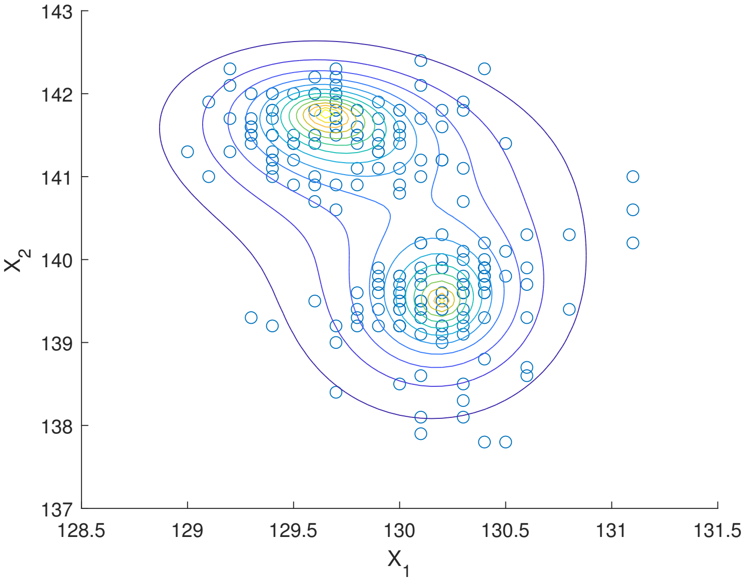


Figure 1: Scatter plot of the Swiss bank data set along with the contour plots of the fitted two-component FM-MSL model.

6. CONCLUSION

In this paper, we have proposed mixtures of MSL distributions and given the EM algorithm to obtain the estimates. A small simulation study has been provided to demonstrate the performance of the proposed mixture model. This shows that the proposed mixture model has accurately estimated the parameters. A real data example has also been offered to compare the mixtures of the MSL distributions with the mixtures of MSN distributions. This comparison proves that the proposed model has the best fit according to the information criterion values. This means that the proposed model can be used as an alternative mixture model to the mixtures of MSN distributions.

ACKNOWLEDGMENTS

We wish to thank the anonymous referee and the Associate Editor, whose comments and suggestions have helped us to improve this paper greatly.

REFERENCES

[1] Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* In "Proceeding of the Second International Symposium on Information Theory" (B.N. Petrov and F. Caski, Eds.), Akademiai Kiado, Budapest, 267–281.

[2] Arellano-Valle, R.B. and Genton, M.G. (2005). On fundamental skew distributions, *Journal of Multivariate Analysis*, **96**(1), 93–116.

[3] Arslan, O. (2010). An alternative multivariate skew Laplace distribution: properties and estimation, *Statistical Papers*, **51**(4), 865–887.

[4] Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**(2), 367–389.

[5] Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution, *Biometrika*, **83**(4), 715–726.

[6] Basford, K.E.; Greenway, D.R.; McLachlan, G.J. and Peel, D. (1997). Standard errors of fitted means under normal mixture, *Computational Statistics*, **12**, 1–17.

[7] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*, Springer, Singapore.

[8] Cabral, C.R.B.; Lachos, V.H. and Prates, M.O. (2012). Multivariate mixture modeling using skew-normal independent distributions, *Computational Statistics & Data Analysis*, **56**(1), 126–142.

[9] Dempster, A.P.; Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

[10] Dias, J.G. and Wedel, M. (2004). An empirical comparison of EM, SEM and MCMC performance for problematic gaussian mixture likelihoods, *Statistics and Computing*, **14**, 323–332.

[11] Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics, a Practical Approach*, Cambridge University Press, Cambridge.

[12] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer, New York.

[13] Gupta, A.K. (2003). Multivariate skew t-distribution, *Statistics: A Journal of Theoretical and Applied Statistics*, **37**(4), 359–363.

[14] Gupta, A.K.; González-Farías, G. and Domínguez-Molina, J.A. (2004). A multivariate skew normal distribution, *Journal of Multivariate Analysis*, **89**(1), 181–190.

[15] Hartigan, J.A. and Wong, M.A. (1979). Algorithm AS 136: a k-means clustering algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **28**(1), 100–108.

[16] Lin, T.I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models, *Journal of Multivariate Analysis*, **100**, 257–265.

[17] Lin, T.I. (2010). Robust mixture modeling using multivariate skew t distributions, *Statistics and Computing*, **20**(3), 343–356.

[18] Lin, T.I.; Ho, H.J. and Lee, C.R. (2014). Flexible mixture modelling using the multivariate skew-t-normal distribution, *Statistics and Computing*, **24**(4), 531–546.

[19] Ma, Y. and Genton, M.G. (2004). Flexible class of skew-symmetric distribtions, *Scandinavian Journal of Statististics*, **31**, 459–468.

[20] McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Application to Clustering*, Marcel Dekker, New York.

[21] McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*, Wiley, New York.

[22] Naik, D.N. and Plungpongpun, K. (2006). *A Kotz-type distribution for multivariate statistical inference.* In "Advances in Distribution Theory, Order Statistics, and Inference" (pp. 111–124), Birkhäuser, Boston.

[23] Peel, D. and McLachlan, G.J. (2000). Robust mixture modelling using the t distribution, *Statistics and computing*, **10**(4), 339–348.

[24] Pyne, S.; Hu, X.; Wang, K.; Rossin, E.; Lin, T.I.; Maier, L.; Baecher-Allan, C.; McLachlan, G.J.; Tamayo, P.; Hafler, D.A.; De Jager, P.L. and Mesirov, J.P. (2009). Automated high-dimensional flow cytometric data analysis, *Proc. Natl. Acad. Sci. USA*, **106**, 8519–8524.

[25] Sahu, S.K.; Dey, D.K. and Branco, M.D. (2003). A new class of multivariate skew distributions with applications to Bayesian regression models, *Canadian Journal of Statistics*, **31**(2), 129–150.

[26] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461–464.

[27] Titterington, D.M.; Smith, A.F.M. and Markov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.

REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 47–69

CHOICE OF SMOOTHING PARAMETER FOR KERNEL TYPE RIDGE ESTIMATORS IN SEMIPARAMETRIC REGRESSION MODELS

Authors: Ersin Yilmaz
– Department of Statistics, Mugla Sitki Kocman University,
Mugla, Turkey
yilmazersin13@hotmail.com

Bahadir Yuzbasi
– Department of Econometrics, Inonu University,
Malatya, Turkey
b.yzb@hotmail.com

Dursun Aydin
– Department of Statistics, Mugla Sitki Kocman University,
Mugla, Turkey
duaydin@hotmail.com

Received: August 2017 Revised: July 2018 Accepted: September 2018

Abstract:

- This paper concerns kernel-type ridge estimators of parameters in a semiparametric model. These estimators are a generalization of the well-known Speckman's approach based on kernel smoothing method. The most important factor in achieving this smoothing method is the selection of the smoothing parameter. In the literature, many selection criteria for comparing regression models have been produced. We will focus on six selection criterion improved version of Akaike information criterion (AIC$_c$), generalized cross-validation (GCV), Mallows' C$_p$ criterion, risk estimation using classical pilots (RECP), Bayes information criterion (BIC), and restricted maximum likelihood (REML). Real and simulated data sets are considered to illustrate the key ideas in the paper. Thus, suitable selection criterion are provided for optimum smoothing parameter selection.

Keywords:

- *semiparametric model; kernel smoothing; ridge type estimator; smoothing parameter generalized cross-validation.*

AMS Subject Classification:

- 62G08, 62J07, 65C60.

1. INTRODUCTION

Let us consider the following semiparametric regression model:

$$y_i = \mathbf{x}_i \beta + f(t_i) + \varepsilon_i, \quad 1 \leq i \leq n, \tag{1.1}$$

where the y_i's are observations, the $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ are known p-vectors with $p < n$ and t_i's have bounded support, say the unit interval and have been reordered so that $t_i \leq t_2 \leq \cdots \leq t_n$. $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^\top$ is an unknown p-dimensional vector of parameters, $f(\cdot)$ is unknown function and ε_i's are the random error terms assumed to be uncorrelated with mean zero and variance σ^2. Note that f symbolizes the smooth part of the model and assume that it shows the unparameterized functional relationship.

The model (1.1) is also called as a partially linear model, due to the connection with the classical linear model (see [8]). In matrix-vector form, the model (1.1) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon}, \tag{1.2}$$

where $\mathbf{y} = (y_1, ..., y_n)^\top$, $\mathbf{X} = (x_1, ..., x_n)^\top$, $\mathbf{f} = (f(t_1), f(t_2), ..., f(t_n))^\top$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^\top$. The key idea is to estimate the unknown parameter vector $\boldsymbol{\beta}$, the nonparametric function $f(t)$ and the mean vector $\mu = \mathbf{X}\boldsymbol{\beta} + \mathbf{f}$ based on the data $y_i, \mathbf{x}_i, t_i$. Note that semiparametric models have received a considerable attention in the past two decades. One of the most important reasons for this is that these models are more flexible than the standard linear model because they combine both parametric and nonparametric components. In this context, a number of authors have studied the model (1.1), including Green and Silverman [12], Speckman [30], Eubank *et al.* [9], Schimek [28], Liang [21], Aydin *et al.* [3], Ahmed [1] and among others.

In many regression problems, there is a perfect or exact relationship between the columns of $\mathbf{X}$. In this case, multicollinearity is a serious problem which can dramatically influence the effectiveness of a regression model. The multicollinearity results in large variances and covariances of the parameter estimates and may lead to lack of statistical significance of individual parameters even though the overall model may be significant. For the purposes of the paper, we will employ the kernel type ridge regression procedure that is designed to deal with multicollinearity in semiparametric regression.

Concerning the collinear data, Gibbons [11] introduced a simulation study of ridge estimators for parametric linear models. Kibria [18] proposed some new estimators based on generalized ridge regression approach and considered some methods to estimate ridge parameter. For the linear regression models Muniz and Kibria [23] reviewed and proposed some estimators based on Kibria [18]. Key references for semiparametric regression based on kernel smoothing are Robinson [26] and Speckman [30]. It should be noted that Robinson [26] introduced an estimator for parametric part of a semiparametric model when nonparametric component is stochastic and of arbitrary dimension. Speckman [30] discussed two estimation method, one related to partial smoothing spline and the other modified by partial residual, in estimating the components of a semiparametric model and examined the asymptotic behaviours for both methods. Chen [6] studied the parametric component of the partial linear model. Foucart [10] used the ridge estimators on partial linear models for combat multicollinearity. Ridge estimation of a semiparametric regression model and a comparison of this ridge estimation with two steps estimation are introduced by Hu [15].

Roozbeh *et al.* [27], Yuzbasi and Ahmed [36] and Yuzbasi *et al.* [37] proposed a semiparametric ridge regression estimator for partially linear models. More recently, semiparametric regression models based on different selection methods were studied and compared by Aydin [2]. Lastly, the pretest and shrinkage ridge regression estimators based on smoothing spline approach for partially linear models was studied by Yuzbasi [35] and modified estimators in semiparametric regression models based on right censored data is studied by Aydin and Yilmaz [4].

The main difference of our study is that we consider various kernel type ridge estimators to estimate the components of a semiparametric regression model with collinear data. The most important issue in this problem is to determine an amount of smoothing. In order to specify an optimum smoothing parameter we use six different selection criteria under simulated and real data settings. The basic idea is to find a useful selection criteria that provides a good estimation of the model (1.1) based on multicollinear data. Due to smoothing parameter selection criteria, we provide a comparison of the different ridge type estimators. To the best of our knowledge, the studies in the literature often address the problem of comparing different ridge type estimators and the selection of ridge parameter, but such a study that includes kernel type ridge estimators based on different selection criteria has not yet been conducted. This paper is organized as follows. Estimation based on kernel smoothing is examined in Section 2. In Section 3, the kernel type ridge estimators in semiparametric models are discussed. Statistical properties of the ridge type estimators are examined in Section 4. Section 5 reviews six different smoothing parameter selection methods. Section 6 compares these methods via a real example. In Section 7, a simulation study is given. Finally, concluding remarks are presented in Section 8. Supplemental technical materials are relegated to the Appendix.

2. ESTIMATION BASED ON KERNEL SMOOTHING

First we consider the nonparametric estimation of the unknown regression function $f(t)$ in (1.1). For convenience, we assume that $\boldsymbol{\beta}$ in equation (1.1) is known. In this case, the relationship between $y_i - \mathbf{X}_i\boldsymbol{\beta}$ and t_i can be denoted by

$$(y_i - \mathbf{X}_i\boldsymbol{\beta}) = f(t_i) + \varepsilon_i, \quad i = 1, ..., n. \tag{2.1}$$

Equation (2.1) can be considered as equivalent to the nonparametric part of a semiparametric model. As expressed in the study of Speckman [30], this leads to the Nadaraya-Watson estimator proposed by Nadaraya [24] and Watson [34], and this is also referred to as the kernel estimator:

$$\hat{f}_\lambda(t) = \sum_{i=1}^{n} w_{i\lambda}(t_i)(y_i - \mathbf{X}_i\boldsymbol{\beta}) = \mathbf{W}_\lambda(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.2}$$

where λ is a smoothing parameter (or bandwidth) and $\mathbf{W}_\lambda$ is a kernel smoother matrix with j-th entries $w_{i\lambda}$, given by

$$w_{i_\lambda}(t_i) = K\left(\frac{t - t_i}{\lambda}\right) \Bigg/ \sum_{i=1}^{n} K\left(\frac{t - t_i}{\lambda}\right) = K(u_i) / \sum_{i} K(u_i). \tag{2.3}$$

As shown in (2.1), kernel smoothing (or regression) uses the appropriate weights $w_{i\lambda}(t)$ to estimate $f(t)$. The weights given to the observations t_i are directed by the kernel function $K(u)$ with a smoothing parameter λ, which controls the size of the neighborhood around t [31]. Note that $K(u)$ in (2.3) is a kernel or weight function such that $\int K(u)du = 1$, and $K(u) = K(-u)$. The kernel function is selected to give most weight to observations close to and least weight to observations far from t.

Using the matrix and vector form of the model (1.2), we can obtain the following partial residuals in matrix form:

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \hat{\mathbf{f}} = (\mathbf{I} - \mathbf{W}_\lambda)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}, \tag{2.4}$$

where $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X}$ and $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y}$. Thus, we obtain a transformed set of data based on kernel residuals. Considering these partial residuals for the vector $\boldsymbol{\beta}$ yields the following weighted least squares (WLS) criterion:

$$\mathrm{WLS}(\boldsymbol{\beta};\boldsymbol{\lambda}) = ((\mathbf{I} - \mathbf{W}_\lambda)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))^{\mathsf{T}} ((\mathbf{I} - \mathbf{W}_\lambda)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) = \left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)' \left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right). \tag{2.5}$$

In analogy with ordinary least squares, the solution to the criterion $\mathrm{WLS}(\boldsymbol{\beta};\lambda)$ given in equation (2.5) is easily seen to be

$$\hat{\boldsymbol{\beta}}_p = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}'\tilde{\mathbf{y}}. \tag{2.6}$$

Moreover, according to the equation (2.3) updating the steps for $f(t)$ simplifies to

$$\hat{f}_\lambda(t) = \sum_{i=1}^{n} K\left(\frac{t-t_i}{\lambda}\right) \Big/ \sum_{i=1}^{n} K\left(\frac{t-t_i}{\lambda}\right) \left(y_i - X_i\hat{\boldsymbol{\beta}}_p\right). \tag{2.7}$$

Equation (2.7) can also be written in a matrix form as

$$\hat{\mathbf{f}}_p = W_\lambda \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_p\right). \tag{2.8}$$

Our estimate of μ_p is then

$$\mu_p = \mathbf{X}\hat{\boldsymbol{\beta}}_p + \hat{\mathbf{f}}_p = \mathbf{X}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} + \mathbf{W}_\lambda(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_p)\right)$$

for

$$\mathbf{H}_p = \mathbf{W}_\lambda + \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right). \tag{2.9}$$

Equations (2.6) and (2.8) are hierarchical in the sense that the adjustment is made for t. Adjusting for $\mathbf{X}$ first would produce a different estimator. One advantage of $\hat{\beta}_p$ is that, there is no iteration in calculation of $\hat{\beta}_p$ even if a non-linear smoother is used. As a result the approach requires only a standard regression routine if a computation of the $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ has been done with smoother matrix $\mathbf{W}_\lambda$.

3. KERNEL TYPE RIDGE ESTIMATORS IN SEMIPARAMETRIC MODELS

Ridge regression has been proposed by Hoerl and Kennard [13], [14] as a solution to the multicollinearity problem. It is well known that a ridge estimator provides a slight improvement on the estimations of partial regression coefficients when the column vectors of the matrix $\mathbf{X}$ in a linear model $y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ are highly correlated. Generally, the linear model can be written in centered and scaled form. For notational convenience, we do not consider an explicitly centered and scaled model here.Then, the ridge estimate of $\boldsymbol{\beta}$ for some $k > 0$ can be written as

$$\hat{\boldsymbol{\beta}}_r(k) = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}, \tag{3.1}$$

where $\mathbf{I}$ is $p \times p$ identity matrix and k is the shrinkage parameter, whose value is specified by the researcher. When $k = 0$ the ridge estimate corresponds to the least squares estimate. To fit the model (1.1) to data, we can use ridge regression that shrinks the regression coefficients by imposing a penalty on their size. This procedure can be related to the idea of hints due to Speckman [30], where the parameter vector β is obtained by minimizing the penalized residual sum of squares criterion

$$\text{PRSS}(\boldsymbol{\beta}; \lambda) = \sum_{i=1}^{n}\left(\tilde{y}_i - \tilde{X}_i\boldsymbol{\beta}\right)^2 + k\sum_{j=1}^{n}\beta_j^2 = \sum_{i=1}^{n}\left(\tilde{y}_i - \tilde{X}_i\boldsymbol{\beta}\right)^2 + \sum_{j=1}^{n}(0 - k\beta_j)^2, \tag{3.2}$$

where $k \geq 0$ is the shrinkage parameter that controls the magnitude of the penalty term. The basic idea is to recast the linear regression problem as a linear smoother problem for another data set. This means that if artificial data having response value zero are introduced, then a fitting procedure can be forced to shrink the coefficients toward zero.

In matrix and vector form, equation (3.2) can be rewritten as

$$\text{PRSS}\left(\boldsymbol{\beta}; \boldsymbol{\lambda}\right) = \left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right)'\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\right) + k\|0 - \boldsymbol{\beta}\|^2. \tag{3.3}$$

The main objective is to find parameter vector $\boldsymbol{\beta}$ such that equation (3.3) is as small as possible. The following theorem gives the estimates.

Theorem 3.1. *Let $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}$ where $\tilde{\boldsymbol{\varepsilon}} = \tilde{\boldsymbol{f}} + \boldsymbol{\varepsilon}^*$, $\tilde{\boldsymbol{f}} = (\boldsymbol{I} - \mathbf{W}_\lambda)\,\boldsymbol{f}$ and $\boldsymbol{\varepsilon}^* = (\boldsymbol{I} - \mathbf{W}_\lambda)\,\boldsymbol{\varepsilon}$. Also, $\tilde{\mathbf{X}}$ is a $n \times p$ matrix and $\tilde{\mathbf{y}}$ is a $n \times 1$ vector, as defined in (2.8), respectively. If $\mathbf{W}_\lambda$ is an arbitrary smoother matrix then the ridge regression estimates may be computed by augmenting data*

$$\boldsymbol{X}_A = \begin{bmatrix}\tilde{\mathbf{X}} \\ \sqrt{k}\boldsymbol{I}_p\end{bmatrix} \qquad \textit{and} \qquad \tilde{\boldsymbol{y}}_A = \begin{bmatrix}\tilde{\mathbf{y}} \\ 0\end{bmatrix}.$$

The kernel type ridge estimator for $\boldsymbol{\beta}$ is indicated by $\hat{\boldsymbol{\beta}}_R(k)$ and given by

$$\hat{\boldsymbol{\beta}}(k) = \left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + k\boldsymbol{I}_p\right)^{-1}\tilde{\boldsymbol{X}}'\tilde{\mathbf{y}}. \tag{3.4}$$

Proof of the Theorem 3.1 is given in Appendix A.1.

As in discussion Theorem 3.1, $\hat{\boldsymbol{\beta}}_R(k)$ is the ridge type estimator of the vector $\boldsymbol{\beta}$ in the model (1.2). When $k = 0$, the ridge estimate reduces to a Speckman estimate problem in the equation (2.8). Also, it is seen that there is a formal similarity between the equation (3.3) and ridge estimator of the linear regression model. Combining equations (3.3) and (3.4) we obtain the estimator of $\mathbf{f}$ as

$$\hat{\mathbf{f}}_R(k) = \mathbf{W}_\lambda\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R(k)\right). \tag{3.5}$$

Thus the estimator (3.5) is defined as the kernel type ridge estimator for the unknown function $\mathbf{f}$ in the model (1.2).

4. FURTHER PROPERTIES OF THE ESTIMATORS

It is easily seen that equation (3.4) is identical to

$$\hat{\boldsymbol{\beta}}_R(k) = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1} \tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta}_p = \left[\mathbf{I}_p + k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right]^{-1} \hat{\boldsymbol{\beta}}_p, \tag{4.1}$$

where $\hat{\boldsymbol{\beta}}_p$ is the Speckman estimate, as defined in (2.8). Using the fact $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, the equation (3.4) also becomes

$$\hat{\boldsymbol{\beta}}_R(k) = \left[\mathbf{I}_p + k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right]^{-1} \hat{\boldsymbol{\beta}}_p = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}' + k\mathbf{I}_p\right)^{-1} \tilde{\mathbf{X}}'\tilde{\mathbf{y}}. \tag{4.2}$$

It appears from (4.2) that the ridge type estimator is clearly biased, since

$$\left[\mathbf{I}_p + k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\right]^{-1} \neq \mathbf{I}_p.$$

Hoerl and Kennard [13], [14] used this interpretation as a basis for the definition of the $\hat{\boldsymbol{\beta}}_R(k)$ with $k \geq 0$, the shrinkage parameter that controls the size of coefficients. Also, equation (4.2) can be viewed as the Speckman estimator for $k = 0$.

Using the abbreviation

$$\mathbf{G}_k = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}. \tag{4.3}$$

Moments of the kernel type ridge estimator can be obtained as follows:

$$E\left(\hat{\boldsymbol{\beta}}_R(k)\right) = \mathbf{G}_k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{X}}'\tilde{\mathbf{f}}\right) = \boldsymbol{\beta} - k\boldsymbol{G}_k\boldsymbol{\beta} + \mathbf{G}_k\tilde{\mathbf{X}}\mathbf{f}, \tag{4.4}$$

$$\operatorname{Bias}\left(\hat{\boldsymbol{\beta}}_R(k)\right) = \mathbf{G}_k\tilde{\mathbf{X}}'\tilde{\mathbf{f}} - k\mathbf{G}_k\boldsymbol{\beta}, \tag{4.5}$$

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right) = \sigma^2\mathbf{G}_k\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right). \tag{4.6}$$

The implementation details of Equations (4.4)–(4.6) are given in Appendix A.2. It should be noted that in practice $\boldsymbol{\beta}$ and σ^2 stated in equations above are replaced by their estimated values.

4.1. Estimating the error variance

The error variance σ^2 is usually unknown. In practice, σ^2 needs to be estimated. In a general semiparametric regression model, the estimate of variance σ^2 can be found by the residual sum of squares

$$\begin{aligned}\text{RSS} &= \left(\mathbf{y} - \hat{\mathbf{y}}'\right)'\left(\mathbf{y} - \hat{\mathbf{y}}'\right) \qquad \text{where} \quad \{\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_R(k) + \hat{\mathbf{f}}_R(k)\} \\ &= \left(\mathbf{y} - \left(\mathbf{X}\hat{\boldsymbol{\beta}}_R(k) + \hat{\mathbf{f}}_R(k)\right)\right)'\left(\mathbf{y} - \left(\mathbf{X}\hat{\boldsymbol{\beta}}_R(k) + \hat{\mathbf{f}}_R(k)\right)\right).\end{aligned}$$

Substituting $\hat{\mathbf{y}} = \left(\mathbf{X}\hat{\boldsymbol{\beta}}_R(k) + \hat{\mathbf{f}}_R(k)\right) = \mathbf{H}_\lambda \mathbf{y}$, we obtain

$$\text{RSS} = (\mathbf{y} - \mathbf{H}_\lambda \mathbf{y})' (\mathbf{y} - \mathbf{H}_\lambda \mathbf{y}) = \parallel (\mathbf{I} - \mathbf{H}_\lambda)\, \mathbf{y} \parallel_2^2, \tag{4.7}$$

where $\mathbf{H}_\lambda$ is called the smoother matrix which depends on $\lambda > 0$. Note that the matrix $\mathbf{H}_\lambda$ is used to estimate the fitted values of the model in (1.2) and is expressed as

$$\mathbf{H}_\lambda = \mathbf{W}_\lambda + (\mathbf{I}_p - \mathbf{W}_\lambda)\, \tilde{\mathbf{X}} \mathbf{G}_k \tilde{\mathbf{X}}'. \tag{4.8}$$

Furthermore, the expected value of RSS is

$$E(\text{RSS}) = \sigma^2 \left[n - \text{tr}(2\mathbf{H}_\lambda - \mathbf{H}_\lambda^2)\right] + E(\mathbf{y}')(\mathbf{I} - \mathbf{H}_\lambda)'(\mathbf{I} - \mathbf{H}_\lambda)E(\mathbf{y}),$$

where the first term measures the variance, while the second term measures bias, respectively. Detailed implementations of the equation (4.7) and E(RSS) are given in Appendix A.3.

Hence, similar to ordinary least squares regression,estimation of the error variance can be defined by

$$\hat{\sigma}^2 = \frac{\text{RSS}}{\text{tr}\,(\mathbf{I} - \mathbf{H}_\lambda)^2} = \frac{\parallel (\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y} \parallel_2^2}{n - p}, \tag{4.9}$$

where $\text{tr}\,(\mathbf{I} - \mathbf{H}_\lambda)^2 = n - \text{tr}\left(2\mathbf{H}_\lambda - \mathbf{H}_\lambda' \mathbf{H}_\lambda\right) = n - p$ is the residual degrees of freedom. From equation (4.9) see that the degrees of freedom for RSS is also known as the number of total observations minus total number of the parameters in the model.

To show that $\hat{\sigma}^2$ is biased or unbiased for σ^2, $E(\hat{\sigma}^2)$ is found as

$$E(\hat{\sigma^2}) = \frac{1}{n-p} E\left(\parallel (\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y} \parallel_2^2\right) = \frac{1}{n-p} E(\text{RSS}).$$

The expected value of E(RSS) implies that the estimator of σ^2 in equation (4.9) has a positive bias. However, it should be noted that the (4.9) yields asymptotically negligible bias. Considering this point of view, it is noteworthy that $\hat{\sigma}^2$ is equivalent to mean square error (MSE) which is a widely used criterion for measuring the quality of estimation (see Speckman [30]).

4.2. Measuring the risk and performance efficiency

This section investigates the superiority of a biased estimator $\hat{\boldsymbol{\beta}}_{R1}(k)$ with respect to any other biased estimator $\hat{\boldsymbol{\beta}}_{R2}(k)$. It is well known that ridge type estimators are biased and need to measure the loss of information. Generally, the expected loss of a vector $\hat{\boldsymbol{\beta}}_R(k)$ estimator is measured by risk (i.e., the bias-variance decomposition). Our task is now to approximate the risk in the models in (1.1) or (1.2). Such approximations have the advantage of being simpler to optimize the practical selection of smoothing parameters. For convenience, we will work with the scalar valued mean dispersion error.

Definition 4.1. The risk is closely related to the matrix valued mean dispersion error (MDE) of an estimator $\hat{\boldsymbol{\beta}}_R(k)$ of $\boldsymbol{\beta}$. The scalar valued version of the MDE matrix is specified as

$$\mathrm{SMDE}\left(\hat{\boldsymbol{\beta}}_R(k), \boldsymbol{\beta}\right) = E\left(\hat{\boldsymbol{\beta}}_R(k) - \boldsymbol{\beta}\right)'\left(\hat{\boldsymbol{\beta}}_R(k) - \boldsymbol{\beta}\right) = \mathrm{tr}\left(\mathrm{MDE}\left(\hat{\boldsymbol{\beta}}_R(k) - \boldsymbol{\beta}\right)\right).$$

Lemma 4.1. *Consider different estimators $\hat{\beta}_{jR}(k)$ of β_j.The mean dispersion error (*MDE*) of these estimators is the sum of the covariance matrix and the squared bias:*

$$E\left(\| \hat{\boldsymbol{\beta}}_R(k) - \boldsymbol{\beta} \|^2\right) = \sum_{j=1}^{k} E\left(\hat{\beta}_{jR}(k) - \beta_j\right)^2 = \mathrm{tr}\left[\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right] + \left[\mathrm{Bias}\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right]^2.$$

Note that $\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right)$ *is the covariance matrix of* $\hat{\boldsymbol{\beta}}_R(k)$ *and its trace can be illustrated as* $\mathrm{tr}\left(\sum_{j=1}^{p} \mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{jR}(k)\right)\right)$.

For the proof, see Appendix A.4.

Applying the equations (4.4), (4.5) and (4.6), we obtain

$$E\left[\left(\hat{\boldsymbol{\beta}}_R(k) - \boldsymbol{\beta}\right)^2\right] = \sigma^2 \mathbf{G}_k \tilde{\mathbf{X}}' \left(\mathbf{I} - \mathbf{W}_\lambda\right)^2 \tilde{\mathbf{X}} \mathbf{G}_k + \mathbf{G}_k \left(\tilde{\mathbf{X}}'\tilde{\mathbf{f}} - k\boldsymbol{\beta}\right)\left(\tilde{\mathbf{X}}'\tilde{\mathbf{f}} - k\boldsymbol{\beta}\right)' \mathbf{G}_k. \tag{4.10}$$

As stated in Definition 4.1, the MDE matrix decomposes into a sum of the squared bias and covariance of the estimator. Also, it can be interpreted as the mean Euclidean distance between the vectors $\hat{\boldsymbol{\beta}}_R(k)$ and $\boldsymbol{\beta}$. Thus, from Definition 4.1, the MDE matrix is written as

$$\mathrm{MDE}\left(\hat{\boldsymbol{\beta}}_R(k), \boldsymbol{\beta}\right) = \mathbf{G}_k \left(\sigma^2 \tilde{\mathbf{X}}' \left(\mathbf{I} - \mathbf{W}_\lambda\right)^2 \tilde{\mathbf{X}} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{f}} - k\boldsymbol{\beta}\right)\left(\tilde{\mathbf{X}}'\tilde{\mathbf{f}} - k\boldsymbol{\beta}\right)'\right) \mathbf{G}_k. \tag{4.11}$$

As in Definition 4.1, the scalar valued version of the MDE matrix in (4.11) is given by

$$\begin{aligned} \mathrm{SMDE}\left(\hat{\boldsymbol{\beta}}_R(k), \boldsymbol{\beta}\right) &= \mathrm{tr}\{\mathrm{MDE}\left(\hat{\boldsymbol{\beta}}_R(k), \boldsymbol{\beta}\right)\} \\ &= \mathrm{tr}\{\mathbf{G}_k \left(\sigma^2 \tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)^2 \tilde{\mathbf{X}} + (\tilde{\mathbf{X}}'\tilde{\mathbf{f}} - k\boldsymbol{\beta})(\tilde{\mathbf{X}}'\tilde{\mathbf{f}} - k\boldsymbol{\beta})'\right) \mathbf{G}_k\}. \end{aligned} \tag{4.12}$$

Hence, we can compare the quality of two estimators by looking at the ratio of their SMDE in (4.12). This ratio gives the following definition concerning the superiority of any two estimators.

Definition 4.2. The relative efficiency of an estimator $\hat{\boldsymbol{\beta}}_{R1}(k)$ compared to another estimator $\hat{\boldsymbol{\beta}}_{R2}(k)$ is obtained by the ratio,

$$\mathrm{RE}\left(\hat{\boldsymbol{\beta}}_{R1}(k), \hat{\boldsymbol{\beta}}_{R2}(k)\right) = \frac{R\left(\hat{\boldsymbol{\beta}}_{R2}(k), \boldsymbol{\beta}\right)}{R\left(\hat{\boldsymbol{\beta}}_{R1}(k), \boldsymbol{\beta}\right)} = \frac{\mathrm{SMDE}\left(\hat{\boldsymbol{\beta}}_{R2}(k)\right)}{\mathrm{SMDE}\left(\hat{\boldsymbol{\beta}}_{R1}(k)\right)}, \tag{4.13}$$

where $R(\cdot)$ denotes the scalar risk that is equivalent to the equation (4.12). $\hat{\boldsymbol{\beta}}_{R2}(k)$ is said to be more efficient than $\hat{\boldsymbol{\beta}}_{R1}(k)$ if $\mathrm{RE}\left(\hat{\boldsymbol{\beta}}_{R1}(k), \hat{\boldsymbol{\beta}}_{R2}(k)\right) < 1$.

5. CHOOSING THE SMOOTHING PARAMETER

The main idea of this paper is how to select the smoothing parameter expressed in a penalized residual sum of squares criterion (3.3). Our task is to select an optimum value of the λ. In practice, this can be achieved by using smoothing parameter selection criteria. A reasonable value of λ can be chosen to minimize the mentioned criteria. Examples of the most widely used selection methods are summarized as follows:

GCV Criterion: The generalized cross validation (GCV) score is specified by (see Craven and Wahba, [7])

$$\mathrm{GCV}(\lambda) = n^{-1} \parallel (\mathbf{I}-\mathbf{H}_\lambda)\mathbf{y} \parallel^2 \Big/ \left[n^{-1}\,\mathrm{tr}(\mathbf{I}-\mathbf{H}_\lambda)\right]^2 ,$$

where $\mathbf{H}_\lambda$, as is defined in (4.8), is the smoother matrix based on λ.

$\mathbf{C}_p$ Criterion: This criterion proposed by Mallows [22] is aimed to provide an estimate of the MSE in (4.9) scaled by σ^2, and given as

$$\mathrm{C}_p(\lambda) = \frac{1}{n}\{\parallel (\mathbf{H}_\lambda-\mathbf{I})\mathbf{y} \parallel^2 + 2\sigma^2\,\mathrm{tr}(\mathbf{H}_\lambda) - \sigma^2\} = \frac{1}{n}\{\parallel \mathbf{y}-\hat{\mathbf{f}}_\lambda \parallel^2 + 2\sigma^2\,\mathrm{tr}(\mathbf{H}_\lambda) - \sigma^2)\}.$$

If σ^2 is unknown, in practice an estimation for σ^2 can be provided by

$$\hat{\sigma}^2 = \hat{\sigma}^2_{\hat{\lambda}_p} = \parallel (\mathbf{H}_{\hat{\lambda}_p}-\mathbf{I})\mathbf{y} \parallel^2 \Big/ \mathrm{tr}\left(\mathbf{I}-\mathbf{H}_{\lambda_p}\right),$$

where $\hat{\lambda}$ is an estimate of λ pre-chosen with any of the selection criterion (for example GCV). For details, see Liang [21], Mallows [22] and Wahba [33].

$\mathbf{AIC_c}$ Criterion: Notice that the classical Akaike information criterion tends to overfit when the sample size is relatively small. Hurvich *et al.* [16] suggested an improved version, called $\mathrm{AIC_c}$, which is defined by

$$\mathrm{AIC_c}(\lambda) = 1 + \log\left[\parallel (\mathbf{H}_\lambda-\mathbf{I})\mathbf{y} \parallel^2 \Big/ n\right] + \left[2\{\mathrm{tr}(\mathbf{H}_\lambda)+1\} \Big/ n - \mathrm{tr}(\mathbf{H}_\lambda) - 2\right].$$

BIC Criterion: Schwarz [29] improved the Bayesian information criterion (BIC) by using Bayes estimators. Thus, the BIC is also called Schwarz Information Criterion (SIC). The criterion is expressed as

$$\mathrm{BIC}(\lambda) = 1/n \parallel (\mathbf{I}-\mathbf{H}_\lambda)\mathbf{y} \parallel^2 + (\log(n)/n)\,\mathrm{tr}(\mathbf{H}_\lambda).$$

RECP Criterion: Risk estimation criteria (RECP) measures the distance between $\mathbf{f}$ and $\hat{\mathbf{f}}_\lambda$. By direct calculation, the $\mathrm{RECP}(\hat{\lambda}_p)$ score is defined as

$$\mathrm{RECP}(\lambda_p) = 1/n\{\parallel (\mathbf{H}_\lambda-\mathbf{I})\hat{\mathbf{f}}_{\lambda_p} \parallel^2 + \hat{\sigma}^2_{\lambda_p}\,\mathrm{tr}(\mathbf{H}_\lambda\mathbf{H}^{\mathsf{T}}_\lambda)\} = 1/nE \parallel \mathbf{f}-\hat{\mathbf{f}}_{\lambda_p} \parallel^2,$$

where $\hat{\sigma}^2_{\lambda_p}$ and $\hat{\mathbf{f}}_{\lambda_p}$ are the appropriate *pilot estimates* for σ^2 and $\mathbf{f}$, respectively. The pilot λ_p selected by classical methods is used for computation of the pilot estimates (see Lee [19], [20]).

REML Criterion: The restricted maximum likelihood (REML) criterion motivates treating λ as a variance parameter. The REML and GCV have a similar form and provide identical values. Moreover, the derivatives of both the REML and the GCV with respect to λ can be determined quite naturally in a common form (see Reiss and Ogden [25]). The REML score can be specified as

$$\text{REML}(\lambda) = \| (\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y} \|^2 \Big/ n - \text{tr}(\mathbf{H}_\lambda).$$

5.1. Comparisons of computational times

In this paper, we discuss different parameter selection techniques proposed in the literature. Generally, they differ in the amount of computational time as well as a priori information required. The four selection methods GCV, AIC$_c$, BIC, and REML need approximately the same computational time for finding their corresponding smoothing parameter λ, as their computations only require one numerical minimization problem. From computational perspective, a causing difficulty term is $\text{tr}(\mathbf{H}_\lambda)$, which takes $O(n^2)$ operations to assess directly, for each set of smoothing parameters. Compared to these four methods, both C$_p$ and RECP require a longer computation time, as they need an estimate of parameter λ pre-chosen with a selection criterion, such as GCV. So, there are two numerical minimization in computations of C$_p$ and RECP. However, it should be noted that some calculations are unnecessary for these two numerical minimizations. For this reason, when careful programming is made, the overall calculation time will not be doubled.

6. REAL DATA EXAMPLE

In this study, to illustrate how ridge type kernel method works on real data, power plant data has been used. The power plant dataset includes 500 data points collected from a Combined Cycle Power Plant. The goal is to predict the net hourly electrical energy output (EP) of the plant from the features consisting of hourly average ambient variables such as temperature (T), ambient pressure (AP), relative humidity (RH) and exhaust vacuum (V).

Tufekci [32] has used the dataset for prediction of electrical power output of a base load operated combined cycle power plant using machine learning methods. Also, Kaya *et al.* [17] have used this data in their study called “Local and Global Learning Methods for Predicting Power of a Combined Gas and Steam Turbine”.

In order to explain the variables clearly, their intervals and units are defined as follows: T, AP, RH, V and EP lie in the range 1.81–37.11 Celsius, 992.89–1033.30 milibar, 25.56%–100.16%, 25.36–81.56 cm-Hg, and 420.26–495.76 MW, respectively. The averages are taken from various locations around the plant. Also, ambient variables are recorded every second.

Scatterplot matrix and Correlogram of these variables are shown in Figures 1–2, respectively. According to Figure 1, V seems to have a curvilinear structure according to response variable EP. In this context, this variable breaches the linearity assumption of the classical regression model. Therefore, V will compose a nonparametric part of the semiparametric regression model. Other variables have considerable linear structure; consequently, T, AP, RH variables will be the parametric component of the semiparametric model.

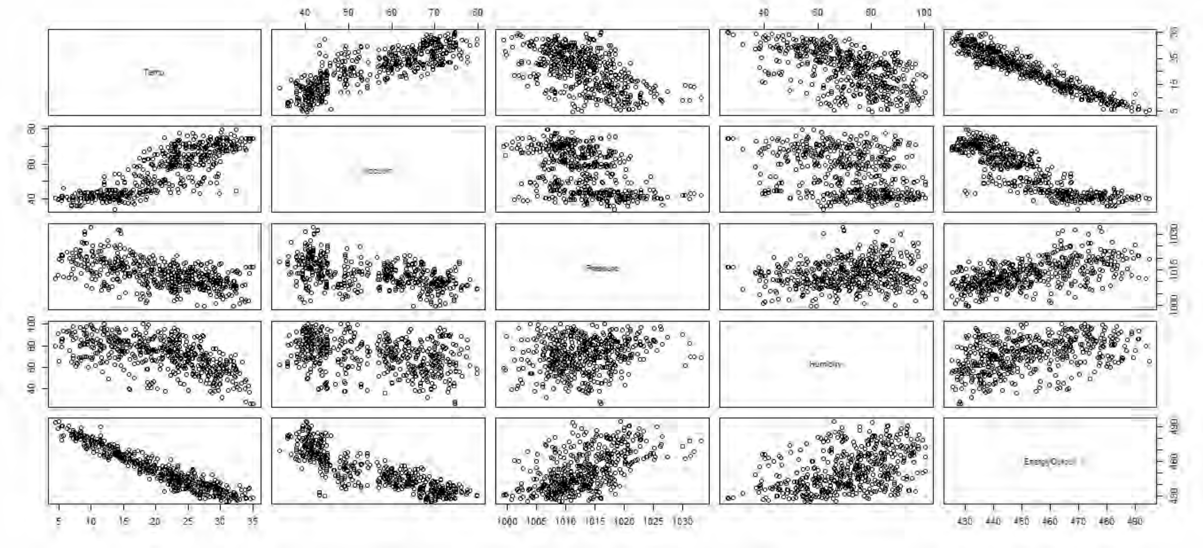


Figure 1: Scatterplot matrix of power plant data.

Thus, the semiparametric regression model in equation (1.1) can be defined the following way:

$$EP_i = \beta_1(T_i1) + \beta_2(AP_i2) + \beta_3(RH_i3) + f(V_i4) + \varepsilon_i, \quad i = 1, ..., 500. \tag{6.1}$$

Collinearity can be checked by simply calculating the correlations of the predictors in the model (6.1). Let $\mathbf{X}$ be a 500×4 matrix of the levels of the predictors in our real data example. A very simple measure of multicollinearity is inspection of the Correlogram given in Figure 2. It can be seen that several predictors have strong relationships with each other.

The eigenvalues of the $\mathbf{X}'\mathbf{X}$ for power plant data are $\lambda_1 = 0.01$, $\lambda_2 = 1613$, $\lambda_3 = 3481$, $\lambda_4 = 138950$, respectively. As is known, small eigenvalues indicate a bad condition in the data and maybe a collinearity problem. In order to determine the existence of multicollinearity, a condition index might be used. Condition Index (CI) is commonly used as an overall collinearity measure (Belsley *et al.*, [5]). If the value of CI exceeds 30, then we conclude that there is a strong multicollinearity in the data. This index is calculated as follows:

$$CI = \left[\lambda_{max}(\mathbf{X}'\mathbf{X})/\lambda_{min}(\mathbf{X}'\mathbf{X})\right]^{1/2} = 3723.10.$$

The value of $CI = 3723.10$ is an indication of potential multicollinearity problems. To combat with the collinearity, researchers use the ridge regression estimators given in (3.1). The illustration here will be based on kernel type ridge estimators given in (3.4) and (3.5). The parameter are chosen by minimizing the $\mathrm{AIC_c}, \mathrm{GCV}, \mathrm{BIC}, \mathrm{REML}, \mathrm{C}_p$ and RECP criteria, respectively. Also, the tuning parameter k is chosen with the generalized ridge regression estimator suggested by Hoerl and Kennard [13], [14]. The outcomes are given in Table 1.

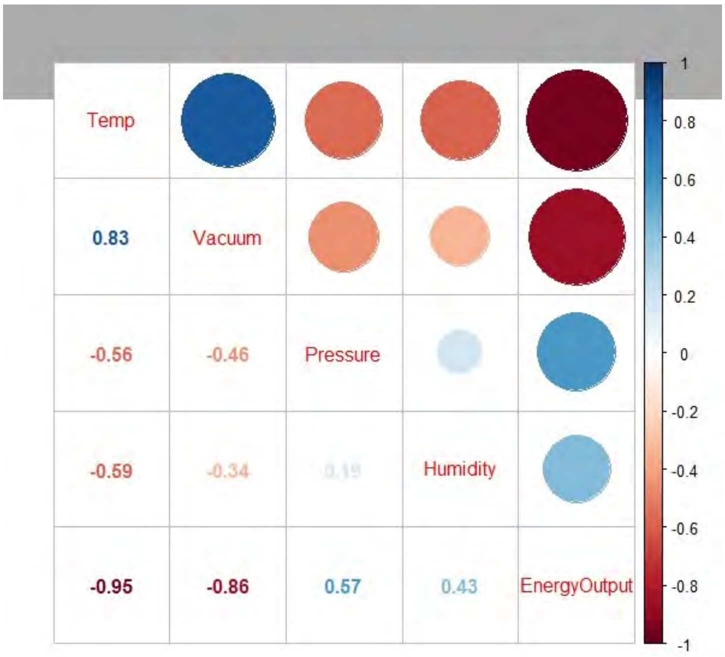


Figure 2: Correlogram for power plant data: Red colour indicates a negative correlation between the variables, while blue colour denotes a positive correlation. Size of the circle and intensity of the colour shows the strength of the relationships between variables.

As denoted in Table 1, slope parameters estimated with AIC$_c$ and RECP are very similar and likewise BIC and GCV. The SMDE values, variances and bias values of the semiparametric model have been obtained from six different selection methods. The SMDE, bias and variance values calculated by RECP criterion smaller than other methods. This is indicated in bold. In this situation, it is obvious that the RECP criterion has a more convincing performance for the selection of the parameter λ and that C$_p$ method does not perform well under study.

Table 1: Estimated coefficients of parametric component of the model.

	$\hat{\beta}_{\mathrm{AIC_c}}$	$\hat{\beta}_{\mathrm{BIC}}$	$\hat{\beta}_{\mathrm{GCV}}$	$\hat{\beta}_{\mathrm{REML}}$	$\hat{\beta}_{\mathrm{RECP}}$	$\hat{\beta}_{\mathrm{C}_p}$
T	-2.21931	-2.1932	-2.1924	-2.1928	-2.20437	-2.20721
AP	0.5024	0.4925	0.5003	0.5023	0.48276	0.47746
RH	-0.1660	-0.1678	-0.1657	-0.1659	-0.16813	-0.16862
SMDE	425.2455	571.9484	565.9700	672.4861	408.2248	696.5099
Bias	19.8500	23.2257	23.1069	25.1973	19.69050	25.67990
Variance	31.2230	32.7831	32.3670	37.9560	20.50790	37.0527

The smooth curves in Figure 3 are the graph of $\hat{y} = \hat{f}(V)$, different nonparametric estimates of the effect of V variable on EP. For smoothed curves the MISE values given in (7.2) are 24.8788, 26.0882, 26.0683, 26.3782, 24.1794 and 26.7481, respectively. Here, all of the selection methods have shown almost the same performance except the C$_p$ criterion. Thus, we can say that the C$_p$ does not provide a good empirical approximation.

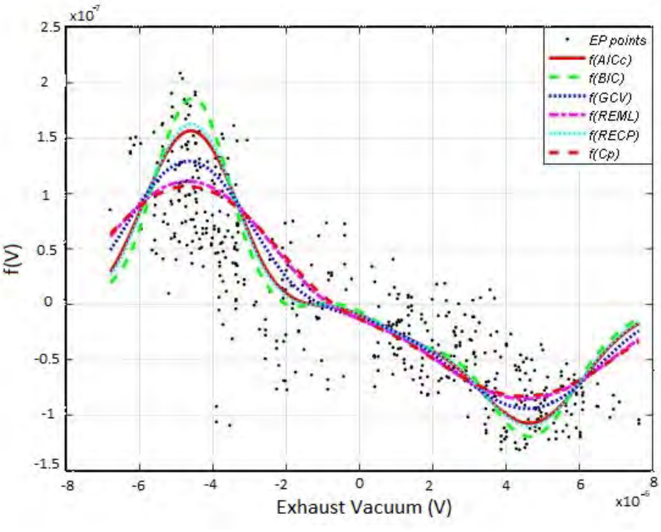


Figure 3: The smoothed curves for different kernel type ridge estimators based on AIC_c, BIC, GCV, REML, RECP and C_p methods, respectively.

7. MONTE CARLO SIMULATION STUDY

In this section, a Monte Carlo simulation study are carried out to compare the performance of the six selection methods expressed in Section 5. In the study, we simulate the response variable for samples of size $n = 50, 100$ and 200 with 103 iteration from the following model:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta x_{i4} + f(t_i) + \varepsilon_i, \quad i = 1, ..., n, \tag{7.1}$$

where $\varepsilon \sim N(0, \sigma^2 \mathbf{I}_n)$ which the values of $\sigma = 0.5$ and 1, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)' = (5, 4, 3, 2)'$, x_1, x_2, x_3 and x_4 are the correlated random variables, from the normal distribution. In here, three correlation (ρ) levels are considered as: 0.85, 0.95 and 0.99. Finally, the function f is represented by

$$f(t_i) = \sqrt{t_i(1-t_i)}\,\sin(2\pi/t_i) \qquad \text{with} \quad t_i = (1-0.5)/n.$$

It should be emphasized that we investigate three correlation levels, as stated above. If $\rho = 0.85$, for instance, this allows us to obtain about the same correlation levels between all pairs of variables. They are displayed in Table 2 for detecting correlations between the explanatory variables. Note that the outcomes from correlated data based on $\rho = 0.95$, and 0.99 are not reported here, because of space limitations.

Table 2: Correlation matrix for $\rho = 0.85$ level.

X	x_1	x_2	x_3	x_4
x_1	1.00	0.83	0.83	0.82
x_2	0.83	1.00	0.83	0.86
x_3	0.83	0.83	1.00	0.84
x_4	0.82	0.86	0.84	1.00

7.1. Evaluating the parametric part

The focus of the study is to estimate the parametric and nonparametric components of the semiparametric model. Additionally, the study is illustrating behaviors and performances of the selection methods with small, medium and large samples under multicollinear data sets. For each of the data sets, 1000 estimates of $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$ are obtained. These estimates are formed through a parametric component of the semiparametric regression model. The following tables and figures summarize the results of the simulation study.

There are four panels in Figure 4. In each panel, "AIC1, AIC2 and AIC3" denote the parametric biases of $\hat{\boldsymbol{\beta}}$ from semiparametric regression using ridge type kernel smoothing based on a smoothing parameter selected by improved AIC$_c$ method for $n = 50$, 100 and 200, respectively; similarly, "BIC1, BIC2 and BIC3" denote the case using BIC method for the sample sizes; "GCV1, GCV2 and GCV3" denote the case for GCV method; "R1, R2 and R3" denote REML method; "P1, P2 and P3" denote the RECP method; "Cp1, Cp2 and Cp3" illustrate Mallows' C$_p$ method . The ordinate indicates the scale of the biases of regression coefficients.

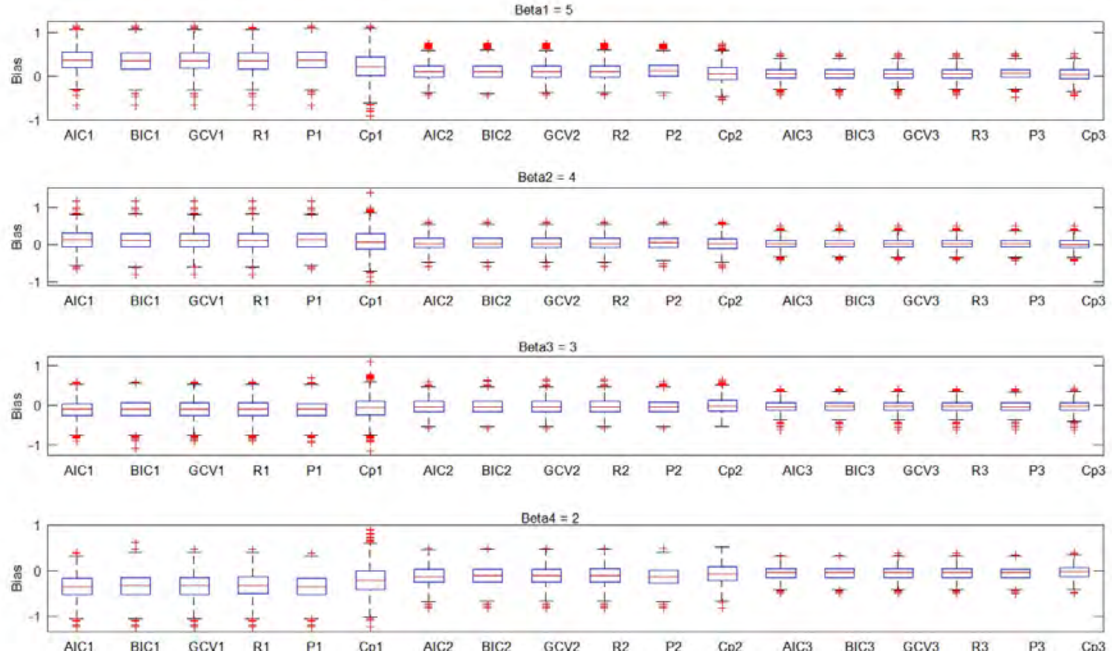


Figure 4: Boxplots of the estimates ($n = 50$, 100 and 200) obtained from semiparametric model for $\rho = 0.95$ and $\sigma = 1$. Panels indicate the boxplots of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ and $\hat{\beta}_4$.

In this study, there are 18 different configurations. Since it is hard to illustrate here all of these configurations, some of them are given in Figure 4 for correlation level $\rho = 0.95$ and $\sigma = 1$. As the sample size n gets larger, the range of estimates are getting narrower. That means that estimates from medium and large sized samples are more stable than those from small sized samples. If there is a correlation between the predictors, then the sample size has an effect on the quality of parametric estimates. We can say that kernel type ridge estimators work well for all samples. The key idea of the study is to compare the SMDEs for the estimators computed with each one of the criteria. The values of SMDE are illustrated in Table 3. The criterion that has the smallest SMDE is the best one.

Table 3: Average SMDEs of the parameters based on 1000 Monte Carlo runs.

n	ρ	*CI*	σ	AIC$_c$	BIC	GCV	REML	RECP	C$_p$
50	0.85	21.62	0.5	0.0043	0.0040	0.0040	0.0040	0.0040	0.0045
		34.03	1.0	0.0050	0.0048	0.0047	0.0047	0.0045	0.0055
	0.95	47.70	0.5	0.0157	0.0155	0.0159	0.0155	0.0132	0.0168
		53.21	1.0	0.0211	0.0209	0.0201	0.0193	0.0199	0.0246
	0.99	98.15	0.5	0.1348	0.1444	0.1267	0.1047	0.1366	0.1195
		100.56	1.0	0.2356	0.2244	0.2457	0.2032	**0.1995**	0.2010
100	0.85	14.78	0.5	0.0007	0.0007	0.0007	0.0007	0.0007	0.0008
		30.16	1.0	0.0010	0.0009	0.0009	0.0009	0.0009	0.0018
	0.95	45.84	0.5	0.0043	0.0043	0.0043	0.0042	0.0038	0.0044
		68.12	1.0	0.0052	0.0051	0.0052	0.0051	0.0050	0.0064
	0.99	75.24	0.5	0.0272	0.0275	0.0278	0.0268	0.0258	0.0278
		91.01	1.0	0.0385	0.0384	0.0398	0.0381	**0.0383**	0.0399
200	0.85	24.42	0.5	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002
		22.31	1.0	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
	0.95	57.71	0.5	0.0012	0.0012	0.0011	0.0011	0.0011	0.0011
		63.94	1.0	0.0014	0.0014	0.0015	0.0014	0.0014	0.0018
	0.99	99.41	0.5	0.0102	0.0110	0.0111	0.0113	0.0088	0.0136
		110.48	1.0	0.0152	0.0154	0.0156	0.0151	**0.0148**	0.0177

As discussed in real data, the values of *CI* presented in Table 3 are given to measure the extent of multicollinearity in the simulated data sets. It is readily seen that we have mostly multicollinear data sets. According to the same table, it is possible to see that the BIC, GCV, REML and RECP outperform AIC$_c$ and C$_p$ criteria for samples of size $n = 50$ and $\rho = 0.85$. Also, we see that the performances of five criteria, except C$_p$, behaviour quite similar in the medium and large sized samples generated by various scenarios. Notice, however, that RECP has a better performance under multi-collinear data sets especially for highly correlation levels. They are indicated in bold in Table 3. A very attractive component here is that as the sample size increase, the SMDE values decrease for all criteria based on correlation level of $\rho = 0.99$.

Table 4: Simulated bias of the slope parameters for $\rho = 0.99$ and $\sigma = 0.5$.

n	$\boldsymbol{\beta}$	AIC$_c$	BIC	GCV	REML	RECP	C$_p$
50	$\hat{\beta}_1$	0.0827	0.0811	0.0866	0.0852	0.0841	0.0933
	$\hat{\beta}_2$	0.0462	0.0458	0.0472	0.0492	0.0428	0.0444
	$\hat{\beta}_3$	0.0240	0.0234	0.0248	0.0242	0.0237	0.0320
	$\hat{\beta}_4$	0.0842	0.0980	0.0951	0.0985	0.0846	0.0874
100	$\hat{\beta}_1$	0.0547	0.0559	0.0558	0.0589	0.0563	0.0393
	$\hat{\beta}_2$	0.0186	0.0177	0.0176	0.0181	0.0127	0.0175
	$\hat{\beta}_3$	0,0138	0.0151	0.0151	0.0146	0.0107	0.0153
	$\hat{\beta}_4$	0.0311	0.0391	0.0390	0.0382	0.0283	0.0321
200	$\hat{\beta}_1$	0.0219	0.0226	0.0219	0.0222	0.0150	0.0164
	$\hat{\beta}_2$	0.0031	0.0037	0.0035	0.0033	0.0019	0.0036
	$\hat{\beta}_3$	0.0032	0.0032	0.0033	0.0032	0.0028	0.0034
	$\hat{\beta}_4$	0.0152	0.0169	0.0173	0.0168	0.0126	0.0177

Table 4 presents a checking of the bias of the slope parameters of the model (7.1). The number of parameters $p = 4$ and the parametric component of the model consists of real parameter vector $\boldsymbol{\beta} = (5, 4, 3, 2)^{\top}$. In general, sample sizes get larger, estimates obtained by six different kernel type estimators give small bias values, as expected. Among six kernel type ridge estimators, the one obtained by using RECP criterion provide the smallest bias of the estimation of real coefficients, especially for samples of size $n = 200$. Results that related to other correlation and sigma levels are similar. So, they are not reported here.

7.2. Measuring and comparing the efficiencies

In order to illustrate and compare the efficiency of the selection methods based on highly correlated data, a relative efficiency values are constructed from the SMDE ratios in (4.13). For each sample size the mentioned values are displayed in Figure 5. As can be seen from Figure 5, relative efficiency values of the RECP are better than others except for samples of size $n = 50$ and $\rho = 0.85$. This case shows that RECP is more efficient than the other selection methods, especially for all samples based on highly correlated data. Note also that outcomes from correlated data based on $\rho = 0.90$ are similar to the results displayed in Figure 5 under $\rho = 0.99$ and are not reported here.

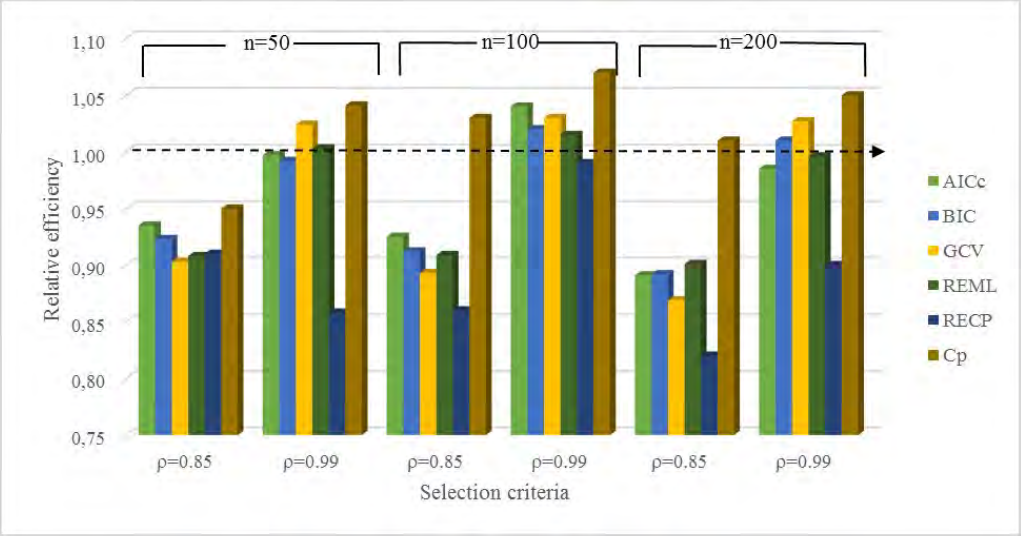


Figure 5: The column chart provides the averaged-relative efficiencies computed by the selection criteria.

Inspection of the relative efficiency values in Figure 5 also reveal that for $\rho = 0.85$, RECP criterion converges at 0.82, the highest rate when sample size is large. This indicates that under multicollinear data and noisy data, RECP criterion has the best performance among all other criteria, making it an ideal selection method for semiparametric regression based on ridge type kernel smoothing method. It can also be observed from Figure 5 in which four criteria, $\mathrm{AIC_c}$, BIC, GCV and REML, perform similarly, and better than the C_p criterion.

7.3. Evaluating the nonparametric part

In order to measure the nonparametric component of the semiparametric model, 1000 estimates of function f are obtained for each selection criterion. Smoothness and appropriateness of curve estimates have been measured by using the mean of the integrated squared error (MISE) value:

$$\text{MISE} = \frac{1}{1000}\sum_{j=1}^{1000} \text{ISE}_j, \tag{7.2}$$

where ISE_j denotes the integrated square error for the sample j, given by

$$\text{ISE}_j = \int \Big(f(t) - \hat{f}_j(t)\Big)^2 dt \approx \frac{1}{n}\sum_{i=1}^{n}\Big(f(t_i) - \hat{f}_j(t_i)\Big)^2 \qquad \text{where} \quad t_i = \frac{i - 0.5}{n},$$

where $f(t_i)$ value at t_i points to the appropriate function f. In our simulation study, because 18 different configurations are carried out, it is very hard to illustrate all of them. Therefore, only four different configurations will be presented in Figure 6. The left panels in the figure represent the smoothed curves together with a real function $f(t)$. In each graph, the smoothed curves, $f(\text{AICc})$, $f(\text{BIC})$, $f(\text{GCV})$, $f(\text{REML})$, $f(\text{Cp})$, respectively, are estimates of function $f(t)$ using ridge type kernel smoothing based on AIC_c, BIC, GCV, REML, RECP and C_p criteria. Also, the right panels of the Figure 6 denote the boxplots of the MISE values in (7.2) for each criterion.

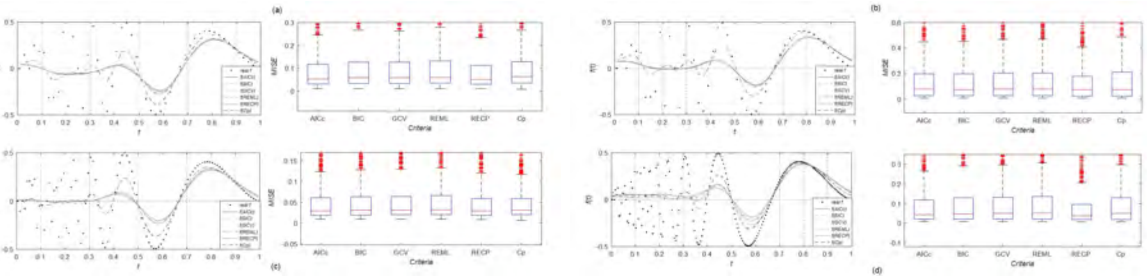


Figure 6: (a) $n = 50$, $\rho = 0.85$, $\sigma = 1$; (b) $n = 50$, $\rho = 0.99$, $\sigma = 0.5$; (c) $n = 100$, $\rho = 0.85$, $\sigma = 1$; (d) $n = 200$, $\rho = 0.99$, $\sigma = 1$.

In Figure 6 we see that the improvements in the MISE values mostly depend on the size of samples used in study. We also see that increasing the levels of correlation leads to poor performance in terms of MISE values, even if the sample sizes are the same. On the other hand, a visual inspection of the boxplots in all panels ((a) to (d)) denoted that RECP criteria maintain their dominance over the remaining selection methods, especially for large sized samples (say $n = 200$) based on data sets with $\rho = 0.99$ and $\sigma = 1$. On the contrary, the C_p criterion similar behaviors to others in terms of performance (see panels (a) and (c) of Figure 6). Notice, however, that the C_p yields poor estimates of the nonparametric component, compared to the estimates obtained by other methods, as in parametric cases.

8. CONCLUDING REMARKS

In this paper, for the parameters of the semiparametric model we proposed, kernel type ridge estimators minimize the penalized residual sum of squares method. Efficient computation of this method requires an optimum smoothing parameter λ. This optimum parameter is provided by means of AIC_{c}, BIC, GCV, REML, RECP and C_p criteria. Accordingly, we obtained six different estimators for the parametric and nonparametric components of the semiparametric model. We considered a real data example and simulated 1000 test observations to compare six different kernel type ridge estimators.

The empirical results confirmed that in the case of multicollinearity the kernel type ridge estimators based on AIC_{c}, BIC, GCV, REML and RECP, criteria have similar values of the SMDEs. The RECP, however, are superior to others in terms of SMDEs, especially when higher correlation levels are used. Throughout this discussion, the estimators based on C_p do not yield better performance in prediction of parametric and nonparametric components. On the other hand, although the REML criterion is more stable than AIC_{c}, GCV and RECP criteria, its performance is not good for all sample sizes and correlation levels. For the simulation studies, the findings of the numerical experiments are summarized in Tables 3–4 and Figures 4–6. We conclude the following statements from these tables and figures:

- For all the selection criteria, the SMDE, variance, and bias values of the slope parameters (or regression coefficients) start to decrease as the sample size n gets larger.
- For small sample sizes, as expected the bias values of slope parameter increase as the correlation and sigma levels increase.
- Also expected, when the lower correlation levels (i.e., $\rho = 0.85$) are used, the MISE values decreases for all selection criteria.
- Finally, when comparing the six selection methods, we see that the kernel type ridge estimators based on RECP method perform better than the others in terms of the SMDE, variance and bias values of the estimates for all sample sizes under collinear data.

A. APPENDIX: SUPPLEMENTAL TECHNICAL MATERIALS

A.1. Proof of Theorem 3.1

Consider data augmentation methods of penalized residual sum of squares fitting. Suppose that $\mathbf{W}_\lambda$ is symmetric smoother matrix. We wish to obtain the vector $\hat{\boldsymbol{\beta}}_R(k)$ that minimizes the penalized residual sum of squares criterion (3.3) by using augmented data sets of the form

$$\mathbf{X}_A = \begin{bmatrix} \tilde{\mathbf{X}}_{n\times p} \\ \left(\sqrt{k}\mathbf{I}\right)_p \end{bmatrix} = \begin{bmatrix} \tilde{x}_{11} & \tilde{x}_{12} & ... & \tilde{x}_{1p} \\ \tilde{x}_{21} & \tilde{x}_{22} & ... & \tilde{x}_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{x}_{n1} & \tilde{x}_{n2} & ... & \tilde{x}_{np} \\ \sqrt{k} & 0 & ... & 0 \\ 0 & \sqrt{k} & ... & \vdots \\ \vdots & ... & \ddots & 0 \\ 0 & ... & 0 & \sqrt{k} \end{bmatrix}_{((n+p)\times p)} \quad \text{and} \quad \mathbf{y}_A = \begin{bmatrix} \tilde{\mathbf{y}}_{n\times 1} \\ \mathbf{0}_p \end{bmatrix} = \begin{bmatrix} \tilde{y}_{11} \\ \tilde{y}_{21} \\ \vdots \\ \tilde{y}_{n1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{((n+p)\times 1)},$$

where $\sqrt{k}\mathbf{I}_p$ is a $p \times p$ new diagonal matrix with diagonal elements equal to the square root of the shrinkage parameter and $\mathbf{0}_p$ is $p \times 1$ new vector of zeros. Also, $\tilde{\mathbf{X}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{X}$ and $\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{W}_\lambda)\mathbf{y}$ as defined in equation (2.8), are partial residuals.

Similar to the ordinary least squares, the kernel ridge type estimators can be conveniently obtained using an augmented data set. A researcher could use this information to construct a penalized least-squares estimator $\hat{\boldsymbol{\beta}}_R(k)$ of $\boldsymbol{\beta}$. The estimator can be derived by

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'_A\mathbf{X}_A\right)^{-1}\mathbf{X}'_A\mathbf{y}_A \\ &= \left(\begin{bmatrix} \tilde{\mathbf{X}}' & (\sqrt{k}\mathbf{I}_p)' \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{X}}' \\ (\sqrt{k}\mathbf{I}_p)' \end{bmatrix}\right)^{-1} \begin{bmatrix} \tilde{\mathbf{X}}' & (\sqrt{k}\mathbf{I}_p)' \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{y}} \\ \mathbf{0}_p \end{bmatrix} \\ &= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + \left(\sqrt{k}\mathbf{I}_p\right)^2\right)^{-1} \left(\tilde{\mathbf{X}}'\tilde{\mathbf{y}} + \left(\sqrt{k}\mathbf{I}_p\right)\mathbf{0}_p\right) \\ &= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{y}}. \end{aligned}$$

Hence, as claimed, this confirms that the kernel type ridge type estimator of the unknown parameters in the models (1.1) or (1.2) is

$$\hat{\boldsymbol{\beta}}(k) = \left(\mathbf{X}'_A\mathbf{X}_A\right)^{-1}\mathbf{X}'_A\mathbf{y}_A = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}. \tag{A.1}$$

A.2. Derivation of the equations (4.4)–(4.6)

Using the definition of $\hat{\boldsymbol{\beta}}_R(k)$ ridge and our modeling assumption on the mean function $E(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}\boldsymbol{\beta}$, we obtain:

$$\begin{aligned}
\left(\hat{\boldsymbol{\beta}}_R(k)\right) &= E\left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}\right] = E\left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right)\mathbf{y}\right] \\
&= E\left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right)\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon}\right)\right] \\
&= E\left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right)\boldsymbol{\varepsilon}\right] \\
&= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} \\
&= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p - k\mathbf{I}_p\right)\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} \\
&= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)\boldsymbol{\beta} - k\mathbf{I}_p\boldsymbol{\beta}\right] + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} \\
&= \left[\mathbf{I}_p - k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\right]\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} \\
&= \boldsymbol{\beta} - k\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}}.
\end{aligned}$$

Equivalently, from (4.1), we obtain

$$\begin{aligned}
E\left(\hat{\boldsymbol{\beta}}_R(k)\right) = E\left(\left(\mathbf{I}_p + k(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\right)^{-1}\hat{\boldsymbol{\beta}}_p\right) &= E\left[\left(\mathbf{I}_p + k(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\right)^{-1}(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right)\mathbf{y}\right] \\
&= \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{X}}'\tilde{\mathbf{f}}\right).
\end{aligned}$$

Hence, using the abbreviation in equation (4.3), as claimed before, it is obtained $E\left(\hat{\boldsymbol{\beta}}_R(k)\right)$, and Bias $\left(\hat{\boldsymbol{\beta}}_R(k)\right)$ in equations (4.4), (4.5), and (4.6), respectively. Also, we denote the variance property of an estimator $\hat{\boldsymbol{\beta}}_R(k)$ by covariance matrix:

$$\begin{aligned}
\text{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right) &= E\left[\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)'\right] \\
&= E\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\left(\mathbf{I} - \mathbf{W}_\lambda\right)\boldsymbol{\varepsilon} \\
&\quad - \left[\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\boldsymbol{\beta} + \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{f}}\right] \\
&= E\left(\left(\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)\boldsymbol{\varepsilon}\right)\left(\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)\boldsymbol{\varepsilon}\right)'\right) \\
&= E\left(\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)^2\right)\left(\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\right)E(\boldsymbol{\varepsilon}^{\mathbf{2}}) \\
&= \sigma^2\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)^2\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}.
\end{aligned}$$

As a result, it can be expressed as the following result with abbreviation

$$\text{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right) = \sigma^2\mathbf{G}_k\tilde{\mathbf{X}}'(\mathbf{I} - \mathbf{W}_\lambda)^2\tilde{\mathbf{X}}\mathbf{G}_k,$$

as claimed.

A.3. The derivation of the smoother matrix and $E(\text{RSS})$

$$\begin{aligned}
\hat{\mathbf{y}} &= \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_R(k) + \hat{\mathbf{f}}_R(k) = \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} + \mathbf{W}_\lambda\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_R(k)\right)\\
&= \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} + \mathbf{W}_\lambda\left[\mathbf{y} - \mathbf{X}\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}} + k\mathbf{I}_p\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}\right]\\
&= \tilde{\mathbf{X}}\mathbf{G}_k\tilde{\mathbf{X}}'\tilde{\mathbf{y}} + \mathbf{W}_\lambda\left[\mathbf{y} - \mathbf{X}\mathbf{G}_k\tilde{\mathbf{X}}'\tilde{\mathbf{y}}\right]\\
&= \mathbf{X}(\mathbf{I}_p - \mathbf{W}_\lambda)\mathbf{G}_k\tilde{\mathbf{X}}'\mathbf{y} + \mathbf{W}_\lambda\left(\mathbf{y} - \mathbf{X}(\mathbf{I}_p - \mathbf{W}_\lambda)\mathbf{G}_k\tilde{\mathbf{X}}'\mathbf{y}\right)\\
&= \tilde{\mathbf{X}}\mathbf{G}_k\tilde{\mathbf{X}}'\mathbf{y} + \mathbf{W}_\lambda\mathbf{y} - \mathbf{W}_\lambda\mathbf{H}\mathbf{y} = \mathbf{W}_\lambda\mathbf{y} + (\mathbf{I}_p - \mathbf{W}_\lambda)\,\mathbf{H}\mathbf{y}\\
&= \left[\mathbf{W}_\lambda + (\mathbf{I}_p - \mathbf{W}_\lambda)\,\mathbf{H}\right]\mathbf{y} = \mathbf{H}_\lambda\mathbf{y},
\end{aligned}$$

where $\mathbf{H} = \tilde{\mathbf{X}}\mathbf{G}_k\tilde{\mathbf{X}}'$. Accordingly, the smoother matrix based on smoothing parameter λ is

$$\mathbf{H}_\lambda = \mathbf{W}_\lambda + (\mathbf{I}_p - \mathbf{W}_\lambda)\tilde{\mathbf{X}}\mathbf{G}_k\tilde{\mathbf{X}}',$$

as defined in the equation (4.8).

The expected value of the RSS in equation (4.9) can be given by

$$\begin{aligned}
E(\text{RSS}) &= E\left((\mathbf{y} - \mathbf{H}_\lambda)'(\mathbf{y} - \mathbf{H}_\lambda)\right)\\
&= E\left(\mathbf{y}'(\mathbf{I} - \mathbf{H}_\lambda)'(\mathbf{I} - \mathbf{H}_\lambda)\mathbf{y}\right) = E\left(\mathbf{y}'(\mathbf{I} - \mathbf{H}_\lambda)^2\mathbf{y}\right)\\
&= \operatorname{tr}\left((\mathbf{I} - \mathbf{H}_\lambda)^2\sigma^2\mathbf{I}\right) + E(\mathbf{y}')(\mathbf{I} - \mathbf{H}_\lambda)^2E(\mathbf{y})\\
&= n\sigma^2\operatorname{tr}(\mathbf{H}_\lambda^2) - 2\sigma^2\operatorname{tr}(\mathbf{H}_\lambda) + E(\mathbf{y}')(\mathbf{I} - \mathbf{H}_\lambda)^2E(\mathbf{y})\\
&= \sigma^2\left[n - \operatorname{tr}(2\mathbf{H}_\lambda - \mathbf{H}_\lambda^2)\right] + E(\mathbf{y}')(\mathbf{I} - \mathbf{H}_\lambda)'(\mathbf{I} - \mathbf{H}_\lambda)E(\mathbf{y}).
\end{aligned}$$

A.4. Proof of Lemma 4.1

Since the MDE equals $\sum_{j=1}^{k} E\left(\hat{\beta}_{jR}(k) - \beta_j\right)^2$ it is sufficient to prove for a scalar $\hat{\boldsymbol{\beta}}_R(k)$

$$\begin{aligned}
E\left[\left(\hat{\boldsymbol{\beta}}_R(k) - \boldsymbol{\beta}\right)^2\right] &= \operatorname{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right) + \operatorname{Bias}^2\left(\hat{\boldsymbol{\beta}}_R(k)\right)\\
&= E\left[\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right) + \left(E\left(\hat{\boldsymbol{\beta}}_R(k)\right) - \boldsymbol{\beta}\right)\right]^2\\
&= E\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)^2 + \left(E\left(\hat{\boldsymbol{\beta}}_R(k)\right) - \boldsymbol{\beta}\right)^2\\
&\quad + 2\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)'\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)\\
&= E\left(\hat{\boldsymbol{\beta}}_R(k) - E\left(\hat{\boldsymbol{\beta}}_R(k)\right)\right)^2 + E\left(E\left(\hat{\boldsymbol{\beta}}_R(k)\right) - \boldsymbol{\beta}\right)^2\\
&= \operatorname{Var}\left(\hat{\boldsymbol{\beta}}_R(k)\right) + \operatorname{Bias}^2\left(\hat{\boldsymbol{\beta}}_R(k)\right).
\end{aligned}$$

This completes the proof of the Lemma 4.1.

REFERENCES

[1] Ahmed, S.E. (2014). *Penalty, Shrinkage and Pretest Strategies: Variable Selection and Estimation*, Springer, New York.

[2] Aydin, D. (2014). Estimation of partially linear model with smoothing spline based on different selection methods: a comparative study, *Pakistan Journal of Statistics*, **30**(1), 35–56.

[3] Aydin, D.; Memeddeli, M. and Omay, R.E. (2013). Smoothing parameter selection for nonparametric regression using smoothing spline, *European Journal of Pure and Applied Mathematics*, **6**(2), 222–238.

[4] Aydin, D. and Yilmaz, E. (2018). Modified estimators in semiparametric regression models with right-censored data, *Journal of Statistical Computation and Simulation*, **88**(8), 1470–1498.

[5] Belsley, D.A.; Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*, John Wiley and Sons, New York.

[6] Chen, H. (1988). Convergence rates for parametric components in a partially linear models, *The Annals of Statistics*, **16**(1), 136–146.

[7] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Num. Math.*, **31**(4), 377–403.

[8] Engle, R.; Granger, C.; Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales, *Journal of American Statistical Association*, **81**(394), 310–320.

[9] Eubank, R.L. (1999). *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

[10] Foucart, T. (1999). Stability of the inverse correlation matrix, partial ridge regression, *Journal of Statistical Planning And Inferences*, **77**(1), 141–154.

[11] Gibbons, D.G. (1981). A simulation study of some ridge estimators, *Journal of the American Statistical Association*, **76**(373), 131–139.

[12] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Model*, Chapman and Hall, New York.

[13] Hoerl, A.E. and Kennard, R.W. (1970a). Ridge regression: biased estimation for non orthogonal problems, *Technometrics*, **12**(1), 55–67.

[14] Hoerl, A.E. and Kennard, R.W. (1970b). Ridge regression: applications to non orthogonal problems, *Technometrics*, **12**(1), 69–82.

[15] Hu, H. (2005). Ridge estimation of a semiparametric regression model, *Journal of Computational and Applied Mathematics*, **176**(1), 215–222.

[16] Hurvich, C.M.; Simonoff, J.S. and Tasi, C.L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *J. R. Statist. Soc. B.*, **60**(2), 271–293.

[17] Kaya, H.; Tufekci, P. and Gurgen, S.F. (2012). Local and global learning methods for predicting power of a combined gas steam turbine, *Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering ICETCEE*, 13–18.

[18] Kibria, B.M. (2003). Performance of some new ridge regression estimators, *Communications in Statistics – Simulation and Computation*, **32**(2), 419–435.

[19] Lee, Thomas C.M. (2003). Smoothing parameter selection for smoothing splines: a simulation study, *Computational Statistics and Data Analysis*, **42**(1–2), 139–148.

[20] Lee, Thomas C.M. (2004). Improved smoothing spline regression by combining estimates of different smoothness, *Statistics and Probability Letters*, **67**(2), 133–140.

[21] Liang, H. (2006). Estimation partially linear models and numerical comparison, *Computational Statistics and Data Analysis*, **50**(3), 675–687.

[22] Mallows, C. (1973). Some comments on C_p, *Technometrics*, **15**(4), 661–675.

[23] Muniz, G. and Kibria, B.M.G. (2009). On some ridge regression estimators: an empirical comparisons, *Communications in Statistics – Simulation and Computation*, **38**(3), 621–630.

[24] Nadaraya, E.A. (1964). On estimating regression, *Theory of Probability and Its Applications*, **9**(1), 141–142.

[25] Reiss, P.T. and Ogden, R.T. (2004). Smoothing parameter selection for a class of semiparametric linear models, *J. R. Statist. Soc. B*, **71**(2), 505–523.

[26] Robinson, M.P. (1988). Root-n-consistent semi-parametric regression, *Econometrica*, **56**(4), 931–954.

[27] Roozbeh, M.; Arashi, M. and Niroumanda, H.A. (2010). Semiparametric ridge regression approach in partially linear models, *Communications in Statistics – Simulation and Computation*, **39**(3), 449–460.

[28] Schimek, G. Michael (2000). *Smoothing and Regression: Approaches, Computation, and Application*, John Willey and Sons, USA.

[29] Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**(2), 461–464.

[30] Speckman, P. (1988). Kernel smoothing in partially linear model, *J. Royal Statist., Soc. B.*, **50**(3), 413–436.

[31] Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models, *Journal of the American Statistical Association*, **84**(405), 276–283.

[32] Tufekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power using machine learning methods, *International Journal of Electrical Power and Energy Systems*, **60**, 505–523.

[33] Wahba, G. (1990). *Spline Model for Observational Data*, SIAM, Philadelphia.

[34] Watson, G.S. (1964). Smooth regression analysis, *Sankhya, Series A*, **26**(4), 359–372.

[35] Yuzbasi, B. (2014). *Penalty and Non-Penalty Estimations Strategies for Linear and Partially Linear Models*, PhD Thesis, Inonu University, Malatya.

[36] Yuzbasi, B. and Ahmed, S.E. (2016). Shrinkage and penalized estimation in semi-parametric models with multicollinear data, *Journal of Statistical Computation and Simulation*, **86**(17), 3543–3561.

[37] Yuzbasi, B.; Ahmed, S.E. and Aydin, D. (2017). Ridge-type pretest and shrinkage estimations in partially linear models, *Statistical Papers*, **61**(2), 869–898. Doi: 10.1007/s00362-017-0967-8.

REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 71–85

A STOCHASTIC STUDY FOR A GENERALIZED LOGISTIC MODEL

Authors: Rafael Luís
– Department of Mathematics, University of Madeira, and
Center for Mathematical Analysis, Geometry, and Dynamical Systems,
University of Lisbon,
Portugal
rafael.luis.madeira@gmail.com

Sandra Mendonça
– Department of Mathematics, University of Madeira, and
Center of Statistics and Applications, University of Lisbon,
Portugal
smfm@uma.pt

Received: December 2017 Revised: July 2018 Accepted: October 2018

Abstract:

- In this paper some properties of a generalized logistic discrete model are studied. Both autonomous and non-autonomous models are addressed, as well as the stochastic model, by varying the sequence of parameters that determine the sequence of mappings of the process. Some results on stability are established and the long-term behaviour of the orbits is studied.

Keywords:

- *periodic difference equations; stochastic difference equations; asymptotic stability; logistic equation; statistical dynamics.*

AMS Subject Classification:

- 37B55, 37C25, 37C75, 37H10, 37H20, 39A23; 39A30, 39A50.

1. INTRODUCTION

Dynamical systems occur in all branches of science. According to Martin Rasmussen [29], “the main goal of the study of a dynamical system is to understand the long behaviour of states in a system for which there is a deterministic rule for how a state evolves”. On the other hand, Christian Pötzsche [28] claims that “an understanding of the asymptotic behaviour of a dynamical system is probably one of the most relevant problems in sciences based on mathematical modeling”.

There are two approaches in the study of such mathematical models. The autonomous model where the system is governed by a single mapping and the non-autonomous model where the evolution in time is, in general, governed by a family of different mappings.

The non-autonomous systems arise naturally in the study of phenomena that evolve in time and cannot be ruled by the a single mapping by the simple fact that such phenomena do not repeat. For a general theory of non-autonomous (periodic) difference equations we refer a recent book by Luís [22] where the author presents the main concepts and results concerning periodic difference equations.

A generalization of discrete non-autonomous systems can be given by stochastic difference equations or random dynamical systems. The study of these systems are appropriate in the situation where the rules that govern the evolution of the system have a random nature.

Some works and authors in the field of random dynamical systems are worth-mention. The book of Arnold [5], where the author explores, separately, both random differential equations and random difference equations. The work of Kifer, [17] where the author studies basic connections between compositions of independent random transformations and corresponding Markov chains together with some applications. Liu in [21] reviews a selection of basic results in smooth ergodic theory and in the thermodynamic formalism of dynamical systems generated by compositions of random maps. An excellent tutorial on the asymptotic behaviours of random orbits of dynamical systems with random parameters may be found in the work of Ohno [27]. In 2009, Marie and Rousseau [25] presented a study of the recurrence behaviour in certain random dynamical systems and randomly perturbed dynamical systems. Baladi [6] uses transfer operators to construct invariant measures of chaotic dynamical systems. And to end this short list of references on random dynamical systems, we refer the excellent survey of Diaconis and Freedman [10] on iterated random functions, where the authors provide several examples under the unifying idea that the iterates of random Lipschitz functions converge if the functions are contracting on the average.

One of the well known models that have a discrete evolution is the quadratic model given by

$$x_{n+1} = \mu_n x_n(1-x_n), \qquad x \in [0,1], \quad \mu_n \in (0,4), \ n = 0,1,2,\dots. \tag{1.1}$$

When the sequence of parameters μ_n is constant, the model given by (1.1) is the well known logistic equation. The modern theory of discrete dynamical systems owns a great part of its development to the understanding of the dynamics of this equation, and may be found in many books on discrete dynamical systems, as the ones by Alligood, Sauer and Yorke [1, Chapter 1], by Devaney [9, Chapter 1], by Elaydi [11, Chapter 1] and by Zhang [30, Chapter 2], among others.

When the sequence of parameters is not constant, the dynamics of equation (1.1) is naturally more complex. Both, non-stochastic model, where the elements of the sequence of parameters are taken with a deterministic rule from the interval $(0, 4)$, and stochastic model, where the referred elements are taken randomly from the same interval, are far from being exhaustively studied. Some partial studies may be found in the literature. Grinfeld *et al.* [13] studied the bifurcation in 2-periodic logistic equations. AlSharawi and Angelos [2] showed that when $\mu_{n+p} = \mu_n$, for all n, the p-periodic logistic equation (1.1) has cycles (periodic solutions) of minimal periods 1, p, $2p$, $3p$, ... The same authors have also extended Singer's theorem to periodic difference equations, and used it to show that the p-periodic logistic equation has at most p stable cycles. Particular attention was given to the cases $p = 2$ and $p = 3$. AlSharawi *et al.* [3] and Alves [4] have, independently, presented an extension of Sharkovsky's theorem to periodic difference equations, where the main example is the periodic logistic equation.

In this paper some properties of a generalized logistic model given by

$$x_{n+1} = \mu_n x_n^k (1 - x_n), \tag{1.2}$$

where $x_n \in [0, 1]$, $k > 1$ and $\mu_n > 0$ for all $n = 0, 1, 2...$, are studied. Some particular studies on the stability in both, non-autonomous (periodic) model (Section 2) and stochastic model (Section 3) are presented. In particular, the dynamical system defined by equation (1.2) when $k = 2$ and $\mu_n \in (0, 27/4]$ is deeply studied. The main focus of this study is the comprehension of the model's dynamics in the parameter space.

Finally, it should be mentioned that Marotto [26] studied the autonomous equation (1.2) when $k = 2$ and $\mu_n = \mu$, for all natural n. When $\mu_n = \mu$, for all n, the dynamical properties of the autonomous equation (1.2) have been addressed by several authors, like Levin and May [20], Hernández-Bermejo and Brenig [14], Briden and Zhang [7], among others.

2. NON-STOCHASTIC MODEL

Let us consider the difference equation given by

$$x_{n+1} = \mu_n x_n^{k_n} (1 - x_n)\,, \tag{2.1}$$

where $x_n \in [0, 1]$, $\mu_n > 0$ and $k_n = 2, 3, 4, ...$ for all non negative integer n.

Equation (2.1) may be represented by the map

$$f_n(x) = \mu_n x^{k_n} (1 - x)\,.$$

In order to insure that $x_n \in I = [0, 1]$ for all n, we make the following assumption concerning the parameters

$$\mathbf{H}\!: \;\; \mu_n \le \left(\frac{k_n + 1}{k_n}\right)^{k_n} (k_n + 1), \qquad n = 0, 1, 2....$$

Assumption **H** guarantees that all the orbits in (2.1) are bounded. Furthermore, it guarantees that f_n maps the interval I into the interval I for all $n = 0, 1, 2....$

2.1. Autonomous equation

Let us first study the dynamics of the particular map $f(x) = \mu x^k (1-x)$, with $x \in I$, $\mu > 0$ and $k = 2, 3, \ldots$. To find the fixed points of f we determine the solutions of the equation $\mu x^k(1-x) = x$. After eliminating the trivial solution, $x = 0$, the positive fixed points are the solutions of

$$\mu x^{k-1}(1-x) = 1, \tag{2.2}$$

or equivalently

$$\ln(\mu) = -(k-1)\ln x - \ln(1-x). \tag{2.3}$$

Letting $g(x) = -(k-1)\ln x - \ln(1-x)$, we see that $g(x) > 0$ for all $x \in (0,1)$. Moreover, g is convex in the unit interval since $g'(x) > 0$, for all $x \in I$, and attains its minimum at $g(c_g)$ where $c_g = \frac{k-1}{k}$ is the unique critical point of g in the unit interval. Let $\mathbf{O}_\mu$ be the immediate basin of attraction of the origin.

1. If $g(c_g) > \ln(\mu)$, then Eq. (2.3) has no solution. Hence, $x^* = 0$ is the unique fixed point of the map f whenever $\mu < k\left(\frac{k}{k-1}\right)^{k-1}$. Under this scenario $x^* = 0$ is globally asymptotically stable, given that it is the unique fixed point in I. Notice that at the origin we have $f'(0) = 0$ and that $\mathbf{O}_\mu = [0,1]$.
2. If $g(c_g) = \ln(\mu)$, then Eq. (2.3) has a unique solution, $x^* = \frac{k-1}{k} = c_g$. Hence, the map f has a unique positive fixed point when $\mu = k\left(\frac{k}{k-1}\right)^{k-1}$. In this case and using (2.2), we obtain $|f'(x^*)| = 1$ and $|f''(x^*)| = -k^2 < 0$, that allows us to conclude that x^* is an unstable fixed point, but semi-stable from the right. Moreover, its immediate basin of attraction is the set $\left[x^*, \max f^{-1}(\{x^*\})\right]$ where $f^{-1}(\{x^*\})$ is the pre-image of $\{x^*\}$. Notice that $\mathbf{O}_\mu = I \setminus \left[x^*, \max f^{-1}(\{x^*\})\right]$.
3. If $g(c_g) < \ln(\mu)$, then Eq. (2.3) has two positive solutions. Hence, the map f possesses two positive fixed points whenever $\mu > k\left(\frac{k}{k-1}\right)^{k-1}$. The smaller, denoted as $\mathbf{A}_\mu$, is known as a threshold point and the greater, denoted by $\mathbf{K}_\mu$, is known as a carrying capacity. Under this scenario, the fixed point $\mathbf{A}_\mu$ is always unstable and the fixed point $\mathbf{K}_\mu$ is locally asymptotically stable in the interval $\left(\mathbf{A}_\mu, \max f^{-1}(\{\mathbf{A}_\mu\}\right)$ if $\left|k - \mu\mathbf{K}_\mu^k\right| < 1$. Moreover, $\mathbf{O}_\mu = [0, \mathbf{A}_\mu) \cup \left(\max f^{-1}(\{\mathbf{A}_\mu\}), 1\right]$.

Notice that the sequence $a_k = \left(\frac{k+1}{k}\right)^k (k+1)$ that is used to define Assumption $\mathbf{H}$ is increasing for $k = 2, 3, \ldots$. We now resume the precedent ideas in the following result, for a general integer $k = 2, 3, \ldots$:

Theorem 2.1. *Let $f(x) = \mu x^k(1-x)$, $k = 2, 3, \ldots$. Then the following yields:*

1. *If $\mu < k\left(\frac{k}{k-1}\right)^{k-1}$, then $x^* = 0$ is a globally asymptotically stable fixed point of f and its basin of attraction is the unit interval.*
2. *If $\mu = k\left(\frac{k}{k-1}\right)^{k-1}$, then the map has two fixed points, the origin and a positive fixed point $x^* = \frac{k-1}{k}$. This last one is locally asymptotically stable from the right and its immediate basin of attraction is the set $\left[x^*, \max f^{-1}(\{x^*\})\right]$. Moreover, $\mathbf{O}_\mu = I \setminus \left[x^*, \max f^{-1}(\{x^*\})\right]$.*

3. *If $\mu > k\left(\frac{k}{k-1}\right)^{k-1}$, then the map has three fixed points, the origin, a threshold fixed point $\mathbf{A}_\mu$ and a carrying capacity $\mathbf{K}_\mu$ such that $\mathbf{A}_\mu < \mathbf{K}_\mu$. The threshold fixed point is always unstable and if $|k - \mu\mathbf{K}_\mu^k| < 1$ the carrying capacity is locally asymptotically stable with a basin of attraction given by the set $\left(\mathbf{A}_\mu, \max f^{-1}(\{\mathbf{A}_\mu\})\right)$. Moreover, $\mathbf{O}_\mu = I \setminus \left[\mathbf{A}_\mu, \max f^{-1}(\{\mathbf{A}_\mu\})\right]$.*

Remark 2.1. Before ending this subsection and having in mind the next section, let us have a particular look in the dynamics of the autonomous equation when $k = 2$, i.e., the dynamics of the equation when the map is given by $f(x) = \mu x^2(1 - x)$. We will be needing these results when studying the corresponding stochastic equation.

1. If $\mu < 4$, then the origin is a globally asymptotically stable fixed point provided that it is the unique fixed point in the unit interval.

2. If $\mu = 4$, then the map possesses two fixed points, the origin and $x^* = \frac{1}{2}$. The basin of attraction of the origin is

$$\mathbf{O}_4 = \left[0, \frac{1}{2}\right) \cup \left(\frac{1+\sqrt{5}}{4}, 1\right], \tag{2.4}$$

while the basin of attraction of the positive fixed point is $\left[\frac{1}{2}, \frac{1+\sqrt{5}}{4}\right]$. Notice that $x^* = \frac{1}{2}$ is a fixed point semi-stable from the right.

3. If $4 < \mu$, then the map has three fixed points, the origin, the threshold point $\mathbf{A}_\mu = \frac{1}{2}\left(1 - \sqrt{\frac{\mu-4}{\mu}}\right)$ and the carrying capacity $\mathbf{K}_\mu = \frac{1}{2}\left(1 + \sqrt{\frac{\mu-4}{\mu}}\right)$.
It is a straightforward computation to see that, when $\mu > 4$,

$$|f'(\mathbf{A}_\mu)| = 3 + \frac{\mu}{2}\left(-1 + \sqrt{\frac{\mu-4}{\mu}}\right) > 1.$$

Hence, the fixed point $\mathbf{A}_\mu$ is unstable.
Similarly, we see that

$$|f'(\mathbf{K}_\mu)| = \left|3 - \frac{\mu}{2}\left(1 + \sqrt{\frac{\mu-4}{\mu}}\right)\right| < 1 \text{ iff } 4 < \mu < \frac{16}{3}.$$

When $\mu = \frac{16}{3}$ we have $f'(\mathbf{K}_\mu) = -1$. Forward computations show that the Schwarzian derivative evaluated at the fixed point is negative, i.e., $Sf(\mathbf{K}_\mu) < 0$. Consequently, from Theorem 2 in [24] it follows that the fixed point $\mathbf{K}_\mu$ is asymptotically stable. Thus, the fixed point $x^* = \mathbf{K}_\mu$ is locally asymptotically stable whenever $4 < \mu \le \frac{16}{3}$ and its basin of attraction is the set $\left(\mathbf{A}_\mu,\ \max f^{-1}(\{\mathbf{A}_\mu\})\right)$. Moreover,

$$\mathbf{O}_\mu = [0, \mathbf{A}_\mu) \cup \left(\max f^{-1}(\{\mathbf{A}_\mu\}), 1\right]. \tag{2.5}$$

2.2. Non-autonomous equation

We start this subsection presenting a result related to the non-autonomous equation (2.1) when $k = 2$ (although it may be extended for other values of the parameter k as well). It is not hard to prove the following:

Lemma 2.1. *Consider the non-autonomous difference equation given by*

$$x_{n+1} = \mu_n x_n^2 (1 - x_n), \tag{2.6}$$

where $x_n \in [0,1]$, $\mu_n \in \left(0, \frac{27}{4}\right]$, *for* $n = 0, 1, 2...$, *and* $\mathbf{O}_\mu$ *the immediate basin of attraction of the origin. Then*

$$4 \leq \mu_1 \leq \mu_2 \leq \frac{27}{4} \Rightarrow \mathbf{O}_4 \supseteq \mathbf{O}_{\mu_1} \supseteq \mathbf{O}_{\mu_2} \supseteq \mathbf{O}_{\frac{27}{4}}, \tag{2.7}$$

where $\mathbf{O}_4$ *is given by (2.4) and*

$$\mathbf{O}_{\frac{27}{4}} = \left[0, \frac{9-\sqrt{33}}{18}\right) \cup \left(\max f^{-1}\left(\left\{\mathbf{A}_{\frac{27}{4}}\right\}\right), 1\right], \tag{2.8}$$

where $\max f^{-1}\left(\left\{\mathbf{A}_{\frac{27}{4}}\right\}\right) \approx 0.971\,62$.

Let us now turn our attention to the non-autonomous periodic equation (2.1). We will study the case where the sequence of maps is p-periodic, i.e., when $f_{n+p} = f_n$, for all $n = 0, 1, 2, \ldots$. Under this scenario, equation (2.1) is p-periodic.

The dynamics of the non-autonomous p-periodic equation (2.1) is completely determined by the following composition operator

$$\Phi_p = f_{p-1} \circ ... \circ f_1 \circ f_0.$$

From assumption $\mathbf{H}$ it follows that $\Phi_p(I) \subseteq I$ with $\Phi_p(0) = 0$ and $\Phi_p(1) = 0$. Hence, by the Brouwer's fixed point theorem [16], the composition operator Φ_p has a fixed point in the unit interval.

It is clear that $x^* = 0$ is a locally asymptotically stable fixed point of Φ_p provided that $|\Phi_p'(0)| = 0$. Now, if $\Phi_p(x) < x$, for all $x \in (0,1)$, then $x^* = 0$ is the unique fixed point of the composition operator Φ_p in the unit interval. In this case, $x^* = 0$ is a globally asymptotically stable fixed point and its basin of attraction is the entire unit interval. This is the case where local stability implies global stability in the sense that every orbit of $x_0 \in I$ converge to the origin.

Notice that, if C_{Φ_p} is the set of critical points of Φ_p, i.e., if C_{Φ_p} contains all the solutions in the unit interval of the p equations $\Phi_i(x) = c_i$, $i = 0, 1, ..., p-1$, where c_i is the critical point of the map f_i, then $\Phi_p(x) < x$, for all $x \in (0,1)$ if $\Phi_p(c_{\Phi_p}) < c_{\Phi_p}$, where $c_{\Phi_p} \in C_{\Phi_p}$.

Now, if $|\Phi_p(x)| > x$ for some $x \in (0,1)$,the composition operator Φ_p has more than one fixed point. We know from Coppel's Theorem [8] that every orbit converges to a fixed point if and only if the equation $\Phi_p \circ \Phi_p(x) = x$ has no solutions with the exception of the fixed points of Φ_p. It is not possible, in general, to say much concerning the number of fixed points of Φ_p since we have many scenarios. However, if all maps f_i have a threshold fixed point $\mathbf{A}_i$ and we let $\mathbf{A}_m = \min\{\mathbf{A}_0, \mathbf{A}_1, ..., \mathbf{A}_{p-1}\}$ and $\mathbf{A}_M = \max\{\mathbf{A}_0, \mathbf{A}_1, ..., \mathbf{A}_{p-1}\}$, then one can show that the minimal positive fixed point of Φ_p, $\mathbf{A}_{\Phi_p}$, lies between $\mathbf{A}_m$ and $\mathbf{A}_M$ and is, in fact, an unstable fixed point. Under this scenario, the immediate basin of attraction of the origin is $\cup_{i\geq 1} J_i$ where $J_i \subset I$ and

$$\Phi_p(J_i) \subset [0, \mathbf{A}_{\Phi_p}).$$

See Figure 1 for an example of this scenario.

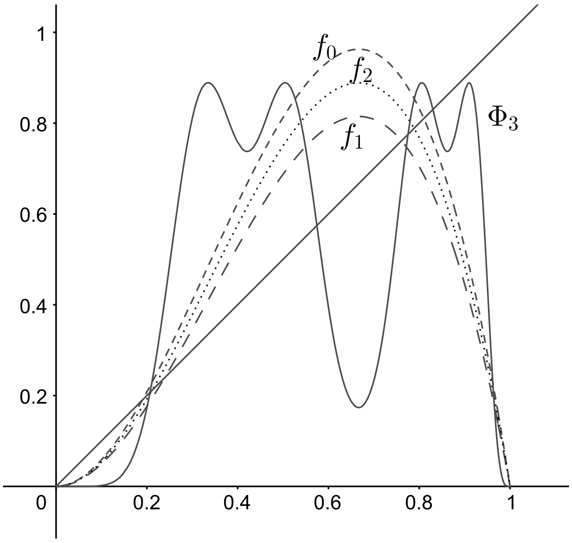


Figure 1: Composition of three generalized logistic maps. The composition map Φ_3 is represented by the solid curve and the individual maps are represented by the dashed curves. The values of parameters are $k = 2$, $\mu_0 = 6.5$ (f_0), $\mu_1 = 5.5$ (f_1) and $\mu_2 = 6$ (f_2).

We remark that each fixed point of the composition map Φ_p, with the exception of $x^* = 0$, generates a periodic orbit in equation (2.1). More precisely, if x^* is a non-trivial fixed point of Φ_p, then

$$\overline{C} = \{\overline{x}_0 = x^*, \overline{x}_1 = f_0(\overline{x}_0), \overline{x}_2 = f_1(\overline{x}_1), ..., \overline{x}_{p-1} = f_{p-2}(\overline{x}_{p-2})\}$$

is a periodic cycle of equation (2.1), which is locally asymptotically stable if

$$|\Phi_p'(x^*)| = \left|\prod_{i=0}^{p-1} f_i'(\overline{x}_i)\right| < 1.$$

Notice that, due the periodicity of the maps f_i, we have $\overline{x}_p = f_{p-1}(\overline{x}_{p-1}) = \overline{x}_0$, $\overline{x}_{p+1} = \overline{x}_1$, and so on.

From the dynamical point of view, it is interesting to know the region where the stability of the fixed points occurs. Since we are not able to find explicitly the fixed points of the composition map Φ_p for general values of the parameters k_i and μ_i, $i = 0, 1, ..., p-1$, we will particularize and study the cases where this is possible as are the cases when $p = 2, 3, 4$ and $k = 2$, i.e., we will study the dynamics of the system when the sequence of maps is 2-periodic and given by

$$f_{n \bmod (2)}(x) = \mu_{n \bmod (2)}\, x^k(1-x), \quad k = 2, 3, 4.$$

Let us start with the case $k = 2$. Following the techniques employed in [23], one can find the region of local stability of the fixed points of the composition map $\Phi_2 = f_1 \circ f_0$ by calculating the boundary where the absolute value of $\Phi_2'(x^*)$ is equal to one. Since the computations are long we will omit them here. The stability regions are depicted in Figure 2, in the parameter space $\mu_0 O \mu_1$.

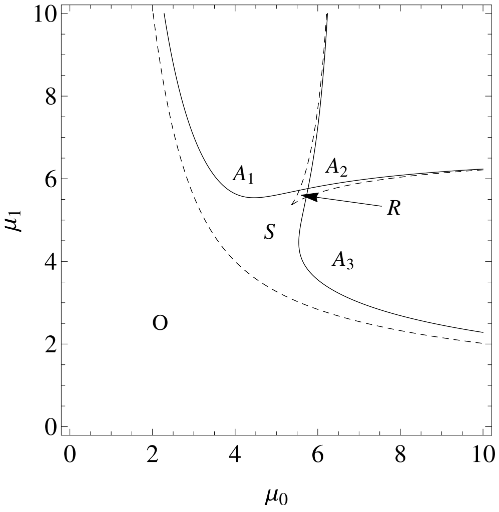


Figure 2: Region of local stability, in the parameter space $\mu_0 O \mu_1$ where the fixed points of $f_1 \circ f_0$ are locally asymptotically stable and the maps are given by $f_i(x) = \mu_i x^2(1-x)$, $i = 0, 2$.

If the parameters μ_0 and μ_1 belong to the region **O**, then the origin is a fixed point globally asymptotically stable. Once the parameters cross the dashed curve, from Region **O** to Region **S**, a bifurcation occurs, known as saddle-node bifurcation. The fixed point $x^* = 0$ becomes unstable and a new locally stable fixed point of Φ_2 is born. This fixed point is, in fact, a 2-periodic cycle of the 2-periodic equation (2.1). Now if the parameters μ_0 and μ_1 cross the dashed curve from Region **S** to Region **R,** a saddle-node bifurcation occurs. The 2-periodic cycle becomes unstable and a new locally asymptotically stable 2-periodic cycle is born.

At the solid curve a new type of bifurcation occurs known as a period-doubling bifurcation. Hence, when the parameters cross the solid curve from Region S to Region A_i, $i = 1,\ 2,\ 3$, the 2-periodic cycle of equation (2.1) becomes unstable and a new locally asymptotically stable 4-periodic cycle is born.

Following a similar idea as before, we are able to find (numerically) the regions of local stability of the 4-periodic cycle identified before. We notice that this scenario of period-doubling bifurcation continues route to chaos.

For a general framework of bifurcation in one-dimensional periodic difference equations, we refer the work of Elaydi, Luís, and Oliveira in [12].

Now, following the same techniques as before, we are able to find the regions of local stability of fixed points when $k = 3$ and $k = 4$. These regions are represented in Figure 3. As we can observe, they are similar to the case $k = 2$ and the conclusions follow in the same fashion.

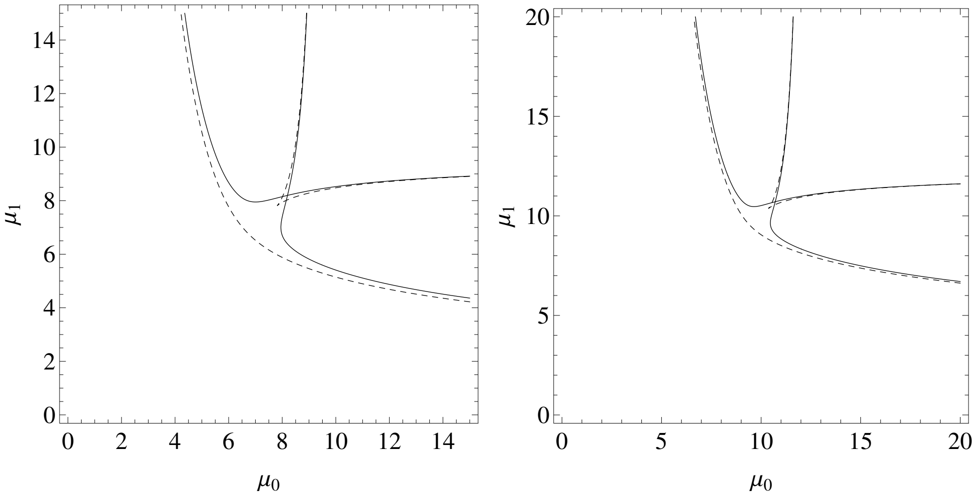


Figure 3: Regions of local stability, in the parameter space, of the 2-periodic equation when $k = 3$ (left) and $k = 4$ (right).

3. STOCHASTIC MODEL

In this section, we will consider the stochastic version of the difference equation (2.1) when $k_n = 2$, for all n, defined by the equation

$$x_{n+1} = f_n(x_n) = b(\mu_n, x_n) = \mu_n x_n^2 (1 - x_n), \tag{3.1}$$

with $x_0 \in I = [0,1]$, $\{\mu_n, n \in \mathbb{N}_0\}$ a sequence of independent and identically distributed random variables with support contained in $S = \left(0, \frac{27}{4}\right]$ and common probability density function ϕ.

3.1. Stochastic kernel and asymptotic behaviour

Notice that x_n, for $n \in \mathbb{N}$, defined by (3.1) is an absolutely continuous random variable (with respect to Lebesgue measure). Let f_n be the probability density function of x_n. For each $n \in \mathbb{N}$, the random variables μ_n and x_n are independent and hence their joint probability density function is the product of the individual probability density functions ϕf_n. Let h be an arbitrary bounded function defined in I $(h \in L^\infty(I))$. We have

$$E[h(x_{n+1})] = \int_I h(x) f_{n+1}(x)\, dx, \tag{3.2}$$

and, on the other hand,

$$E[h(x_{n+1})] = E[h(b(\mu_n, x_n))] = \int_I \int_S h(b(u,x))\, \phi(u)\, f_n(x)\, du dx.$$

Letting $y = b(u,x) = ux^2(1-x)$ in the inner integral, we obtain

$$E[h(x_{n+1})] = \int_I \left[\int_0^{\frac{27}{4}x^2(1-x)} h(y)\, \phi\left(\frac{y}{x^2(1-x)}\right) f_n(x) \frac{1}{x^2(1-x)} dy \right] dx. \tag{3.3}$$

Let $\gamma_1 : [0,1] \to \left[0, \frac{2}{3}\right]$ be the inverse function of $\gamma : \left[0, \frac{2}{3}\right] \to [0,1]$ and $\gamma_2 : [0,1] \to \left[\frac{2}{3}, 1\right]$ the inverse function of $\gamma : \left[\frac{2}{3}, 1\right] \to [0,1]$, i.e.,

$$\gamma_1(y) = \frac{1}{3}\left(\sqrt[3]{2\sqrt{y^2-y}-2y+1} + \frac{1}{\sqrt[3]{2\sqrt{y^2-y}-2y+1}} + 1\right)$$

and

$$\gamma_2(y) = -\frac{1}{6}\left(1+i\sqrt{3}\right)\sqrt[3]{2\sqrt{y^2-y}-2y+1} - \frac{1-i\sqrt{3}}{6\sqrt[3]{2\sqrt{y^2-y}-2y+1}} + \frac{1}{3}.$$

The functions γ, γ_1 and γ_2 are represented in Figure 4.

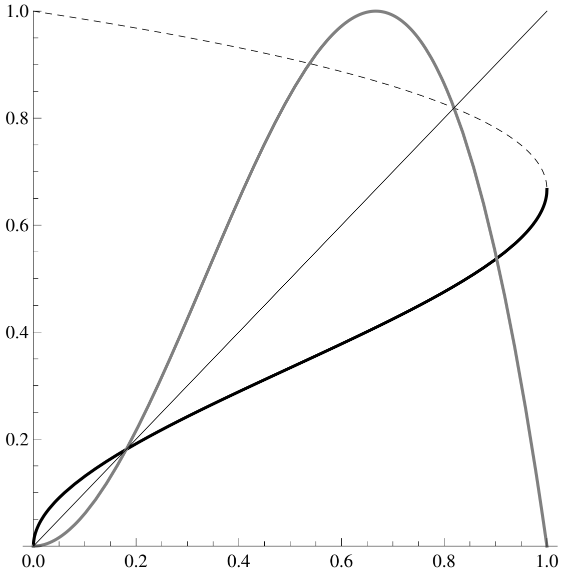


Figure 4: Graphs of γ (grey solid line), γ_1 (black solid line) and γ_2 (dashed line) in the unit interval.

Inverting the integration order in (3.3) we obtain

$$E\left[h\left(x_{n+1}\right)\right] = \int_I h(y)\left[\int_{\gamma_1(y)}^{\gamma_2(y)} \phi\left(\frac{y}{x^2(1-x)}\right) f_n(x) \frac{1}{x^2(1-x)} dx\right] dy. \tag{3.4}$$

Comparing (3.2) and (3.4), since h is arbitrary, it follows that

$$f_{n+1}(y) = \int_I \phi\left(\frac{27}{4}\frac{y}{\gamma(x)}\right) f_n(x) \frac{27}{4}\frac{1}{\gamma(x)} I_{[\gamma_1(y),\gamma_2(y)]}(x)\, dx$$

(where $I_A(v) = 1$ if $v \in A$, $I_A(v) = 0$, otherwise).

It is not difficult to prove that if f_n is supported on $S_n \subseteq I$, then f_{n+1} is supported on $S_{n+1} \subseteq I$.

Let $f \in L^1(I)$, i.e., such that $\int_I |f(x)|\, dx < +\infty$ and $P : L^1(I) \to L^1(I)$ the operator defined by

$$Pf(u) = \int_I L(u,v) f(v)\, dv, \tag{3.5}$$

were L is defined for $(u,v) \in$ on $I \times I$ by

$$L(u,v) = \phi\left(\frac{u}{v^2(1-v)}\right)\frac{1}{v^2(1-v)} I_{[\gamma_1(u),\gamma_2(u)]}(v)\,. \tag{3.6}$$

Notice that

$$\int_I L(u,v)\,du = \int_0^{\frac{27}{4}} \phi(y)\,dy = 1,$$

i.e., L is a stochastic kernel on $I \times I$, since, in addition, $L \geq 0$, and also that

$$P^{n+1} f(u) \;=\; \int_I f(v)\, L_{n+1}(u,v)\,dv$$

with

$$L_{n+1}(v_0, v_{n+1}) \;=\; \int_{I^n} \prod_{i=1}^{n+1} L(v_{i-1}, v_i)\, dv_n ... dv_2 dv_1.$$

In the sequel will study the asymptotically behaviour of the sequence $\{P^n, n \in \mathbb{N}\}$. Suppose ϕ is a bounded probability density function with support $[a,b] \subset \left(0, \frac{27}{4}\right]$ and consider the function

$$h_u(v) = \frac{u}{v^2(1-v)},$$

defined for $v \in (0,1)$ and $u \in I$ (cf. Figure 5 for some graphical examples). The minimum of $h_u(v)$ is obtained when $v = \frac{2}{3}$ and is given by $h_u\left(\frac{2}{3}\right) = u\frac{27}{4}$.

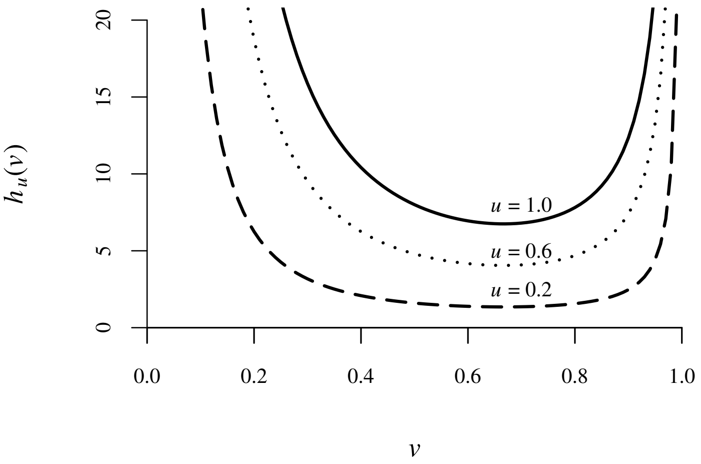


Figure 5: Graphs of h_u when $u = 1$ (solid line), $u = 0.6$ (dotted line) and $u = 0.2$ (dashed line).

Notice that (cf. (3.6))

$$L(u,v) = \phi(h_u(v))\frac{1}{v^2(1-v)} I_{[\gamma_1(u),\gamma_2(u)]}(v)\,.$$

There are three possibilities, for a given u:

1. If u is such that $h_u\left(\frac{2}{3}\right) > b$, i.e., if $u > \frac{4}{27}b$, then $L(u,v) = 0$, for all $v \in I$.

2. If u is such that $a \le h_u\left(\frac{2}{3}\right) \le b$, i.e., if $\frac{4}{27}a \le u \le \frac{4}{27}b$, then

$$L(u,v) = \phi(h_u(v)) \frac{1}{v^2(1-v)} I_{[\gamma_1(u),\gamma_2(u)]\cap V(u)}(v) \le \frac{27}{4}\frac{b}{a}M, \tag{3.7}$$

where $V(u) = \left[\min h_u^{-1}(\{b\}), \max h_u^{-1}(\{b\})\right]$ and $M = \sup\limits_{u,v\in[0,1]} \phi(h_u(v))$.

3. Finally, for u such that $h_u\left(\frac{2}{3}\right) < a$, L is null if $v \notin \{v : a \le h_u(v) \le b\}$, and the same condition (3.7) is obtained.

We can then conclude that $\forall u, v \in I$ we have

$$L(u,v) \le \frac{27}{4}\frac{b}{a}M.$$

Since $\int_I \frac{b}{a}\frac{27}{4}M dx < +\infty$, we have proven the following result (cf. [18], p. 99 and Theorem 5.7.3 in p. 118):

Theorem 3.1. *The sequence $\{P^n, n \in \mathbb{N}\}$, where P is defined by (3.5), is asymptotically periodic.*

This means that there exists a finite sequence of densities $g_1, ...g_r$, a sequence of linear functionals $\lambda_1, ..., \lambda_r$, and a permutation ω of the integers $1, ..., r$ such that

$$Pg_i = g_{\omega(i)}, \quad g_i g_j = 0 \quad \text{for } i \ne j$$

and

$$\lim_{n\to\infty} \left\| P^n f - \sum_{i=1}^{r} \lambda_i(f) g_{\omega^n(i)} \right\| = 0 \quad \text{for} f \in L^1.$$

For better understanding the behaviour of the sequence $\{P^n, n \in \mathbb{N}\}$, where P is defined by (3.5), let the parameters μ_n, for $n \in \mathbb{N}$ from the stochastic difference equation (3.1) be uniform in an interval $C \subseteq S = (0, 27/4]$, i.e., let $\phi(x) = \frac{1}{|C|} I_C(x)$. The asymptotic behaviour of the process depends on the set C. For example, if $C = S$, i.e., if $\phi(x) = \frac{4}{27} I_S(x)$, then at the instant n the system can be in one of the following intervals:

$$E_1 = \left[0, \mathbf{A}_{\frac{27}{4}}\right), \quad E_2 = \left(\mathbf{A}_{\frac{27}{4}}, \frac{1}{2}\right), \quad E_3 = \left[\frac{1}{2}, \frac{1+\sqrt{5}}{4}\right],$$

$$E_4 = \left(\frac{1+\sqrt{5}}{4}, \max f^{-1}\left(\left\{\mathbf{A}_{\frac{27}{4}}\right\}\right)\right), \quad E_5 = \left(\max f^{-1}\left(\left\{\mathbf{A}_{\frac{27}{4}}\right\}\right), 1\right],$$

where, recall, $\mathbf{A}_{\frac{27}{4}} = \frac{9-\sqrt{33}}{18}$. Consider $P_n = [p_{i,j,n}]_{i,j\in\{1,\ldots,5\}}$ where $p_{i,j,n} = P(x_{n+1} \in E_j | x_n \in E_i)$.

We have

$$P_n = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ p_{2,1,n} & p_{2,2,n} & p_{2,3,n} & p_{2,4,n} & p_{2,5,n} \\ p_{3,1,n} & p_{3,2,n} & p_{3,3,n} & p_{3,4,n} & p_{3,5,n} \\ p_{4,1,n} & p_{4,2,n} & p_{4,3,n} & p_{4,4,n} & p_{4,5,n} \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Since $p_{i,j,n} \neq 0$ for $i \in \{2,3,4\}$ and $j \in \{1,..,5\}$, the fixed point zero will attract all points with probability one. Also, if there exists a natural number n_0 such that $p_{i,j,n_0} = 0$, then $p_{i,j,n} = 0$, for all $n \geq n_0$.

On the other hand, if, e.g., $C = (4, 16/3)$ and $x_0 \in E_3$, the system will remain in E_3 (Figure 6 represents two samples of the position of the system after 20000 steps). Hence, in this case, there exists a set of positive Lebesgue measure where the inequality $P^n f > 0$ holds for $n \geq n_0(f)$, for every probability density function, f, with support on the positive real numbers set. Using, e.g., Lemma 1 from [19], we can then conclude the following result:

Corollary 3.1. *If ϕ is the uniform distribution based on a non null subset of $(4, \frac{16}{3})$, the sequence $\{P^n, n \in \mathbb{N}\}$, where P is defined by (3.5) and (3.6), is asymptotically stable, i.e., there exists a probability density function f^* on R^+ such that $Pf^* = f^*$ and*

$$\lim_{x \to \infty} \|P^n f - f^*\| = 0,$$

for any probability density function f on R^+, where $\|.\|$ denotes the norm in L^1.

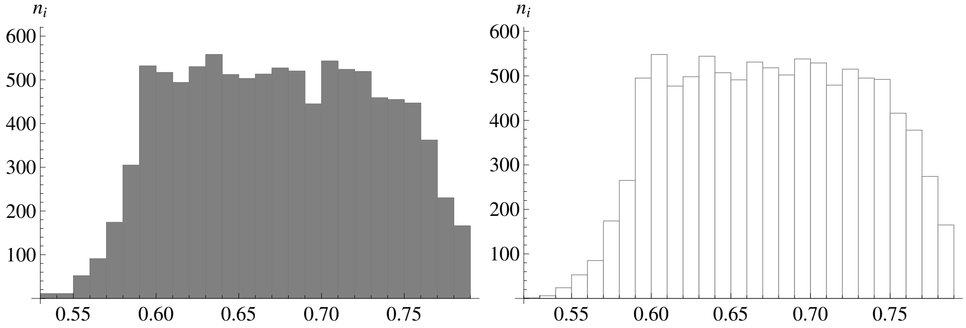


Figure 6: Two samples of size 10000 of the random variable x_{20000} when the sequence μ_n is uniformly distributed in $(4, 10/3)$ and $x_0 = 0.6$.

ACKNOWLEDGMENTS

This work was partially supported by FCT/Portugal through the projects UID/MAT/04459/2013 and UID/MAT/00006/2013.

REFERENCES

[1] Alligood, K.T.; Sauer, T. and Yorke, J. (1996). *Chaos: An Introduction to Dynamical Systems*, Springer, New York.

[2] AlSharawi, Z. and Angelos, J. (2006). On the periodic logistic equation, *Applied Mathematics and Computation*, **180**(1), 342–352.

[3] AlSharawi, Z.; Angelos, J.; Elaydi, S. and Rakesh, L. (2006). An extension of Sharkovsky's theorem to periodic difference equations, *Journal of Mathematical Analysis and Applications*, **316**(1), 128–141.

[4] Alves, J.F. (2009). What we need to find out the periods of a periodic difference equation, *Journal of Difference Equations and Applications*, **15**(8–9), 833–847.

[5] Arnold, L. (1998). *Random Dynamical Systems*, Springer Monographs in Mathematics, Springer, Berlin.

[6] Baladi, V. (2000). *Positive Transfer Operators and Decay of Correlations*, Vol. 16 of Advanced Series in Nonlinear Dynamics, World Scientific.

[7] Briden, W. and Zhang, S. (1994). Stability of solutions of generalized logistic difference equations, *Periodica Mathematica Hungarica*, **29**(1), 81–87.

[8] Coppel, W.A. (1955). The solution of equations by iteration, *Mathematical Proceedings of the Cambridge Philosophical Society*, **51**(1), 41–43.

[9] Devaney, R.L. (2003). *An Introduction to Chaotic Dynamical Systems*, 2nd ed., CRC Press.

[10] Diaconis, P. and Freedman, D. (1999). Iterated random functions, *SIAM Review*, **41**(1), 45–76.

[11] Elaydi, S. (2007). *Discrete Chaos: With Applications in Science and Engineering*, 2nd ed., Chapman & Hall/CRC.

[12] Elaydi, S.; Luís, R. and Oliveira, H. (2013). Local bifurcation in one-dimensional nonautonomous periodic difference equations, *International Journal of Bifurcation and Chaos*, **23**(3), 1350049(1–18).

[13] Grinfeld, M.; Knight, P.A. and Lamba, H. (1996). On the periodically perturbed logistic equation, *Journal of Physics A: Mathematical and General*, **29**(24), 8035–8040.

[14] Hernández-Bermejo, B. and Brenig, L. (2006). Some global results on quasipolynomial discrete systems, *Nonlinear Analysis: Real World Applications*, **7**(3), 486–496.

[15] Keller, G. and Liverani, C. (1999). Stability of the spectrum for transfer operators, *Ann. Scuola Norm. Sup. Pisa Cl. Sci., Série 4*, **28**(1), 141–152.

[16] Kellogg, R.B.; Li, T.Y. and Yorke, J. (1976). A constructive proof of the Brouwer fixed-point theorem and computational results, *SIAM Journal on Numerical Analysis*, **13**(4), 473–483.

[17] Kifer, Y. (1986). *General analysis of random maps.* In "Ergodic Theory Of Random Transformations" (Y. Kifer, Ed.), Birkhäuser Basel, Boston, 7–32.

[18] Lasota, A. and Mackey, M.C. (1994). *Chaos, Fractals, and Noise – Stochastic Aspects of Dynamics*, 2nd ed., Springer.

[19] Lasota, A.; Mackey, M.C. and Tyrcha, J. (1992). The statistical dynamics of recurrent biological events, *Journal of Mathematical Biology*, **30**(8), 775–800.

[20] Levin, S.A. and May, R.M. (1976). A note on difference-delay equations, *Theoretical Population Biology*, **9**(2), 178–187.

[21] Liu, P.D. (2001). Dynamics of random transformations: smooth ergodic theory, *Ergodic Theory and Dynamical Systems*, **21**(5), 1279–1319.

[22] Luís, R. (2017). *Nonautonomous Periodic Difference Equations with Applications to Populations Dynamics and Economics*, Lambert Academic Publishing, Germany.

[23] Luís, R.; Elaydi, S. and Oliveira, H. (2010). Non-autonomous periodic systems with Allee effects, *Journal of Difference Equations and Applications*, **16**(10), 1179–1196.

[24] Luís, R. and Mendonça, S. (2016). A note on the bifurcation point of a randomized Fibonacci model, *Chaotic Modeling and Simulation (CMSIM)*, **4**, 445–458.

[25] Marie, P. and Rousseau, J. (2011). Recurrence for random dynamical systems, *Discrete Continuous Dynamical Systems – A*, **30**(1), 1–16.

[26] Marotto, F.R. (1982). The dynamics of a discrete population model with threshold, *Mathematical Biosciences*, **58**(1), 123–128.

[27] Ohno, T. (1983). Asymptotic behaviors of dynamical systems with random parameters, *Publ. Res. Inst. Math. Sci.*, **19**, 83–98.

[28] Pötzsche, C. (2010). *Geometric Theory of Discrete Nonautonomous Dynamical Systems*, Lecture Notes in Mathematics, Springer.

[29] Rasmussen, M. (2007). *Attractivity and Bifurcation for Nonautonomous Dynamical Systems*, Lecture Notes in Mathematics, Springer.

[30] Zhang, W.-B. (2006). *Discrete Dynamical Systems, Bifurcations and Chaos in Economics*, Mathematics in Science and Engineering, Elsevier, Amsterdam.

REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 87–97

CONFIDENCE INTERVAL FOR QUANTILE RATIO OF THE DAGUM DISTRIBUTION

Authors: Alina Jędrzejczak
– Institute of Statistics and Demography, University of Lodz,
Rewolucji 1905 r. 41/43, 90-214 Łódź, Poland
alina.jedrzejczak@uni.lodz.pl

Dorota Pekasiewicz
– Institute of Statistics and Demography, University of Lodz,
Rewolucji 1905 r. 41/43, 90-214 Łódź, Poland
dorota.pekasiewicz@uni.lodz.pl

Wojciech Zieliński
– Department of Econometrics and Statistics, Warsaw University of Life Sciences,
Nowoursynowska 159, 02-776 Warszawa, Poland
wojciech_zielinski@sggw.edu.pl

Received: March 2018 Revised: September 2018 Accepted: October 2018

Abstract:

- Inequality measures based on ratios of quantiles are frequently applied in economic research, especially to the analysis of income distributions. In the paper, we construct a confidence interval for such measures under the Dagum distribution which has widely been assumed as a model for income and wage distributions in empirical analysis and theoretical considerations. Its properties are investigated on the basis of computer simulations. The constructed confidence interval is further applied to the analysis of income inequality in Poland in 2015.

Keywords:

- *ratio of quantiles; confidence interval; Dagum distribution; quintile share ratio.*

AMS Subject Classification:

- 62F25, 62P20.

1. INTRODUCTION

In the Eurostat regional yearbook (2016), one of the basic measures of income distribution inequality is defined as the income quintile share ratio or the $S80/S20$ ratio. It is calculated as the ratio of total income received by the 20% of the population with the highest income (the top quintile) to that received by the 20% of the population with the lowest income (the bottom quintile), i.e. income quintile share ratio is defined as

$$r_{0.2,0.8} = \frac{F^{-1}(0.8)}{F^{-1}(0.2)},$$

where F denotes the distribution of the population income. The natural estimator of $r_{0.2,0.8}$ is the ratio of appropriate sample quintiles. However, the problem is in interval estimation. According to the best knowledge of the Authors such a problem has never been considered in the literature. In the paper a confidence interval for the population ratio of quintiles is constructed. The proposed confidence interval is based on the asymptotic distribution of the ratio of sample quintiles.

We confine ourselves to the Dagum ([1]) distribution as a probabilistic model of income. The Dagum distribution is widely used for income modeling in many countries all over the world (see for example Domański and Jędrzejczak [5], Jędrzejczak [10]). The Dagum distribution has many good mathematical as well as statistical properties. Basic properties of this distribution are presented in Appendix A; for more see Kleiber ([11]), Dey *et al.* ([4]). See also Encyclopedia ([6]) (pp. 3363–3378, also 3236–3248) and the references therein.

The paper is organized as follows. In the second section confidence interval for a ratio of quantiles is constructed. It is based on the ratio of sample quantiles of the Dagum distribution. It appears that the ends of the proposed confidence interval depend on a shape parameter which should be estimated from a sample. In the third section a short simulation study is provided. In this study two estimators of the shape parameter were applied. Namely, the estimator obtained by the method of moments and the one obtained by the method of probability-weighted moments. Results of the simulations are very similar for these two estimators. In the fourth section an application to income inequality analysis based on the data coming from the Polish Household Budget Survey is presented. In the last section some conclusions are presented as well as some remarks on further research on the subject.

We consider a more general set-up, namely a confidence interval for a ratio of α and β quantiles is constructed. To obtain a confidence interval for the quintile ratio it is enough to put $\alpha = 0.2$ and $\beta = 0.8$. The results of the paper may easily be generalized to other distributions applied in personal income modeling, such as Pareto, Burr Type XII, Beta, etc.

2. CONFIDENCE INTERVAL

Let $0 < \alpha < \beta < 1$ be given numbers and let

$$r_{\alpha,\beta} = \frac{F^{-1}(\beta)}{F^{-1}(\alpha)}$$

be the quantile ratio of interest, where $F(\cdot)$ is the cumulative distribution function (CDF) of income distribution. Let $X_1, ..., X_n$ be a sample of incomes of randomly drawn n individuals. Let $X_{1:n} \leq \cdots \leq X_{n:n}$ denote the ordered sample. As an estimator of $r_{\alpha,\beta}$ it is taken

$$r^*_{\alpha,\beta} = \frac{X_{\lfloor n\beta \rfloor + 1:n}}{X_{\lfloor n\alpha \rfloor + 1:n}},$$

where $\lfloor x \rfloor$ denotes the greatest integer not greater than x.

In our considerations we confine ourselves to the Dagum distribution, i.e. throughout the paper it will be assumed that the distribution of the population income is the Dagum one. As it was mentioned above, the Dagum distribution fits population income quite well for many countries around the world.

Consider the Dagum distribution with parameters $a, v > 0$ and $\lambda > 0$. Its cumulative distribution function (CDF) and probability density function (PDF) are as follows

$$F_{a,v,\lambda}(x) = \left(1 + \left(\frac{x}{\lambda}\right)^{-v}\right)^{-a} \quad \text{for } x > 0$$

and

$$f_{a,v,\lambda}(x) = \frac{av}{\lambda}\left(\frac{x}{\lambda}\right)^{av-1}\left(1 + \left(\frac{x}{\lambda}\right)^{v}\right)^{-a-1} \quad \text{for } x > 0.$$

Its quantile function equals

$$Q_{a,v,\lambda}(q) = \lambda\left(q^{-1/a} - 1\right)^{-1/v} \quad \text{for } 0 < q < 1.$$

For other interesting properties of the Dagum distribution see Appendix A.

The problem is in constructing a confidence interval at the confidence level δ for a ratio of quantiles of the Dagum distribution

$$r_{\alpha,\beta} = \frac{Q_{a,v,\lambda}(\beta)}{Q_{a,v,\lambda}(\alpha)} = \left(\frac{\beta^{-1/a} - 1}{\alpha^{-1/a} - 1}\right)^{-1/v}$$

on the basis of a random sample $X_1, ..., X_n$.

In what follows "large" sample sizes are considered, i.e. it is assumed that $n \to \infty$. There are two reasons for such an approach. The first one is that real sample sizes usually comprise many thousands of observations. The second one is rather technical — the finite sample size distribution of the ratio of sample quantiles of the Dagum distribution is analytically untractable (for exact distribution see Maswadah 2013).

Theorem 2.1. *For $0 < \alpha < \beta < 1$ the random variable $r^*_{\alpha,\beta}$ is strongly consistent estimator of $r_{\alpha,\beta}$, for all a, v, λ.*

Proof: The proof follows form the fact (David and Nagaraja [2]; Serfling [15]) that $X_{\lfloor n\alpha \rfloor + 1:n}$ is strongly consistent estimator of the α's quantile of the underlying distribution. Application of Slutsky theorem gives the thesis. □

Theorem 2.2. *For $0 < \alpha < \beta < 1$ the estimator $r^*_{\alpha,\beta}$ is asymptotically normally distributed random variable.*

Proof: Let $Y_i = \ln X_i$. Of course $Y_{i:n} = \ln X_{i:n}$. Let γ^Y_α and γ^Y_β denote the quantiles of Y. For $\alpha < \beta$ we have (Serfling [15], th. 2.3.3; David and Nagaraja [2], th. 10.3):

$$\sqrt{n}\begin{bmatrix} Y_{\lfloor n\alpha\rfloor+1:n} - \gamma^Y_\alpha \\ Y_{\lfloor n\beta\rfloor+1:n} - \gamma^Y_\beta \end{bmatrix} \to N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix} \frac{\alpha(1-\alpha)}{(f_Y(\gamma^Y_\alpha))^2} & \frac{\alpha(1-\beta)}{(f_Y(\gamma^Y_\alpha)f_Y(\gamma^Y_\beta))} \\ \frac{\alpha(1-\beta)}{(f_Y(\gamma^Y_\alpha)f_Y(\gamma^Y_\beta))} & \frac{\beta(1-\beta)}{(f_Y(\gamma^Y_\beta))^2} \end{bmatrix}\right),$$

where $f_Y(\cdot)$ is the PDF of Y.

Hence

$$\sqrt{n}\left[\left(Y_{\lfloor n\beta\rfloor+1:n} - Y_{\lfloor n\alpha\rfloor+1:n}\right) - \left(\gamma^Y_\beta - \gamma^Y_\alpha\right)\right] \to N\left(0, \sigma^2\right),$$

where

$$\sigma^2 = \frac{\beta(1-\beta)}{\left(f_Y(\gamma^Y_\beta)\right)^2} + \frac{\alpha(1-\alpha)}{\left(f_Y(\gamma^Y_\alpha\ )\right)^2} - 2\frac{\alpha(1-\beta)}{\left(f_Y(\gamma^Y_\beta)f_Y(\gamma^Y_\alpha)\right)}.$$

So we have

$$\sqrt{n}\left(\ln\frac{X_{\lfloor n\beta\rfloor+1:n}}{X_{\lfloor n\alpha\rfloor+1:n}} - \left(\gamma^Y_\beta - \gamma^Y_\alpha\right)\right) \to N\left(0, \sigma^2\right).$$

Applying Delta method (Greene [8], p. 913) with $g(t) = e^t$:

$$\sqrt{n}\left(\frac{X_{\lfloor n\beta\rfloor+1:n}}{X_{\lfloor n\alpha\rfloor+1:n}} - e^{\gamma^Y_\beta - \gamma^Y_\alpha}\right) \to e^{(\gamma^Y_\beta - \gamma^Y_\alpha)}N\left(0, \sigma^2\right).$$

Since in the Dagum distribution $\gamma^Y_\alpha = \ln\gamma_\alpha$ we have

$$\sqrt{n}\left(\frac{X_{\lfloor n\beta\rfloor+1:n}}{X_{\lfloor n\alpha\rfloor+1:n}} - \frac{\gamma_\beta}{\gamma_\alpha}\right) \to \left(\frac{\gamma_\beta}{\gamma_\alpha}\right)N\left(0, \sigma^2\right),$$

i.e.

$$(*) \qquad \sqrt{n}\left(r^*_{\alpha,\beta} - r_{\alpha,\beta}\right) \to r_{\alpha,\beta}N\left(0, \sigma^2\right). \qquad \square$$

Simple calculations show that

$$\sigma^2 = \frac{1}{(av)^2}\left(\frac{1-\beta}{\beta}\frac{1}{(1-\beta^{\frac{1}{a}})^2} + \frac{1-\alpha}{\alpha}\frac{1}{(1-\alpha^{\frac{1}{a}})^2} - 2\frac{1-\beta}{\beta}\frac{1}{(1-\alpha^{\frac{1}{a}})(1-\beta^{\frac{1}{a}})}\right).$$

Since we are interested in the estimation of the ratio $r_{\alpha,\beta}$ of quantiles, we reparametrize the considered model. It can be seen that

$$v = \frac{\log\left(\frac{\alpha^{-1/a}-1}{\beta^{-1/a}-1}\right)}{\log r_{\alpha,\beta}}.$$

The CDF of the Dagum distribution may be written in the following form

$$F_{a,r_{\alpha,\beta},\lambda}(x) = \left(1 + \left(\frac{x}{\lambda}\right)^{-\frac{\log\left(\frac{\alpha^{-1/a}-1}{\beta^{-1/a}-1}\right)}{\log r_{\alpha,\beta}}}\right)^{-a}$$

for $x > 0$ and $a > 0$, $r_{\alpha,\beta} > 0$ and $\lambda > 0$.

We have $\sigma^2 = (\log r_{\alpha,\beta})^2 w^2(a)$, where

$$w^2(a) = \left(\frac{1}{a\,\log\left(\frac{\alpha^{-1/a}-1}{\beta^{-1/a}-1}\right)}\right)^2 \left(\frac{1-\beta}{\beta}\frac{1}{\left(1-\beta^{\frac{1}{a}}\right)^2} + \frac{1-\alpha}{\alpha}\frac{1}{\left(1-\alpha^{\frac{1}{a}}\right)^2} - 2\frac{1-\beta}{\beta}\frac{1}{\left(1-\alpha^{\frac{1}{a}}\right)\left(1-\beta^{\frac{1}{a}}\right)}\right).$$

Let δ be a given confidence level. From ($*$) we have (the scale parameter λ is omitted)

$$P_{a,r_{\alpha,\beta}}\left\{\sqrt{n}\left|\frac{r^*_{\alpha,\beta}-r_{\alpha,\beta}}{w(a)r_{\alpha,\beta}\log r_{\alpha,\beta}}\right| \le u_{(1+\delta)/2}\right\} = \delta,$$

where $u_{(1+\delta)/2}$ is the quantile of $N(0,1)$ distribution.

Solving the above inequality with respect to $r_{\alpha,\beta}$ we obtain confidence interval with the ends

$$\frac{r^*_{\alpha,\beta}z_\pm(a)}{W\left(r^*_{\alpha,\beta}z_\pm(a)\exp\left(z_\pm(a)\right)\right)},$$

where $z_\pm(a) = \frac{\sqrt{n}}{u_{(1\pm\delta)/2}w(a)}$ and $W(\cdot)$ is the Lambert W function (see Appendix B).

Note that the ends of the confidence interval depend on an unknown shape parameter a. This parameter is a nuisance parameter and must be eliminated. There are at least two methods of eliminating such nuisance parameters: estimating or appropriate averaging. In our considerations the shape parameter a is to be estimated. Therefore, a problem arises what estimation method should be chosen. Because theoretical considerations seem to be impossible, a simulation study was carried out.

3. SIMULATION STUDY

The simulation study was performed for different values of quantile ratios $r_{\alpha,\beta}$ and shape parameter a (since scale parameter λ is not important in the problem of ratio of quantiles estimation, it has been set to 1). We take $\alpha = 0.2$, $\beta = 0.8$ and the nominal confidence level equal to 0.95.

From among various methods of parameter estimation for the Dagum distribution (Dey *et al.* [4]) two methods were chosen. The first one is the classical method of moments (MM). In this method theoretical moments of the distribution are compared with the empirical ones. Estimators obtained by this method are solutions of the following system of equations

$$\lambda^m\frac{\Gamma\left(a+\frac{m}{v}\right)\Gamma\left(1-\frac{m}{v}\right)}{\Gamma\left(a\right)} = \frac{1}{n}\sum_{i=1}^{n}x_i^m, \quad \text{for } m=1,2,3.$$

The left-hand side is the m^{th} moment of the Dagum distribution (see Appendix A).

The second method applied in the study was the probability-weighted moments (PWM) (see eg. Hosking *et al.* [9]; Małecka and Pekasiewicz [12]; Pekasiewicz [14]). Probability-weighted moments of the Dagum distribution are equal to (see Appendix A)

$$E_{a,v,\lambda}\left[XF^m_{a,v,\lambda}(X)\right] = \lambda\frac{\Gamma\left((m+1)a+\frac{1}{v}\right)\Gamma\left(1-\frac{1}{v}\right)}{(m+1)\Gamma\left((m+1)a\right)}, \quad \text{for } m \ge 0.$$

Estimators obtained by this method are the solutions of the following system of equations (for $m = 0, 1, 2$)

$$\begin{cases} \dfrac{\lambda\Gamma\left(a+\frac{1}{v}\right)\Gamma\left(1-\frac{1}{v}\right)}{\Gamma\left(a\right)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_{i:n}, \\ \dfrac{\lambda\Gamma\left(2a+\frac{1}{v}\right)\Gamma\left(1-\frac{1}{v}\right)}{2\Gamma\left(2a\right)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \dfrac{(i-1)}{(n-1)} x_{i:n}, \\ \dfrac{\lambda\Gamma\left(3a+\frac{1}{v}\right)\Gamma\left(1-\frac{1}{v}\right)}{3\Gamma\left(3a\right)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \dfrac{(i-1)(i-2)}{(n-1)(n-2)} x_{i:n}. \end{cases}$$

Estimated coverage probabilities based on 10000 repetitions of samples of size $n = 1000$ are given in Table 1 (MM) and Table 3 (PWM). In Table 2 (MM) and in Table 4 (PWM) average lengths of confidence intervals are presented.

Table 1: Coverage probability.

a	$r_{\alpha,\beta}$		
	1.2	1.6	2
0.1	0.9493	0.9492	0.9494
0.5	0.9497	0.9500	0.9530
1.0	0.9501	0.9518	0.9558
1.5	0.9491	0.9496	0.9549
2.0	0.9475	0.9477	0.9492

Table 2: Average length.

a	$r_{\alpha,\beta}$		
	1.2	1.6	2
0.1	0.03793	0.13175	0.24484
0.5	0.03205	0.11137	0.20909
1.0	0.03025	0.10541	0.20010
1.5	0.03014	0.10457	0.19901
2.0	0.03023	0.10442	0.19688

Table 3: Coverage probability.

a	$r_{\alpha,\beta}$		
	1.2	1.6	2
0.1	0.9496	0.9494	0.9494
0.5	0.9496	0.9491	0.9490
1.0	0.9495	0.9497	0.9492
1.5	0.9484	0.9481	0.9486
2.0	0.9479	0.9477	0.9483

Table 4: Average length.

a	$r_{\alpha,\beta}$		
	1.2	1.6	2
0.1	0.03803	0.13182	0.24483
0.5	0.03204	0.11077	0.20529
1.0	0.03020	0.10433	0.19326
1.5	0.03009	0.10393	0.19249
2.0	0.03021	0.10435	0.19325

Since in practical applications the samples usually comprise many thousands of observations (cf. Section 4), in our simulations samples of size 1000 have been used. It appears that such a size may be treated as large enough to do asymptotics: the simulated coverage probability is very close to the nominal confidence level. Of course, for larger sample sizes the coverage probability should be almost equal to the assumed confidence level.

It can also be noticed that whatever method of estimation (the method of moments or of probability-weighted moments) is applied, probability of covering the true value of the quintile share ratio is near the nominal confidence level. It is also seen that the lengths of obtained confidence intervals are similar; it may be concluded that the length does not depend on the applied method of estimation.

It is worth noting that the method of probability-weighted moments has an advantage over the classical method of moments. Namely, the method of moments is applicable for the distributions which have at least three moments, while the method of probability-weighted moments can be applied for the distributions which have at least the expected value (and thus present heavier tails). In the light of the presented results of the simulations, the method of probability-weighted moments may be recommended to the estimation of the shape parameter a of the Dagum distribution in the construction of the confidence interval for quintile share ratio.

4. EXAMPLE OF APPLICATION

In this section we present the application of the inequality measures based on the first and the fourth quintile, i.e. $r_{0.2,0.8}$, to income inequality analysis in Poland. Calculations are based on the sample coming form the Household Budget Survey (HBS) 2015 provided by the Statistics Poland and being the main source of information on income and expenditure of the population of households.

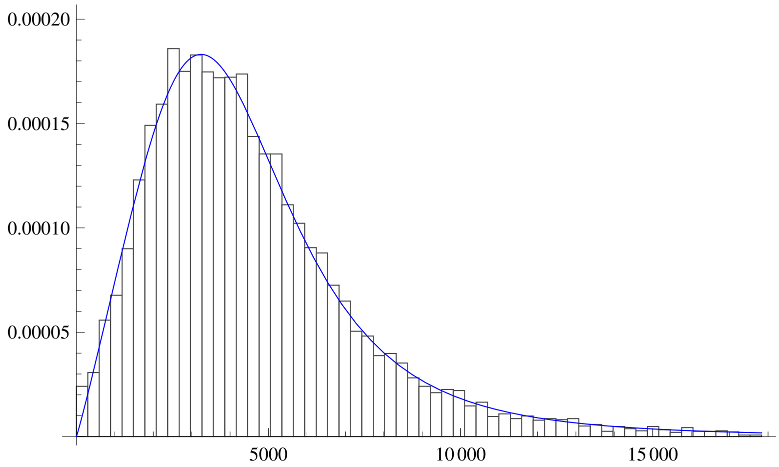


Figure 1: Income distribution in Poland and fitted Dagum distribution ($a = 0.6396$, $v = 3.2403$, $\lambda = 4961.36$).

Within the survey, the sample of size $n = 13420$ was drawn. Firstly, it was checked whether the Dagum distribution fits the data. In Figure 1 the histogram of collected data is shown along with the fitted Dagum distribution (the probability-weighted moments method was applied). The p-value of the standard Kolmogorov-Smirnov test equals 0.8983. Hence it may be concluded that the income distribution in Poland follows the Dagum model.

The sample quintile share ratio $r^*_{0.2,0.8}$ is 2.7600. Application of the formula $(*)$ gives the confidence interval $(2.7081, 2.8160)$ for the population quintile share ratio $r_{0.2,0.8}$ (confidence level equals 0.95). It may be concluded that the income distribution in Poland is quite homogeneous, i.e. the poorest among the richest is about 2.76 times (at least 2.71 but at most 2.82) reacher then the richest among the poorest.

5. CONCLUSIONS

The main goal of the paper was to construct a confidence interval for the ratio of quantiles of the Dagum distribution. According to the best knowledge of the Authors, such a confidence interval has never been constructed. The confidence interval we propose is asymptotic. The first reason for such an approach is lack of finite sample results on the distribution of the ratio of sample quantiles for the Dagum model. Unfortunately, the distribution of the ratio of sample quantiles derived by Maswadah ([13]) was found to be analytically untractable. The second reason for considering asymptotics was that in practise the samples of income are really of large sizes. In a short simulation study it has been shown that sample size of 1000 may be treated as large enough to do asymptotics.

The ends of the obtained asymptotic confidence interval depend on shape parameter a of the Dagum distribution. This parameter should be estimated from a sample. In a simulation study two estimators of this parameter were applied. Both estimators gave similar results.

It will be interesting to check whether the length of the confidence interval depends on the choice of the estimation method (Maximum Likelihood, Method of L-Moments, Method of Maximum Product of Spacings and others) of the shape parameter a. Theoretical solutions seem unavailable, so relevant simulation studies are needed. Such studies are in preparation and will be published separately.

The confidence interval constructed above is symmetrical in the following sense: the risks of underestimation and overestimation are the same. It may also be interesting to consider the problem of constructing the shortest confidence interval. The idea of building such intervals is explained in detail in Zieliński ([16], [17]).

A. APPENDIX

Random variable X follows the Dagum distribution with parameters a, v, λ if its probability density function is given by the formula:

$$f_{a,v,\lambda}(x) = \frac{av}{\lambda}\left(\frac{x}{\lambda}\right)^{av-1}\left(1+\left(\frac{x}{\lambda}\right)^{v}\right)^{-a-1} \quad \text{for } x > 0.$$

Parameters a, v, λ are positive reals. Parameters a and v are shape parameters and λ is a scale parameter.

The distribution is unimodal if $av > 1$. Otherwise it is non-modal. If $av > 1$ the mode value is equal to

$$\lambda\left(\frac{av-1}{v+1}\right)^{\frac{1}{v}}.$$

Moments of the random variable X equal

$$E_{a,v,\lambda}X^m = \lambda^m \frac{\Gamma\left(1-\frac{m}{v}\right)\Gamma\left(a+\frac{m}{v}\right)}{\Gamma\left(a\right)}, \quad \text{for } m < v.$$

Empirical moment from a sample $X_1, ..., X_n$, i.e.

$$\frac{1}{n}\sum_{i=1}^{n} X_i^m$$

is the unbiased estimator of m^{th} moment of the random variable X.

Coefficient of skewness is equal to (for $v > 3$)

$$\frac{\Gamma^2(a)\,\Gamma\big(a+\frac{3}{v}\big)\,\Gamma\big(1-\frac{3}{v}\big) - 3\,\Gamma(a)\,\Gamma\big(a+\frac{1}{v}\big)\,\Gamma\big(a+\frac{2}{v}\big)\,\Gamma\big(1-\frac{2}{v}\big)\,\Gamma\big(1-\frac{1}{v}\big) + 2\,\Gamma^3\big(a+\frac{1}{v}\big)\,\Gamma^3\big(1-\frac{1}{v}\big)}{\Big(\Gamma(a)\,\Gamma\big(a+\frac{2}{v}\big)\,\Gamma\big(1-\frac{2}{v}\big) - \Gamma^2\big(a+\frac{1}{v}\big)\,\Gamma^2\big(1-\frac{1}{v}\big)\Big)^{3/2}}$$

and its kurtosis (for $v > 4$) is

$$\frac{\Gamma^2(a)\,\Big(\Gamma(a)\,\Gamma\big(a+\frac{4}{v}\big)\,\Gamma\big(1-\frac{4}{v}\big) + 3\,\Gamma^2\big(a+\frac{2}{v}\big)\,\Gamma^2\big(1-\frac{2}{v}\big) - 4\,\Gamma\big(a+\frac{1}{v}\big)\,\Gamma\big(a+\frac{3}{v}\big)\,\Gamma\big(1-\frac{3}{v}\big)\,\Gamma\big(1-\frac{1}{v}\big)\Big)}{\Big(\Gamma(a)\,\Gamma\big(a+\frac{2}{v}\big)\,\Gamma\big(1-\frac{2}{v}\big) - \Gamma^2\big(a+\frac{1}{v}\big)\,\Gamma^2\big(1-\frac{1}{v}\big)\Big)^2}.$$

The probability-weighted moments are equal to (for $m \geq 0$ and $v > 1$)

$$E_{a,v,\lambda}\left[XF^m_{a,v,\lambda}(X)\right] = \lambda\frac{\Gamma\left((m+1)a+\frac{1}{v}\right)\Gamma\left(1-\frac{1}{v}\right)}{(m+1)\Gamma\left((m+1)a\right)}.$$

Unbiased estimators (from a sample $X_1, ..., X_n$) of probability-weighted moments are

$$\frac{1}{n}\sum_{i=1}^{n} X_{i:n} \text{ (for } m = 0) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n}\frac{(i-1)\cdots(i-m)}{(n-1)\cdots(n-m)}X_{i:n} \text{ (for } m \geq 1),$$

where $X_{1:n} \leq \cdots \leq X_{n:n}$ are ordered statistics (Hosking *et al.* [9]).

B. APPENDIX

Lambert function $W(\cdot)$ is defined as a solution with the respect to t of the equation

$$te^t = z \quad \Rightarrow \quad t = W(z).$$

It is seen that

$$W(z)e^{W(z)} = z \quad \Rightarrow \quad W(z) = \ln\left(\frac{z}{W(z)}\right) \quad \Rightarrow \quad z = \frac{z}{W(z)}\ln\left(\frac{z}{W(z)}\right).$$

Since the solution with respect to r of the equation $r\ln r = z$ is $r = \frac{z}{W(z)}$, hence

$$A\frac{x-r}{r\ln r} = 1 \quad \Rightarrow \quad Ax = r(\ln r + A) \quad \Rightarrow \quad e^A Ax = \left(re^A\right)\ln\left(re^A\right) \quad \Rightarrow \quad r = \frac{Ax}{W(Axe^A)}.$$

Application of the above to the equation

$$\sqrt{n}\frac{r^*_{\alpha,\beta} - r_{\alpha,\beta}}{w(a)r_{\alpha,\beta}\log r_{\alpha,\beta}} = u_{(1+\delta)/2}$$

gives the confidence interval for the ratio $r_{\alpha,\beta}$.

ACKNOWLEDGMENTS

The authors would like to thank Professor Anna Dembińska (Warsaw University of Technology) for helpful discussion.

REFERENCES

[1] Dagum, C. (1977). A new model of personal income distribution: specification and estimation, *Economie Appliquee*, **30**, 413–437.

[2] David, H.A. and Nagaraja, H.N. (2003). *Order Statistics*, Third Edition, John Wiley & Sons, Inc.

[3] Davidson, R. (2009). Reliable inference for the Gini index, *Journal of Econometrics*, **150**, 30-40.

[4] Dey, S.; Al-Zahrani, B. and Basloom, S. (2017). Dagum distribution: properties and different methods of estimation, *International Journal of Statistics and Probability*, **6**, 74–92.

[5] Domański, Cz. and Jędrzejczak, A. (1998). Maximum likelihood estimation of the Dagum model parameters, *International Advances in Economic Research*, **4**, 243–252.

[6] Encyclopedia (2006). *Encyclopedia of Statistical Sciences*, Second Edition, Volume 5, John Wiley & Sons, Inc.

[7] Eurostat (2016). *The Eurostat Regional Yearbook*, ISBN: 978-92-79-60090-6, ISSN: 2363$\bar{1}$716, `doi: 10.2785/29084`, cat. number: KS-HA-16-001-EN-N, `http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Income_quintile_share_ratio`.

[8] Greene, W.H. (2003). *Econometric Analysis*, 5th ed., Prentice Hall.

[9] Hosking, J.R.M.; Wallis, J.R. and Wood, E.F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, *Technometrics*, **27**, 251–261.

[10] Jędrzejczak, A. (1994). Application of Dagum coefficients in investigating income inequalities in Poland, *Statistical Review*, **41**, 55–66, (in polish).

[11] Kleiber, Ch. (2008). *A guide to the Dagum distributions.* In “Modeling Income Distributions and Lorenz Curves”, Springer.

[12] Małecka, M. and Pekasiewicz, D. (2013). A modification of the probability weighted method of moments and its application to estimate the financial return distribution tail, *Statistics in Transition*, **14**, 495–506.

[13] Maswadah, M. (2013). On the product and ratio of two generalized order statistics from the generalized Burr type-II distribution, *Journal of Mathematics and Statistics*, **9**, 129–136.

[14] Pekasiewicz, D. (2015). *Order Statistics in Estimation Procedures and Their Applications in Socio-economic Research*, University of Lodz (in polish).

[15] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, Inc.

[16] Zieliński, W. (2010). The shortest Clopper–Pearson confidence interval for binomial probability, *Communications in Statistics – Simulation and Computation*, **39**, 188–193.

[17] Zieliński, W. (2017). The shortest Clopper–Pearson randomized confidence interval for binomial probability, *REVSTAT – Statistical Journal*, **15**, 141–153.

REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 99–110

MINIMUM AREA CONFIDENCE REGION FOR WEIBULL DISTRIBUTION BASED ON RECORDS

Authors: Junmei Zhou
– School of Mathematics and Statistics, Hainan Normal University,
Haikou, China

Fen Jiang
– School of Statistics and Mathematics, Yunnan University of Finance and Economics,
Kunming, China

Jin Zhang
– School of Mathematics and Statistics, Yunnan University,
Kunming, China
jinzhang@ynu.edu.cn

Received: September 2017 Revised: August 2018 Accepted: October 2018

Abstract:

- Record values are commonly seen in real life applications, and many important studies on record values relate to Weibull distributions. Based on record values, we establish the minimum area confidence region for the two-parameter Weibull distribution, which is shown to be superior to the classical confidence regions for having smaller expected area.

Keywords:

- *order statistic; parameter; record value; simulation; sufficient statistic.*

AMS Subject Classification:

- 62F25.

1. INTRODUCTION

Weibull distribution has wide applications in survival analysis, reliability engineering, weather forecasting, hydrology, meteorology and insurance (e.g, Murthy *et al.* [16], Ye *et al.* [26]). The cumulative distribution function (cdf) of the two-parameter Weibull distribution, denoted by Weibull(β, η), is

$$F(x; \beta, \eta) = 1 - e^{-(x/\eta)^{\beta}}, \; x > 0,$$

where $\beta > 0$ is the shape parameter and $\eta > 0$ is the scale parameter. In particular, if $\beta = 1$, then the Weibull distribution simplifies as the exponential distribution Exp(η) with mean η, and it becomes the Rayleigh distribution when $\beta = 2$. In the case of $\beta \geq 10$, the shape of Weibull distribution is close to that of the smallest extreme value distribution (e.g, Nelson [17]).

Record values were first introduced by Chandler [11] as special order statistics from random samples, which can be simply described as follows. (For more description, refer to Ahsanullah [1] and Arnold *et al.* [2].) Let $\{X_n, \; n = 1, 2, ...\}$ be an iid (independent and identically distributed) sequence of continuous random samples. Observation X_j is called an upper record if $X_j > X_i$ for each $i < j$. In addition, the record times sequence $\{U_n, \; n \geq 1\}$ is defined by $U_1 = 1$ with probability 1 and $U_n = \min\{j : j > U_{n-1}, \; X_j > X_{U_{n-1}}\}$ for $n \geq 2$. Thus, the sequence $\{X_{U_n}, \; n \geq 1\}$ is called a sequence of upper record statistics. Lower record statistics can be defined analogously.

Record values are commonly seen in real life applications, such as those in meteorology, sports, economics and life tests (e.g., Ahsanullah [1] and Arnold *et al.* [2]), where joint confidence region for unknown parameters is of great practical significance. In the recent years, joint confidence regions based on records were investigated by many authors, and most of their studies on record values are related to Weibull distributions. For references, see, for example, Chan [10], Chen [12], Murthy *et al.* [16], Soliman *et al.* [21], Wu and Tseng [25], Soliman and Al-Aboud [20], Asgharzadeh *et al.* [7], Asgharzadeh and Abdi [3, 4, 5], Teimouri and Nadarajah [22], Wang and Shi [23], Jafari and Zakerzadeh [13], Wang and Ye [24], Zakerzadeh and Jafari [27], and Zhao *et al.* [30].

In the next section, we discuss the classical methods to build joint confidence regions for parameters of Weibull(β, η) distribution, based on (upper) record values. Then the minimum area confidence region (MACR) for (β, η) based on records is established in Section 3 and Section 4. Comparison of these confidence regions is given in Section 5, showing that the proposed MACR is superior to the classical confidence regions for having smaller expected area.

2. CLASSICAL CONFIDENCE REGIONS BASED ON RECORDS

Let $X_{U_1} < X_{U_2} < \cdots < X_{U_n}$ be the upper record values coming from Weibull(β, η). For simplicity, we write X_{U_i} as R_i and let $Y_i = (R_i/\eta)^{\beta}$ $(i = 1, 2, ..., n)$. Then $Y_1 < Y_2 < \cdots < Y_n$ are the first n upper record values from the standard exponential distribution. Arnold *et al.* [2]

showed that $Z_1, ..., Z_n$ are iid from Exp(1), that is, $Z_1, ..., Z_n \overset{\text{iid}}{\sim} \text{Exp}(1)$, where $Z_i = Y_i - Y_{i-1}$ $(i = 1, 2, ..., n;\ Y_0 \equiv 0)$. It follows that for $j = 1, 2, ..., n-1$,

- **(i)** $U_j = 2\sum_{i=1}^{j} Z_i = 2(\frac{R_j}{\eta})^\beta \sim \chi^2_{2j}$, $V_j = 2\sum_{i=j+1}^{n} Z_i = 2[(\frac{R_n}{\eta})^\beta - (\frac{R_j}{\eta})^\beta] \sim \chi^2_{2(n-j)}$ and the two pivotal quantities are independent, where χ^2_m denotes the chi-square distribution with m degrees of freedom;
- **(ii)** $U_j + V_j = 2(\frac{R_n}{\eta})^\beta \sim \chi^2_{2n}$, $\frac{V_j/2(n-j)}{U_j/2j} = \frac{j}{n-j}[(\frac{R_n}{R_j})^\beta - 1] \sim F_{2(n-j),2j}$, and the two pivotal quantities are independent (see Asgharzadeh and Abdi [4], Johnson *et al.* [15], p. 350) where F_{n_1,n_2} stands for the F-distribution with n_1 and n_2 degrees of freedom.

To build a joint confidence region for β and η, we have from (ii) that

$$P\left[F_{2(n-j),2j}(\alpha_1) \le \frac{j}{n-j}[(\frac{R_n}{R_j})^\beta - 1] \le F_{2(n-j),2j}(\alpha_2)\right] = \sqrt{1-\alpha}$$

for $j = 1, 2, ..., n-1$, and

$$P\left[\chi^2_{2n}(\alpha_1) \le 2(\frac{R_n}{\eta})^\beta \le \chi^2_{2n}(\alpha_2)\right] = \sqrt{1-\alpha},$$

where $\alpha_1 = \frac{1-\sqrt{1-\alpha}}{2}$, $\alpha_2 = \frac{1+\sqrt{1-\alpha}}{2}$, $F_{n_1,n_2}(p)$ is the p quantile of F_{n_1,n_2} and $\chi^2_m(p)$ is the p quantile of χ^2_m. Then one type of the classical level $1-\alpha$ confidence region for (β, η) is given by (Asgharzadeh and Abdi [4])

$$A_j : \begin{cases} \dfrac{\log[1 + \frac{n-j}{j}F_{2(n-j),2j}(\alpha_1)]}{\log(R_n/R_j)} \le \beta \le \dfrac{\log[1 + \frac{n-j}{j}F_{2(n-j),2j}(\alpha_2)]}{\log(R_n/R_j)}, \\ R_n[\dfrac{2}{\chi^2_{2n}(\alpha_2)}]^{\frac{1}{\beta}} \le \eta \le R_n[\dfrac{2}{\chi^2_{2n}(\alpha_1)}]^{\frac{1}{\beta}}, \end{cases} \tag{2.1}$$

where $j = 1, 2, ..., n-1$, and each A_j produces a level $1-\alpha$ confidence region for (β, η). Based on Monte Carlo simulation, Asgharzadeh and Abdi [4] observed that $A_{\lfloor \frac{n}{5} \rfloor}$ and $A_{\lfloor \frac{n}{5} \rfloor + 1}$ provide the smallest confidence areas in most cases, where $\lfloor x \rfloor$ denotes the largest integer value smaller than x.

Noticing that $U = 2\beta\sum_{i=1}^{n} \log(R_n/R_i) \sim \chi^2_{2n-2}$ and $V = 2(R_n/\eta)^\beta \sim \chi^2_{2n}$, which are independent, Jafari and Zakerzadeh [13] derived another type of the classical level $1-\alpha$ confidence region for (β, η):

$$B : \begin{cases} \dfrac{\chi^2_{2n-2}(\alpha_1)}{2\sum_{i=1}^{n} \log(R_n/R_i)} \le \beta \le \dfrac{\chi^2_{2n-2}(\alpha_2)}{2\sum_{i=1}^{n} \log(R_n/R_i)}, \\ R_n[\dfrac{2}{\chi^2_{2n}(\alpha_2)}]^{\frac{1}{\beta}} \le \eta \le R_n[\dfrac{2}{\chi^2_{2n}(\alpha_1)}]^{\frac{1}{\beta}}, \end{cases} \tag{2.2}$$

where $\alpha_1 = \frac{1-\sqrt{1-\alpha}}{2}$ and $\alpha_2 = \frac{1+\sqrt{1-\alpha}}{2}$. By simulation study, Jafari and Zakerzadeh [13] concluded that the expected area of the confidence region in (2.2) is smaller than that in (2.1) proposed by Asgharzadeh and Abdi [4].

3. A BASIC THEOREM ON THE MACR

Let $T = T(X)$ be a sufficient statistic of parameter $\theta = (\beta, \eta)$ with pdf (probability density function) $f(t;\theta)$, where $t \in T(\mathcal{X})$, $\theta \in \Theta$. Here, X denotes the random sample with sample space $\mathcal{X}$, and Θ is the parameter space.

According to the Sufficiency Principle in mathematical statistics (e.g., Bickel and Doksum [8], Casella and Berger [9]), we only need to consider the confidence region $C(T)$ based on sufficient statistic $T = T(X)$, without loss of information from the sample X. The purpose of using the sufficient statistic to simplify or reduce the sample X to $T = T(X)$, so that we have $T(\mathcal{X}) = \Theta$ to be used in the following theorem. This theorem creates the MACR for θ under some restriction, where $|C|$ denotes the area of any confidence region C.

Theorem 3.1. *Suppose that for any $\theta \in \Theta$,*

1. *$T = T(X)$ is a sufficient statistic of θ with pdf $f(t;\theta)$, such that $T(\mathcal{X}) = \Theta$;*
2. *There exists some $p(\theta) > 0$, such that $\tilde{f}(T;\theta) = f(T;\theta)/p(\theta)$ is a pivotal quantity;*
3. *The confidence region $C_k(T)$ is defined by*

$$C_k(T) = \{\theta :\ \tilde{f}(T;\theta) \geq k,\ \theta \in \Theta\},$$

where $k > 0$ is the critical value determined by $P[\theta \in C_k(T)] = 1 - \alpha$ for any $\alpha \in (0,1)$.

Then $C_k(T)$ is the level $1-\alpha$ MACR of θ, under restriction

$$\int\limits_{\theta\in C(t)} dt \leq r_k(\theta)|C(\theta)| \tag{3.1}$$

for any $C(T)$ and $\theta \in \Theta$, where $r_k(\theta)= \int\limits_{\theta\in C_k(t)} dt/|C_k(\theta)|$.

Proof: Let $C(T)$ be any level $1-\alpha$ confidence region of θ, satisfying $\int\limits_{\theta\in C(t)} dt \leq r_k(\theta)|C(\theta)|$. Then for any $\theta \in \Theta$,

$$1-\alpha \ \leq\ P[\theta \in C(T)] = \int\limits_{\theta\in C(t)} [f(t;\theta) - kp(\theta)]dt + kp(\theta) \int\limits_{\theta\in C(t)} dt.$$

It follows from $P[\theta \in C_k(T)] = 1 - \alpha$ that

$$\begin{aligned} 0 \ &\leq\ P[\theta \in C(T)] - P[\theta \in C_k(T)] \\ &=\ d_k(\theta) + kp(\theta)\Big[\int\limits_{\theta\in C(t)} dt - \int\limits_{\theta\in C_k(t)} dt\Big] \\ &\leq\ kp(\theta)r_k(\theta)\big[|C(\theta)| - |C_k(\theta)|\big], \end{aligned}$$

where

$$\begin{aligned} d_k(\theta) &= \Big(\int\limits_{\theta \in C(t)} dt - \int\limits_{\theta \in C_k(t)} dt \Big) [f(t;\theta) - kp(\theta)] dt \\ &= \Big(\int\limits_{\theta \in C(t) \cap \overline{C_k(t)}} - \int\limits_{\theta \in C_k(t) \cap \overline{C(t)}} \Big) [f(t;\theta) - kp(\theta)] dt \leq 0 \end{aligned}$$

where $\overline{C}$ denotes the complementary set of C, and $f(t;\theta) - kp(\theta) \geq$ or ≤ 0 if $\theta \in C_k(t)$ or $\theta \in \overline{C_k(t)}$. It follows that $|C(\theta)| - |C_k(\theta)| \geq 0$ or $|C_k(\theta)| \leq |C(\theta)|$ for any $\theta \in \Theta$, which, together with $T(\mathcal{X}) = \Theta$, implies that $|C_k(T)| \leq |C(T)|$ for any T. The proof is complete. □

This theorem extends the basic theorems in Zhang [28, 29], which are valid for building the MACRs of parameters for normal and exponential distributions, but are not for the Weibull(β, η) distribution. By Theorem 3.1, $C_k(T)$ is the level $1-\alpha$ optimal confidence region of θ, minimizing the area of any level $1-\alpha$ confidence region $C(T)$ under the restriction in (3.1). This restricted condition may look complicated, but the MACR $C_k(T)$ does satisfy this condition, due to

$$\int\limits_{\theta \in C_k(t)} dt = r_k(\theta)|C_k(\theta)|.$$

Moreover, there is no need to check which $C(T)$ is under the restriction. The situation is like that of using Lehmann-Scheffé theorem to build the UMVUE (uniformly minimum variance unbiased estimator), without need to check which estimator is unbiased (e.g., Bickel and Doksum [8], Casella and Berger [9]).

A similar theorem was established in Jeyaratnam [14]. The minimum volume confidence region built by Jeyaratnam is based on a pivotal quantity $T(X, \theta)$ such that for each x, $T(x, \theta)$ is a one-to-one map on Θ whose Jacobian J does not depend on θ, and it is optimal for any level $1-\alpha$ confidence region based on the special pivotal quantity.

4. FORMULATION OF THE MACR BASED ON RECORDS

Based on n record values $R_1 < R_2 < \cdots < R_n$ from Weibull(β, η), we now apply Theorem 3.1 to derive the MACR for parameter $\theta = (\beta, \eta)$. Let

$$Z = \sum_{i=1}^{n-1} \log(R_n/R_i).$$

Then (Z, R_n) is a sufficient statistic for (β, η), according to Wang and Ye [24]. Being its equivalent statistic, $T = (Z, \log R_n/Z)$ is also sufficient for (β, η). By Section 2, $U = 2\beta Z \sim \chi^2_{2n-2}$ and $V = 2(\frac{R_n}{\eta})^\beta \sim \chi^2_{2n}$, which are independent. Thus, (U, V) has pdf $f_{\chi^2_{2n-2}}(u) f_{\chi^2_{2n}}(v)$, $u, v > 0$, and the pdf of $T = (T_1, T_2)$ is

$$f(t_1, t_2; \beta, \eta) = f_{\chi^2_{2n-2}}(2\beta t_1) f_{\chi^2_{2n}}[2(\frac{e^{t_1 t_2}}{\eta})^\beta] \Big| \frac{\partial(u,v)}{\partial(t_1,t_2)} \Big|, \ t_1 > 0,$$

where $f_{\chi^2_m}(x)$ $(F_{\chi^2_m}(x))$ denotes the pdf (cdf) of χ^2_m, $U = 2\beta T_1$, $V = 2(\frac{e^{T_1T_2}}{\eta})^\beta$ and Jacobian $|\frac{\partial(u,v)}{\partial(t_1,t_2)}| = 4\beta^2 t_1(\frac{e^{t_1t_2}}{\eta})^\beta$.

Treating $R = (R_1, R_2, ..., R_n)$ as the random sample X in Section 3, we can see that $T = (Z, \log R_n/Z)$ is a sufficient statistic for $\theta = (\beta, \eta)$, satisfying the conditions 1 and 2 in Theorem 3.1, where the pivotal quantity is

$$\tilde{f}(t_1, t_2; \beta, \eta) = f_{\chi^2_{2n-2}}(2\beta t_1) f_{\chi^2_{2n}}[2(\frac{e^{t_1t_2}}{\eta})^\beta] \cdot (2\beta t_1) \cdot 2(\frac{e^{t_1t_2}}{\eta})^\beta, \ t_1 > 0.$$

Hence, the level $1-\alpha$ MACR for (β, η) is $C_k(T) = \{(\beta, \eta) : \ \tilde{f}(T_1, T_2; \beta, \eta) \geq k\}$ or

$$C_k(T) = \{(\beta, \eta) : \ g(\beta Z) + h((R_n/\eta)^\beta) \leq k_\alpha\}, \tag{4.1}$$

where $g(x) = x - (n-1)\log x$ and $h(y) = y - n\log y$ are both convex functions, and k_α is a critical value to be determined.

Let $k(x) \equiv k_\alpha - g(x)$, $\tilde{k} \equiv k_\alpha - h_{\min}$ and $h_{\min} = h(n)$. Then the confidence region in (4.1) can be equivalently expressed as

$$C_k(T) = \begin{cases} g(\beta Z) \leq \tilde{k}, \\ h((R_n/\eta)^\beta) \leq k(\beta Z), \end{cases}$$

for computational purpose. From the property of convex function, $g(\beta Z) \leq \tilde{k}$ is equivalent to $k_1 \leq \beta Z \leq k_2$ with $g(k_1) = g(k_2) = \tilde{k}$, and $h((R_n/\eta)^\beta) \leq k(\beta Z)$ means $k_{11}(\beta Z) \leq (R_n/\eta)^\beta \leq k_{12}(\beta Z)$ with $h(k_{11}(\beta Z)) = h(k_{12}(\beta Z)) = k(\beta Z)$. Finally, the level $1-\alpha$ MACR for (β, η) in (4.1) can be written as

$$C_k(T) = \begin{cases} k_1/Z \leq \beta \leq k_2/Z, \\ R_n/[k_{12}(\beta Z)]^{\frac{1}{\beta}} \leq \eta \leq R_n/[k_{11}(\beta Z)]^{\frac{1}{\beta}}, \end{cases} \tag{4.2}$$

where $g(x) = x - (n-1)\log x$ with $g(k_1) = g(k_2) = \tilde{k}$, $h(y) = y - n\log y$ with $h(k_{11}(\beta Z)) = h(k_{12}(\beta Z)) = k(\beta Z)$, and the critical value k_α is determined by

$$\begin{aligned} 1 - \alpha &= P[(\beta, \eta) \in C_k(T)] \\ &= P[g(\beta Z) + h((R_n/\eta)^\beta) \leq k_\alpha] \\ &= \underset{g(x)+h(y)\leq k_\alpha}{\int_0^\infty \int_0^\infty} 4 f_{\chi^2_{2n-2}}(2x) f_{\chi^2_{2n}}(2y) dx dy \\ &= \int_{k_1}^{k_2} \int_{k_{11}(x)}^{k_{12}(x)} 4 f_{\chi^2_{2n-2}}(2x) f_{\chi^2_{2n}}(2y) dx dy \\ &= \int_{k_1}^{k_2} 2 f_{\chi^2_{2n-2}}(2x)[F_{\chi^2_{2n}}(2k_{12}(x)) - F_{\chi^2_{2n}}(2k_{11}(x))] dx, \end{aligned}$$

where $k_\alpha > g_{\min} + h_{\min}$ and $h(k_{11}(x)) = h(k_{12}(x)) = k_\alpha - g(x)$. A short R code (R Core Team [18]) for computing k_α, k_1, k_2, $k_{11}(x)$, $k_{12}(x)$ in (4.2) is given in Appendix A, where the last integral in the above equation is computed by using Simpson's rule for numerical integration (the interval $[k_1, k_2]$ is split up into 1000 subintervals).

5. COMPARISON OF CONFIDENCE REGIONS

In the statistical literature, the commonly used measure of accuracy for a confidence region is its volume (area). Clearly, the smaller the volume (area), the more accurate the confidence region. To compare the MACR in (4.1) or (4.2) with the classical confidence regions in (2.1) and (2.2), we now discuss their areas as follows.

Given the sample data of upper record values: $R = (R_1, R_2, ..., R_n)$, the area of the classical confidence region in (2.1) is

$$|A_j| = \int_{\frac{\log[1+\frac{n-j}{j}F_{2(n-j),2j}(\alpha_1)]}{\log(R_n/R_j)}}^{\frac{\log[1+\frac{n-j}{j}F_{2(n-j),2j}(\alpha_2)]}{\log(R_n/R_j)}} R_n[(\frac{2}{\chi^2_{2n}(\alpha_1)})^{\frac{1}{\beta}} - (\frac{2}{\chi^2_{2n}(\alpha_2)})^{\frac{1}{\beta}}]d\beta,$$

where the integral can be computed by using Simpson's rule for numerical integration.

Similarly, the area of the classical confidence region in (2.2) is

$$|B| = \int_{\frac{\chi^2_{2n-2}(\alpha_1)}{2\sum_{i=1}^{n}\log(R_n/R_i)}}^{\frac{\chi^2_{2n-2}(\alpha_2)}{2\sum_{i=1}^{n}\log(R_n/R_i)}} R_n[(\frac{2}{\chi^2_{2n}(\alpha_1)})^{\frac{1}{\beta}} - (\frac{2}{\chi^2_{2n}(\alpha_2)})^{\frac{1}{\beta}}]d\beta,$$

and the area of the MACR in (4.1) or (4.2) is

$$\begin{aligned} |C_k(T)| &= \int_{k_1/Z}^{k_2/Z} R_n[(\frac{1}{k_{11}(\beta Z)})^{\frac{1}{\beta}} - (\frac{1}{k_{12}(\beta Z)})^{\frac{1}{\beta}}]d\beta \\ &= \frac{R_n}{Z}\int_{k_1}^{k_2}[(\frac{1}{k_{11}(x)})^{\frac{Z}{x}} - (\frac{1}{k_{12}(x)})^{\frac{Z}{x}}]dx. \end{aligned}$$

Monte Carlo simulation is conducted to compute the expected areas of confidence regions in (2.1), (2.2) and (4.1). Since η is the scale parameter of Weibull(β, η), we can set $\eta = 1$ without loss of generality. We generate $N = 1000$ independent upper record values $R^{(i)} = (R_1^{(i)}, R_2^{(i)}, ..., R_n^{(i)})$ from Weibull$(\beta, 1)$, where $i = 1, 2, ..., N$. Then $\sum_{i=1}^{N} |C(R^{(i)})|/N$ is used to simulate $E|C(R)|$, the expected area of $C(R)$.

Table 1 lists the expected areas of confidence regions in (2.1), (2.2) and (4.1), where A_* stands for the smallest-area confidence region in (2.1), B represents the confidence region in (2.2), and $C_k(T)$ is the MACR in (4.1). We see from Table 1 that the MACR is always the best for having the smallest expected area.

Example 5.1. Roberts [19] gave the monthly maximal of one-hour average concentration of sulfur dioxide in pphm (parts per hundred million) from Long Beach, California, for the years 1956 to 1974. The related upper record values for the month of October is 26, 27, 40 and 41, where $n = 4$ and $R_4 = 41$.

Table 1: Expected areas of confidence regions for (β, η) with $\eta = 1$.

$1-\alpha$	n	Region	β							
			0.25	0.5	1.0	1.2	1.5	2.0	3.0	5.0
0.90	5	A_*	403.8	17.92	6.110	5.250	4.788	4.275	4.054	3.911
		B	336.7	14.93	5.347	4.660	4.127	3.876	3.620	3.565
		$C_k(T)$	315.1	14.01	5.113	4.403	3.890	3.569	3.415	3.364
	10	A_*	214.6	10.17	2.982	2.525	2.253	2.064	1.926	1.893
		B	140.0	6.644	2.384	2.123	1.881	1.756	1.648	1.618
		$C_k(T)$	108.7	6.392	2.232	1.889	1.793	1.651	1.549	1.546
	15	A_*	152.4	6.399	1.970	1.717	1.503	1.364	1.283	1.264
		B	73.04	4.232	1.491	1.291	1.174	1.122	1.068	1.069
		$C_k(T)$	67.13	4.129	1.464	1.267	1.128	1.075	1.031	1.003
	20	A_*	108.1	4.788	1.434	1.233	1.110	1.028	0.973	0.952
		B	55.59	2.741	1.067	0.962	0.900	0.854	0.788	0.780
		$C_k(T)$	45.46	2.605	1.035	0.908	0.876	0.789	0.767	0.744
	30	A_*	76.49	2.806	0.906	0.812	0.741	0.666	0.635	0.629
		B	37.69	1.712	0.684	0.617	0.571	0.539	0.521	0.518
		$C_k(T)$	21.18	1.505	0.659	0.601	0.550	0.517	0.491	0.494
0.95	5	A_*	697.9	27.58	8.752	7.528	6.373	5.800	5.417	5.228
		B	578.3	22.50	7.254	6.413	5.559	5.167	4.787	4.743
		$C_k(T)$	509.4	21.82	6.919	6.050	5.258	4.756	4.471	4.408
	10	A_*	420.0	13.81	4.003	3.490	3.108	2.673	2.538	2.520
		B	228.6	8.852	3.135	2.772	2.490	2.289	2.184	2.142
		$C_k(T)$	198.0	8.648	2.931	2.668	2.342	2.129	2.048	2.004
	15	A_*	336.2	9.370	2.714	2.326	2.006	1.817	1.682	1.670
		B	138.0	6.083	2.057	1.756	1.615	1.494	1.412	1.393
		$C_k(T)$	111.1	5.490	1.911	1.685	1.530	1.420	1.311	1.299
	20	A_*	179.8	7.024	1.998	1.694	1.457	1.356	1.271	1.251
		B	78.08	4.305	1.429	1.275	1.195	1.097	1.038	1.022
		$C_k(T)$	70.62	3.659	1.328	1.218	1.112	1.034	0.988	0.965
	30	A_*	112.8	3.929	1.278	1.093	0.987	0.890	0.845	0.830
		B	50.17	2.508	0.925	0.825	0.766	0.714	0.693	0.675
		$C_k(T)$	39.10	2.265	0.911	0.780	0.722	0.676	0.646	0.643
0.99	5	A_*	3731	71.42	16.96	13.87	11.68	9.785	8.808	8.413
		B	2566	56.98	14.70	12.25	10.13	8.881	8.004	7.626
		$C_k(T)$	1541	45.17	11.68	10.10	8.757	7.533	6.943	6.690
	10	A_*	1022	29.63	7.443	5.990	5.039	4.394	4.002	3.943
		B	568.6	19.05	5.510	4.588	4.046	3.675	3.424	3.349
		$C_k(T)$	434.5	16.75	5.099	4.293	3.778	3.368	3.147	3.054
	15	A_*	750.5	19.30	4.685	3.835	3.263	2.866	2.687	2.600
		B	367.2	11.27	3.356	2.927	2.561	2.321	2.225	2.178
		$C_k(T)$	309.7	10.25	3.104	2.684	2.396	2.160	2.013	2.002
	20	A_*	477.4	13.45	3.479	2.863	2.507	2.159	1.993	1.946
		B	212.1	7.883	2.496	2.174	1.889	1.741	1.634	1.600
		$C_k(T)$	179.1	7.291	2.292	1.968	1.772	1.624	1.533	1.482
	30	A_*	306.3	8.630	2.204	1.892	1.597	1.460	1.330	1.300
		B	115.9	4.887	1.497	1.370	1.239	1.141	1.070	1.068
		$C_k(T)$	79.47	4.170	1.471	1.290	1.125	1.044	1.002	0.981

Chan [10] showed that Weibull(β, η) is a reasonable model for the data set. Then the level 95% MACR for (β, η) in (4) is given by

$$C_k(T) = \{(\beta, \eta) : 0.8979\beta - 3\log(0.8979\beta) + (41/\eta)^{\beta} - 4\beta\log(41/\eta) \leq k_\alpha\}$$

with area 154.908, where $k_\alpha = 1.297$, $k_1 = 0.451$ and $k_2 = 9.640$ are obtained by using the R code in Appendix A.

The level 95% confidence regions for (β, η) in (2.1) are

$$\begin{aligned} A_1 &= \{(\beta,\eta) : 0.5826 \leq \beta \leq 11.9955,\ 41(0.1029)^{\frac{1}{\beta}} \leq \eta \leq 41(1.1318)^{\frac{1}{\beta}}\}, \\ A_2 &= \{(\beta,\eta) : 0.1646 \leq \beta \leq 6.4905,\ 41(0.1029)^{\frac{1}{\beta}} \leq \eta \leq 41(1.1318)^{\frac{1}{\beta}}\}, \\ A_3 &= \{(\beta,\eta) : 0.1720 \leq \beta \leq 58.9824,\ 41(0.1029)^{\frac{1}{\beta}} \leq \eta \leq 41(1.1318)^{\frac{1}{\beta}}\}, \end{aligned}$$

with areas 195.118, 166.671 and 369.396 respectively.

The level 95% confidence region for (β, η) in (2.2) is

$$B = \{(\beta,\eta) : 0.5305 \leq \beta \leq 9.0277,\ 41(0.1029)^{\frac{1}{\beta}} \leq \eta \leq 41(1.1318)^{\frac{1}{\beta}}\}$$

with area 172.502. The plots of the confidence regions for MACR, A_2 and B are displayed in Figure 1, where the MACR has the smallest area and better shape.

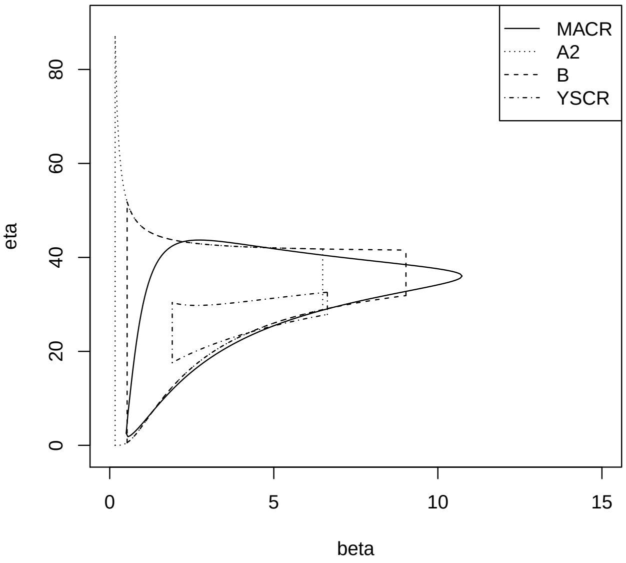


Figure 1: 95% confidence regions MACR, A_2, B and YSCR for (β, η).

For comparison, consider the confidence region (YSCR) of (β, η) in Chen [12] for a complete sample $X = (X_1, X_2, ..., X_n)$. Here the original data set of X is $(n = 19)$

$$26, 14, 27, 15, 16, 16, 11, 10, 14, 12, 15, 40, 29, 13, 20, 41, 31, 28, 11.$$

Then Chen's level 95% confidence region (YSCR) for (β, η) is

$$\begin{cases} 1.9056 \leq \beta \leq 6.6327, \\ (\dfrac{2\sum_{i=1}^{n} X_{(i)}^{\beta}}{60.0972})^{\frac{1}{\beta}} \leq \eta \leq (\dfrac{2\sum_{i=1}^{n} X_{(i)}^{\beta}}{21.2138})^{\frac{1}{\beta}}, \end{cases}$$

which has area 34.2436 and is also plotted in Figure 1. Clearly, the YSCR is much more accurate, but it is based on a complete sample with $n = 19$.

A. APPENDIX: R code for computing k_α, k_1, k_2, $k_{11}(x)$, $k_{12}(x)$ in (4.2)

```
# Compute the critical value k, k1, k2, k11(x), k12(x) at level Pk=1-c.
# n= sample size
f <- function(n,c) {a <- (n-1)*(1-log(n-1))+n*(1-log(n)); b <- 50
g <- function(x) x-(n-1)*log(x)
h <- function(y) y-n*log(y)
# Step 1: find k1 < k2 so that g(k1)=g(k2)=k-h(n).
k1k2 <- function(n,k) {a <- 0; b <- n-1
 kk<- k-(n-n*log(n))
 for (i in 1:50) if (g((a+b)/2)<kk) b <- (a+b)/2 else a <- (a+b)/2
 k1 <- (a+b)/2; a <- n-1; b <- n+100
 for (i in 1:50) if (g((a+b)/2)<kk) a <- (a+b)/2 else b <- (a+b)/2
 k2 <- (a+b)/2; c(k1,k2)}
# Step 2: find k11(x) < k12(x) so that h(k11(x))=h(k12(x))=k-g(x).
k11k12 <- function(n,k,x) {a <- 0; b <- n
 kk<- k-g(x)
 for (i in 1:50) if (h((a+b)/2)<kk) b <- (a+b)/2 else a <- (a+b)/2
 k11x <- (a+b)/2; a <- n; b <- n+100
 for (i in 1:50) if (h((a+b)/2)<kk) a <- (a+b)/2 else b <- (a+b)/2
 k12x <- (a+b)/2; c(k11x,k12x)}
# Step 3: find k so that Pk=1-c.
Int<- function(n,k) {N <- 1000
 K <- k1k2(n,k)
 H <- (K[2]-K[1])/N; df <- 2*(n-1)
 P <- function(x) {
 KK <-k11k12(n,k,x)
 2*dchisq(2*x,df)*(pchisq(2*KK[2],2*n)-pchisq(2*KK[1],2*n))}
 x1 <- K[1]+((1:N)-0.5)*H ; x2 <- K[1]+(1:(N-1))*H
 s1<-0; s2<-0
 for (j in 1:N) s1<- s1+P(x1[j])
 for (j in 1:(N-1)) s2<- s2+P(x2[j])
 Pk <- H/6*(P(K[1])+P(K[2])+4*s1+2*s2); c(Pk,K) }
 for (i in 1:100) {
 R <- Int(n,(a+b)/2)
 if (R[1]<1-c) a <-(a+b)/2 else b<-(a+b)/2}
 k <- (a+b)/2; list(k=k,k1=R[2],k2=R[3])}
```

ACKNOWLEDGMENTS

This research is supported by the National Natural Science Foundation of China (NSFC, Grant No. 11561073, 11461079), and by the Young Teachers Development Program at School of Mathematics and Statistics, Hainan Normal University, China (hnkfj201804).

REFERENCES

[1] Ahsanullah, M. (1995). *Introduction to Record Statistics*, NOVA Science Publishers Inc., Huntington, New York.

[2] Arnold, B.C.; Balakrishnan, N. and Nagaraja, H.N. (1998). *Records*, John Wiley & Sons, New York.

[3] Asgharzadeh, A. and Abdi, M. (2011a). Exact confidence intervals and joint confidence regions for the parameters of the Gompertz distribution based on records, *Pakistan Journal of Statistics*, **27**(1), 55–64.

[4] Asgharzadeh, A. and Abdi, M. (2011b). Joint confidence regions for the parameters of the Weibull distribution based on records, *ProbStat Forum*, **4**, 12–24.

[5] Asgharzadeh, A. and Abdi, M. (2011c). Confidence intervals and joint confidence regions for the Two-Parameter Exponential distribution based on records, *Communications of the Korean Statistical Society*, **18**(1), 103–110.

[6] Asgharzadeh, A. and Abdi, M. (2012). Confidence intervals for the parameters of the Burr Type XII distribution based on records, *International Journal of Statistics and Economics*, **8**, 96–104.

[7] Asgharzadeh, A.; Abdi, M. and Kuş, C. (2011). Interval estimation for the Two-Parameter Pareto distribution based on record values, *Selçuk Journal of Applied Mathematics*, special issue, 149–161.

[8] Bickel, P.J. and Doksum, K.A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics*, Vol. I, 2nd ed., New Jersey, Prentice Hall.

[9] Casella, G. and Berger, R.L. (2002). *Statistical Inference*, 2nd ed., Pacific Grove, CA, Duxbury Press.

[10] Chan, P.S. (1998). Interval estimation of location and scale parameters based on record values, *Statistics and Probability Letters*, **37**, 49–58.

[11] Chandler, K.N. (1952). The distribution and frequency of record values, *Journal of the Royal Statistical Society Series B*, **14**, 220–228.

[12] Chen, Z.M. (1998). Joint estimation for the parameters of Weibull distribution, *Journal of Statistical Planning and Inference*, **66**, 113–120.

[13] Jafari, A.A. and Zakerzadeh, H. (2015). Inference on the parameters of the Weibull distribution using records, *SORT*, **39**(1), 3–18.

[14] Jeyaratnam, S. (1985). Minimum volume confidence regions, *Statistics & Probability Letters*, **3**, 307–308.

[15] Johnson, N.L.; Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1, 2nd ed., Wiley & Sons, New York.

[16] Murthy, D.N.P.; Xie, M. and Jiang, R. (2004). *Weibull Models*, Wiley, Hoboken.

[17] Nelson, W. (1982). *Applied Life Data Analysis*, John Wiley & Sons, INC., New York.

[18] R Core Team (2014). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[19] Roberts, E. (1979). Review of statistics of extreme values with applications to air quality data: Part II. Applications, *Journal of the Air Pollution Control Association*, **29**, 733–740.

[20] Soliman, A.A. and Al-Aboud, F.M. (2008). Bayesian inference using record values from Rayleigh model with application, *European Journal of Operational Research*, **185**, 659–672.

[21] Soliman, A.A.; Abd Ellah, A.H. and Sultan, K.S. (2006). Comparison of estimates using record statistics from Weibull model: bayesian and non-bayesian approaches, *Computational Statistics & Data Analysis*, **51**, 2065–2077.

[22] Teimouri, M. and Nadarajah, S. (2013). Bias corrected MLEs for the Weibull distribution based on records, *Statistical Methodology*, **13**, 12–24.

[23] Wang, L. and Shi, Y.M. (2013). Reliability analysis of a class of exponential distribution under record values, *Journal of Computational and Applied Mathematics*, **239**, 367–379.

[24] Wang, B.X. and Ye, Z.S. (2015). Inference on the Weibull distribution based on record values, *Computational Statistics* & *Data Analysis*, **83**, 26–36.

[25] Wu, J.W. and Tseng, H.C. (2006). Statistical inference about the shape parameter of the Weibull distribution by upper record values, *Statistical Papers*, **48**, 95–129.

[26] Ye, Z.S.; Hong, Y. and Xie, Y. (2013). How do heterogeneities in operating environments affect field failure predictions and test planning, *The Annals of Applied Statistics*, **7**(4), 2249–2271.

[27] Zakerzadeh, H. and Jafari, A.A. (2015). Inference on the parameters of two Weibull distributions based on record values, *Statistical Methods and Applications*, **24**, 25–40.

[28] Zhang, J. (2017). Minimum-volume confidence sets for parameters of normal distributions, *AStA-Advances in Statistical Analysis*, **101**, 309–320.

[29] Zhang, J. (2018). Minimum volume confidence sets for two-parameter exponential distributions, *The American Statistician*, **72**(3), 213–218.

[30] Zhao, X.; Cheng, W.B; Zhang, Y.; Zhang, Q.N. and Yang, Z.H. (2015). New statistical inference for the Weibull distribution, *The Quantitative Methods for Psychology*, **11**(3), 139–147.

REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 111–130

PARAMETER ESTIMATION BASED ON CUMULATIVE KULLBACK–LEIBLER DIVERGENCE

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

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

Received: March 2018 Revised: August 2018 Accepted: October 2018

Abstract:

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

Keywords:

- *estimation; generalized estimating equations; information measures; generalized Pareto distribution; censoring.*

AMS Subject Classification:

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

1. INTRODUCTION

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

$$D\left(f||g\right)=\int\limits_{\mathbb{R}} f\left(x\right)\log\frac{f\left(x\right)}{g\left(x\right)}dx,$$

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

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

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

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

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

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

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

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

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

They also proved that

$$E\left(h\left(X\right)\right)=\int_{0}^{\infty}\bar{F}\left(x;\boldsymbol{\theta}\right)\log\bar{F}\left(x;\boldsymbol{\theta}\right)dx,$$

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

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

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

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

and

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

They proposed a general cumulative Kullback–Leibler information as

$$\operatorname{GCKL}_{\alpha}\left(F:G\right)=\alpha\operatorname{CKL}\left(F:G\right)+\left(1-\alpha\right)\operatorname{CRKL}\left(F:G\right),\quad 0\le\alpha\le1,$$

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

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

2. AN EXTENSION OF MCKLE

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

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

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

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

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

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

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

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

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

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

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

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

If X is nonnegative, then $g(\boldsymbol{\theta})$ in (2.1) reduces to (1.4). So the results of Liu [20], Yari and Saghafi [35], Yari *et al.* [34], Park *et al.* [26] and Hwang and Park [16] yield as special cases. It should be noted that by the law of large numbers F_n converges to $F_{\boldsymbol{\theta}}$ and $\bar{F}_n$ converges to $\bar{F}_{\boldsymbol{\theta}}$ as n tends to infinity. Consequently $\mathrm{CKL}\left(\bar{F}_n||\bar{F}_{\boldsymbol{\theta}}\right)$ converges to zero as n tends to infinity.

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

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

for $x < 0$, and

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

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

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

Using the same steps, we have

$$\int_{0}^{\infty} \bar{F}_n(x) \log \bar{F}(x;\boldsymbol{\theta})\,dx = \frac{1}{n} \sum_{i=k+1}^{n} h\left(x_{(i)}\right).$$

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

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

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

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

In the following, we give some examples.

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

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

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

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

and

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

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

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

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

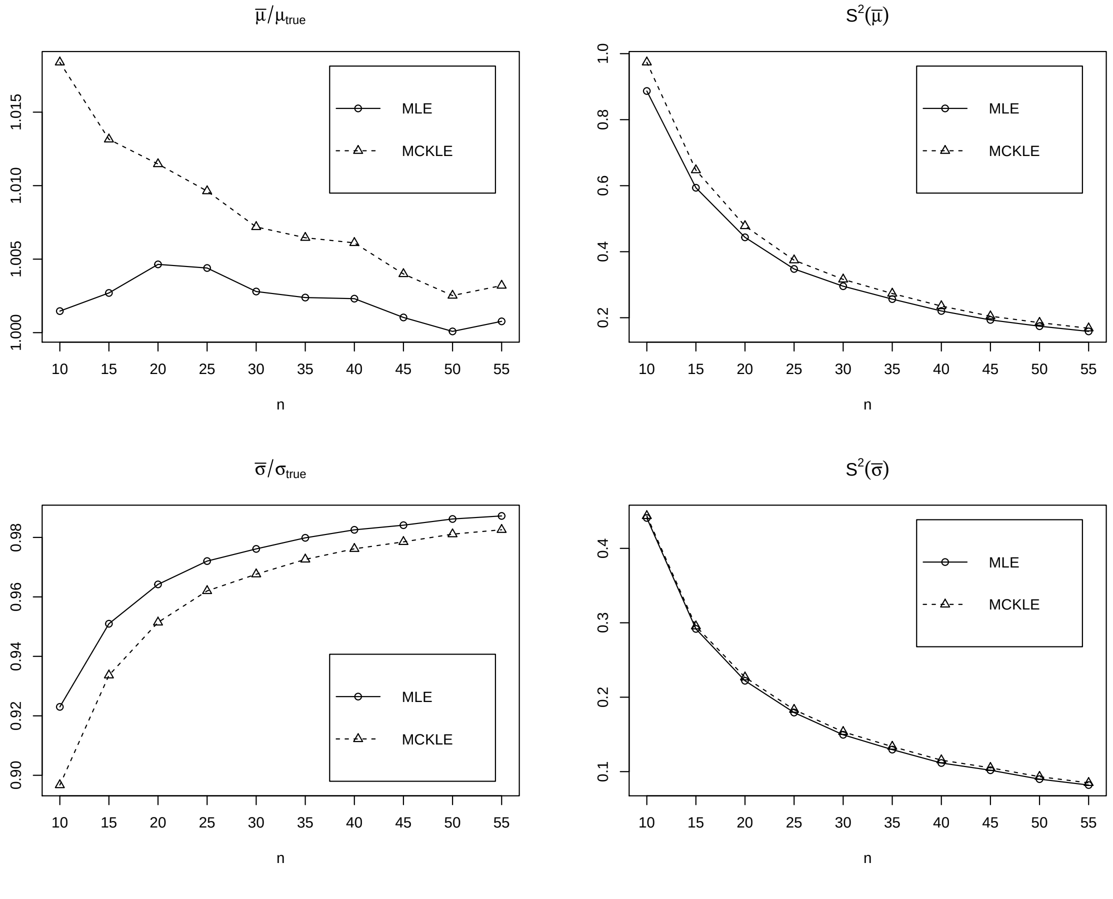


Figure 1: $\bar{\mu}/\mu_{\text{true}}$, $S^2\left(\bar{\mu}\right)$, $\bar{\sigma}/\sigma_{\text{true}}$ and $S^2\left(\bar{\sigma}\right)$ as functions of sample size.

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

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

We simply have MCKLE of θ as

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

This is exactly the moment estimator of θ.

3. ASYMPTOTIC PROPERTIES OF ESTIMATORS

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

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

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

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

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

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

3.1. Consistency and asymptotic normality of the MCKLE

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

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

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

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

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

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

with

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

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

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

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

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

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

with

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

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

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

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

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

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

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

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

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

and

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

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

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

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

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

and

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

So

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

Now, using Theorem 5.13 of Shao [32], σ_{F}^{2} is given as

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

Similar to Theorem 3.2 it can be shown in the case that $\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^{p}$ is vector and $X_{1},...,X_{n}$ are i.i.d., under the conditions of Theorem 5.14 of Shao [32],

$$V_{n}^{-1/2}\left(\widehat{\boldsymbol{\theta}}_{n}-\boldsymbol{\theta}\right)\xrightarrow{d}N_{p}\left(0,I_{p}\right),$$

where $V_{n}=\frac{1}{n}B^{-1}AB^{-1}$ with

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

and

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

provided that B is invertible matrix.

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

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

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

and

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

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

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

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

and

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

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

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

where

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

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

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

and

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

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

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

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

We simply have MCKLE of λ as

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

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

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

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

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

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

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

Notice that using asymptotic bias of $\widehat{\lambda}$, we can find some unbiasing factors to improve our estimator. Since the MLE has inverse Gamma distribution, the unbiased estimator of λ is $\widehat{\lambda}_{um} = (n-1)/n\bar{X}$ [10]. In Liu approach an approximately unbiased estimator of λ is

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

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

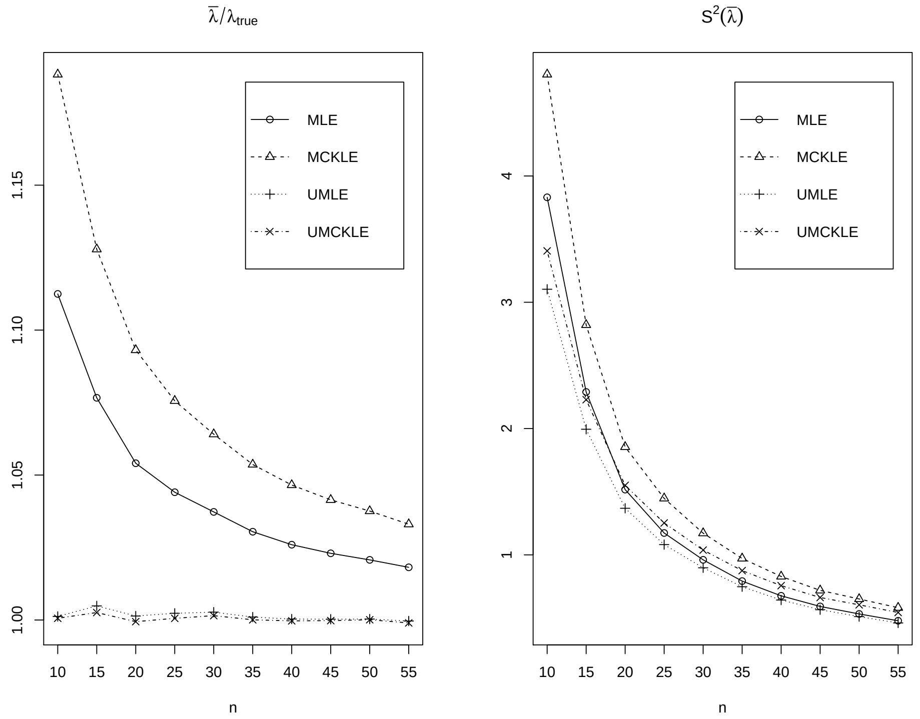


Figure 2: $\bar{\lambda}/\lambda_{\text{true}}$ and $S^2\left(\bar{\lambda}\right)$ as functions of sample size.

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

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

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

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

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

and MCKLE of μ and σ are, respectively,

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

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

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

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

where

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

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

$$g\left(\mu,\sigma\right)=2\sigma\exp\left(\frac{\mu}{\sigma}\right)-\mu-\sigma+\frac{1}{n\sigma}\left[\sum_{\substack{i=k+1\\ x_i\geq 0}}^{n}\frac{x_i^2}{2}-\mu\sum_{\substack{i=k+1\\ x_i\geq 0}}^{n}x_i\right]$$
$$+\frac{\sigma}{n}\left[\sum_{\substack{i=1\\ x_i>0}}^{k}\mathrm{Li}_2\left(\exp\left(-\frac{x_i-\mu}{\sigma}\right)\right)-k\cdot\mathrm{Li}_2\left(\exp\left(\frac{\mu}{\sigma}\right)\right)\right],$$

where $\mathrm{Li}_2\left(\cdot\right)$ is the dilogarithm function. In this case, the MCKLE of μ and σ can be found numerically.

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

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

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

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

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

where

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

and MCKLE estimator $\widehat{k}$ can be found numerically. It should be noted that in this case, for $k\leq-1$, $\widehat{k}$ does not exist. Recently Zhang [37] considered the estimation of for k based on the likelihood method and empirical Bayesian [36], [38]. Denoting the Zhang's estimator by

$\widehat{k}_{\text{Zhang}}$, the cited author shows that the performance of $\widehat{k}_{\text{Zhang}}$ is better than other existing methods for $-6 \leq k \leq 1/2$. In order to compare our estimator ($\widehat{k}_{\text{MCKLE}}$) and Zhang's estimator $\widehat{k}_{\text{Zhang}}$, we evaluated them using simulated samples of sizes $15, 20, 50, 100, 200, 500$ and 1000 with 10000 replicates, considering different true values of the population parameter as $k = -0.75, -0.5, -0.25, 0, 0.25, 0.5, 1, 3, 5$ and 7. Tables 1 and 2 compare bias and root mean squared error (RMSE) of estimators, respectively. It is evident from Table 1 that for all values $k > 0.25$, $\widehat{k}_{\text{MCKLE}}$ has less bias than $\widehat{k}_{\text{Zhang}}$. Also for $k = 0.25$, $n = 15, 20, 500, 1000$, the performance of our estimator is better than the Zhang's estimator. On the other hand, it is seen from Table 2 that except for $k = -0.75$, $n = 100, 200, 500, 1000$, and $k = -0.5$, $n = 500, 1000$, for all values of k, $\widehat{k}_{\text{MCKLE}}$ has less RMSE than $\widehat{k}_{\text{Zhang}}$.

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

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

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

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

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

In this section, we extend MCKLE for the case when the data are collected in censored type I scheme, in continuous case. Some authors such as Lim and Park [18], Cherfi [8], Baratpour and Habibi Rad [2], Park and Shin [27], Park *et al.* [22] Park and Lim [23] and Park and Pakyari [25] studied some forms of KL divergences in different censored data cases. Let $T_1, ..., T_n$ be i.i.d. nonnegative continuous random variables from a c.d.f. F, p.d.f. f and survival function $\bar{F}$. In a variety of applications in biostatistics and life testing, we are only able to observe $X = \min(T, C)$ where C is the constant censoring point. The density function of X can be written as

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

It is known that

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

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

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

and

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

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

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

They also defined the CKL_C of F_n relative to F as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

provided that B is invertible matrix.

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

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

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

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

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

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

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

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

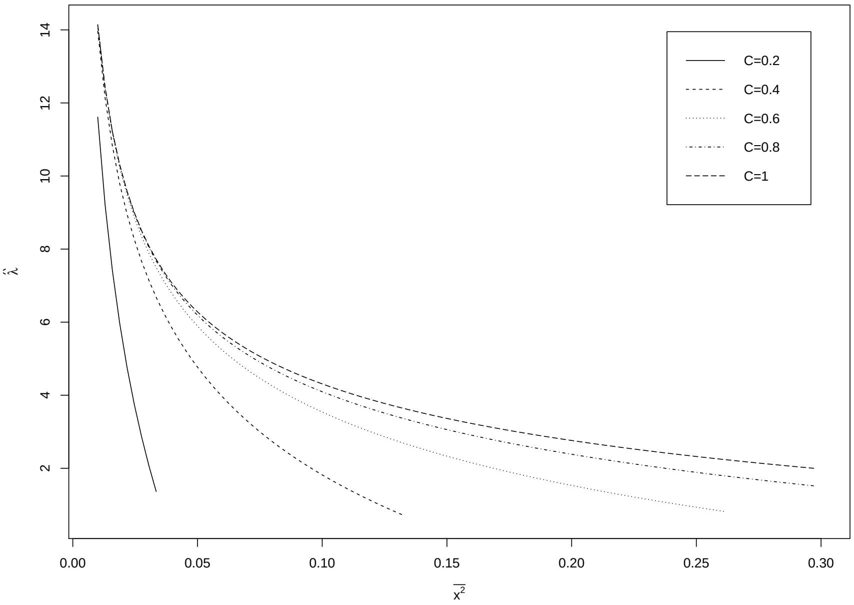


Figure 3: $\widehat{\lambda}$ as a decreasing function of $\overline{x^2}$.

Now, using Theorem 4.1, one can show that

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

where

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

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

ACKNOWLEDGMENTS

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[35] Yari, G. and Saghafi, A. (2012). Unbiased Weibull modulus estimation using differential cumulative entropy, *Communications in Statistics – Simulation & Computation*, **41**(8), 1372–1378.

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

[37] Zhang, J. (2010). Improving on estimation for the generalized Pareto distribution, *Technometrics*, **52**(3), 335–339.

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

REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 131–143

PROOF OF CONJECTURES ON THE STANDARD DEVIATION, SKEWNESS AND KURTOSIS OF THE SHIFTED GOMPERTZ DISTRIBUTION

Author: Fernando Jiménez Torres
– Dpto. de Métodos Estadísticos, Universidad de Zaragoza,
María de Luna 3, Campus Río Ebro,
50018, Zaragoza, Spain
fjimenez@unizar.es

Received: March 2018 Revised: September 2018 Accepted: October 2018

Abstract:

- Three conjectures on the standard deviation, skewness and kurtosis of the shifted Gompertz distribution, as the shape parameter increases to $+\infty$, are proved. In this regard, the exponential integral function and polygamma functions are used in the proofs. In addition, an explicit expression for the ith moment of this probabilistic model is obtained. These results allow to place the shifted Gompertz distribution in the Skewness–Kurtosis diagram, providing a valuable help in the decision to choose the shifted Gompertz distribution among the models to fit data. Their usefulness is illustrated by fitting a real malaria data set using the maximum likelihood method for estimating the parameters of the shifted Gompertz distribution and some classical models. Goodness-of-fit measures are used to compare their performance.

Keywords:

- *shifted Gompertz distribution; moment generating function; skewness–kurtosis diagram; polygamma functions.*

AMS Subject Classification:

- 60E05, 62E15, 26B20.

1. INTRODUCTION

One of the most important references in models of adopting timing of innovations is the model of Bass [1]. From this model, Bemmaor [3] formulated that the individual-level model of adopting timing of a new product in a market is randomly distributed according to the shifted Gompertz distribution. More recently, Lover *et al.* [8] show that modeling studies of period of time to first relapse in human infections with malaria in the New World tropical region, can support the shifted Gompertz distribution.

Some statistical properties of the shifted Gompertz distribution were obtained in Bemmaor [3]. Jiménez Torres and Jodrá [7] gave explicit expressions for the first and second moment, a closed form expression for the quantile function was derived, and the limit distributions of extreme order statistics were considered.

In Jiménez Torres [6] the method of least squares, method of maximum likelihood and method of moments to estimate the parameters of the shifted Gompertz distribution were used. In this paper we want to expand and complete the knowledge and statistical properties of the shifted Gompertz distribution, solving the three conjectures presented in Jiménez Torres and Jodrá [7] and obtaining a general expression for the moments.

Although the Gompertz distribution Z has been given in different forms in the literature, the cumulative distribution function (cdf) $F_Z(z) = P(Z \leq z) = e^{-\alpha e^{-\beta z}}$, $-\infty < z < +\infty$, found in Bemmaor [3], satisfies that its standard deviation, skewness and excess kurtosis are equals to $\pi/(\sqrt{6}\beta)$, $12\sqrt{6}\zeta(3)/\pi^3$ and 2.4, respectively, where $\zeta(\cdot)$ denotes the Riemann zeta function. The skewness of a random variable X is defined by $\gamma_1 = E[(X-\mu)^3]/\sigma^3$ and is a measure of the asymmetry of the probability distribution. The excess kurtosis of X is given by $\gamma_2 = E[(X-\mu)^4]/\sigma^4 - 3$ and it describes the shape of the tails of the probability distribution.

Let X be a random variable having the shifted Gompertz distribution with parameters α and β, where $\alpha > 0$ is a shape parameter and $\beta > 0$ is a scale parameter. The probability density function of X is

$$f_X(x) = \beta e^{-(\beta x + \alpha e^{-\beta x})}\left(1 + \alpha\left(1 - e^{-\beta x}\right)\right), \qquad x > 0. \tag{1.1}$$

This model can be characterized as the maximum of two independent random variables with Gompertz distribution (parameters $\alpha > 0$ and $\beta > 0$) and exponential distribution (parameter $\beta > 0$). From (1.1), given that $\lim_{\alpha\to 0} f_X(x) = \beta e^{-\beta x}$, it may be noted that the shifted Gompertz distribution gets close to an exponential distribution with mean $1/\beta$, as the parameter α decreases to 0. So, for a fixed value of β, $\lim_{\alpha\to 0}\sigma = 1/\beta$, where σ is the standard deviation of X. For the shifted Gompertz distribution we have $\lim_{\alpha\to 0}\gamma_1 = 2$ and $\lim_{\alpha\to 0}\gamma_2 = 6$, which are the skewness and kurtosis of the exponential distribution. If the shape parameter α increases to infinity, the asymptotic behavior of the shifted Gompertz distribution is nontrivial and these limits require analytic tools for their calculation.

Based on numerical evidence showed in Jiménez Torres and Jodrá [7] the next three conjectures were presented:

Conjecture 1: $\lim\limits_{\alpha\to+\infty} \sigma = \dfrac{\pi}{\sqrt{6}\beta}$;

Conjecture 2: $\lim\limits_{\alpha\to+\infty} \gamma_1 = \dfrac{12\sqrt{6}\zeta(3)}{\pi^3}$;

Conjecture 3: $\lim\limits_{\alpha\to+\infty} \gamma_2 = 2.4.$

The remainder of this note is organized as follows. In Section 2, we prove Conjecture 1. In Section 3, we provide an explicit expression for the i-th moment of the shifted Gompertz distribution. In Section 4 and Section 5, we prove Conjecture 2 and Conjecture 3, respectively. In Section 6 we show the importance of these results in the choice of the shifted Gompertz distribution among the models to fit a real data set and finally, the main conclusions are presented in Section 7.

2. PROOF OF CONJECTURE 1

In Jiménez Torres and Jodrá [7] explicit expressions for the moments of orders 1 and 2 of X were obtained. The first moment of X, or mean μ of X, is

$$E[X] = \frac{1}{\beta}\left(\gamma + \log(\alpha) + E_1(\alpha) + \frac{1-e^{-\alpha}}{\alpha}\right), \tag{2.1}$$

where $\gamma \approx 0.57721$ is the Euler–Mascheroni constant and $E_1(x)$ is the exponential integral function, defined by $E_1(x) = \int_x^{+\infty} \frac{e^{-t}}{t}dt$, $x > 0$. The second moment of X is

$$E[X^2] = \frac{2}{\alpha\beta^2}\Big(\gamma + \log(\alpha) + E_1(\alpha) + {}_3F_3[1,1,1;2,2,2;-\alpha]\alpha^2\Big), \tag{2.2}$$

where ${}_3F_3[1,1,1;2,2,2;-\alpha] = \sum_{k=1}^{+\infty} \frac{(-\alpha)^{k-1}}{k!k^2}$ is a generalized hypergeometric function. Moreover, we need the next expression (see Geller and Ng [5]) for $a > 0$ and $b > 0$

$$\int_b^{+\infty} \frac{E_1(ax)}{x}dx = \frac{1}{2}\big((\gamma + \log(ab))^2 + \zeta(2)\big) + \sum_{k=1}^{+\infty} \frac{(-ab)^k}{k!k^2}, \tag{2.3}$$

where $\zeta(2) = \dfrac{\pi^2}{6}$. In particular, using (2.3) with $a = 1$ and $b = \alpha$, we obtain

$$\int_\alpha^{+\infty} \frac{E_1(x)}{x}dx = \frac{1}{2}\big((\gamma + \log(\alpha))^2 + \zeta(2)\big) + \sum_{k=1}^{+\infty} \frac{(-\alpha)^k}{k!k^2}, \tag{2.4}$$

and in the next theorem, we prove Conjecture 1.

Theorem 2.1. *The limit of the standard deviation, σ, of the shifted Gompertz distribution X as the shape parameter α increases to $+\infty$ is finite and its value is*

$$\lim_{\alpha\to+\infty}\sigma = \frac{\pi}{\sqrt{6}\beta}. \tag{2.5}$$

Proof: The variance of a random variable X is $\sigma^2 = E[X^2] - (E[X])^2$. From (2.1), (2.2) and (2.4) we have

$$\begin{aligned}\sigma^2 &= \frac{2}{\alpha\beta^2}\left[\gamma+\log(\alpha)+E_1(\alpha)-\alpha\int_\alpha^{+\infty}\frac{E_1(x)}{x}dx+\frac{\alpha}{2}\big((\gamma+\log(\alpha))^2+\zeta(2)\big)\right]\\ &\quad -\frac{1}{\beta^2}\left(\gamma+\log(\alpha)+E_1(\alpha)+\frac{1-e^{-\alpha}}{\alpha}\right)^2\\ &= \frac{\zeta(2)}{\beta^2}+R(\alpha),\end{aligned} \tag{2.6}$$

where

$$\begin{aligned}R(\alpha) &= \frac{2}{\alpha\beta^2}\big(\gamma+\log(\alpha)+E_1(\alpha)\big)-\frac{2}{\beta^2}\int_\alpha^{+\infty}\frac{E_1(x)}{x}dx\\ &\quad+\frac{1}{\beta^2}\big((\gamma+\log(\alpha)\big)^2-\frac{1}{\beta^2}\left(\gamma+\log(\alpha)+E_1(\alpha)+\frac{1-e^{-\alpha}}{\alpha}\right)^2.\end{aligned} \tag{2.7}$$

So, $\lim\limits_{\alpha\to+\infty}\sigma^2 = \zeta(2)/\beta^2 + \lim\limits_{\alpha\to+\infty}R(\alpha)$. Now, in (2.7) we take limit as α increases to $+\infty$, taking into account the next limits related to the exponential integral function (see Geller and Ng [5]):

$$\lim_{x\to+\infty}\big(\log(x)E_1(x)\big) = \lim_{x\to+\infty}\big(e^{-x}E_1(x)\big) = \lim_{x\to+\infty}\big(x^pE_1(x)\big) = 0. \tag{2.8}$$

So, $\lim\limits_{\alpha\to+\infty}R(\alpha) = 0$, and Conjecture 1 is proved. □

To prove Conjecture 2 and Conjecture 3 we need expressions of the moments of orders 3 and 4, respectively. In Section 3 we are more ambitious and obtain a general expression for the moment of order i of the shifted Gompertz distribution.

3. MOMENT OF ORDER i OF X

The i-th moment of X, denoted and defined by $E[X^i] = \int_0^{+\infty} x^i f_X(x)dx$, $i = 1, 2, ...$, where $f_X(x)$ is given in (1.1), does not seem to have a closed-form expression in terms of elementary functions, but we can find a series expansion. Let $\gamma(a, b)$ be the lower incomplete gamma function defined for any $a > 0$ and $b > 0$ by

$$\gamma(a,b) = \int_0^b v^{a-1}e^{-v}dv, \tag{3.1}$$

and let $M_X(t)$ be the moment generating function of X, i.e., $M_X(t) = E\big[e^{tX}\big]$. In the next theorem we obtain an expression of this function.

Theorem 3.1. *The moment generating function of the shifted Gompertz distribution X for $|t| < \beta$ is*

$$M_X(t) = \alpha^{t/\beta-1}(\alpha + t/\beta)\gamma(1 - t/\beta, \alpha) + e^{-\alpha}. \tag{3.2}$$

Proof: By definition, we have

$$\begin{aligned} E\big[e^{tX}\big] &= \int_0^{+\infty} e^{tx} f_X(x)dx = \beta \int_0^{+\infty} e^{tx-\beta x-\alpha e^{-\beta x}}(1+\alpha(1-e^{-\beta x}))dx \\ &= (1+\alpha)\beta \int_0^{+\infty} e^{tx-\beta x-\alpha e^{-\beta x}}dx - \alpha\beta \int_0^{+\infty} e^{tx-2\beta x-\alpha e^{-\beta x}}dx. \qquad (3.3) \end{aligned}$$

The change of variable $v = \alpha e^{-\beta x}$ in (3.3) provides

$$\begin{aligned} E\big[e^{tX}\big] &= \alpha^{t/\beta-1}(1+\alpha)\int_0^{\alpha} e^{-v}v^{-t/\beta}dv - \alpha^{t/\beta-1}\int_0^{\alpha} e^{-v}v^{1-t/\beta}dv \\ &= \alpha^{t/\beta-1}\big((1+\alpha)\gamma(1-t/\beta,\alpha) - \gamma(2-t/\beta,\alpha)\big). \qquad (3.4) \end{aligned}$$

Integrating by parts in (3.1) yields the recurrence relation $\gamma(a+1,b) = a\gamma(a,b) - b^a e^{-b}$. So, we have

$$E\big[e^{tX}\big] = \alpha^{t/\beta-1}\big((\alpha+t/\beta)\gamma(1-t/\beta,\alpha) + \alpha^{1-t/\beta}e^{-\alpha}\big), \tag{3.5}$$

thereby completing the proof. □

According to Theorem 3.1, the moment generating function of the shifted Gompertz distribution, $M_X(t)$, is finite in the open neighborhood $(-\beta, \beta)$ of 0. In particular, it implies that moments of all orders exist. In the next result, we provide an explicit expression of the moment of order i.

Theorem 3.2. *The moment of order i, $i = 1, 2, ...$, of the shifted Gompertz distribution X is*

$$E[X^i] = \frac{i!}{\beta^i}\left(1 + \sum_{k=1}^{+\infty}\left(\frac{1}{(k+1)^i} - \frac{1}{k^i}\right)\frac{(-\alpha)^k}{k!}\right). \tag{3.6}$$

Proof: Since $M_X(t)$ is finite for t in $(-\beta, \beta)$, it can be expanded in a Taylor series about 0 and the moments of X can be computed by differentiation of $M_X(t)$ at $t = 0$, i.e., $M_X^{(i)}(t)|_{t=0} = M_X^{(i)}(0) = E[X^i]$, $i = 1, 2, ...$, where $M_X^{(i)}(t)$ denotes the i-th derivative of the moment generating function of X. That is,

$$M_X(t) = 1 + \sum_{i=1}^{+\infty}\frac{E[X^i]}{i!}t^i \qquad |t| < \beta. \tag{3.7}$$

Given the Taylor series of the exponential function e^{-v} in (3.1), we have the following series expansion of the lower incomplete gamma function

$$\gamma(a,b) = \int_0^b \sum_{k=0}^{+\infty}(-1)^k\frac{v^{a+k-1}}{k!}dv = \sum_{k=0}^{+\infty}\frac{(-1)^k b^{a+k}}{(a+k)k!}. \tag{3.8}$$

From (3.8), we have

$$\gamma(1-t/\beta,\alpha)=\sum_{k=0}^{+\infty}\frac{(-1)^k\alpha^{1-t/\beta+k}}{(1-t/\beta+k)k!}\qquad |t|<\beta, \tag{3.9}$$

and substituting (3.9) in (3.2), we obtain

$$M_X(t)=(\alpha+t/\beta)\sum_{k=0}^{+\infty}\frac{(-1)^k\alpha^{k}}{(1-t/\beta+k)k!}+e^{-\alpha}\qquad |t|<\beta. \tag{3.10}$$

But the real number $(1-t/\beta+k)^{-1}$ can be expressed as the sum of the terms of a geometric series, i.e.,

$$\frac{1}{1-t/\beta+k}=\frac{1}{k+1}\sum_{i=0}^{+\infty}\frac{t^i}{(k+1)^i\beta^i}\qquad |t|<\beta. \tag{3.11}$$

Finally, substituting (3.11) in (3.10),

$$M_X(t)=(\alpha+t/\beta)\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{(k+1)!}\sum_{i=0}^{+\infty}\frac{t^i}{(k+1)^i\beta^i}+e^{-\alpha}\qquad |t|<\beta. \tag{3.12}$$

Identifying term to term of (3.7) and (3.12), we have

$$E[X^i]=\frac{i!}{\beta^i}\left(1-\sum_{k=1}^{+\infty}\left(\frac{1}{k!k^i}-\frac{1}{(k+1)!(k+1)^{i-1}}\right)(-\alpha)^k\right), \tag{3.13}$$

thereby completing the proof of Theorem 3.2. □

4. PROOF OF CONJECTURE 2

To prove Conjecture 2 we need the next expression (see Geller [4]) for $a>0$ and $\rho>0$

$$\begin{aligned}\int_0^{\rho} e^{-ax}\log^3(x)dx \;=\; &-6\rho\left(\sum_{k=0}^{+\infty}\frac{(-a\rho)^k}{k!(k+1)^4}-\log(\rho)\sum_{k=0}^{+\infty}\frac{(-a\rho)^k}{k!(k+1)^3}\right)\\ &-\frac{3}{a}\log^2(\rho)\Big(\gamma+\log(a\rho)+E_1(a\rho)-\frac{1}{3}\log(\rho)(1-e^{-a\rho})\Big).\end{aligned} \tag{4.1}$$

It may be noted that (4.1) corrects one misprint in Geller [4] (the sign of $\frac{1}{3}\log(\rho)(1-e^{-a\rho})$). In particular, using (4.1) with $a=1$ and $\rho=\alpha$, we have

$$\begin{aligned}\int_0^{\alpha} e^{-x}\log^3(x)dx \;=\; &-6\alpha\left(\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^4}-\log(\alpha)\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^3}\right)\\ &-3\log^2(\alpha)\Big(\gamma+\log(\alpha)+E_1(\alpha)-\frac{1}{3}\log(\alpha)(1-e^{-\alpha})\Big).\end{aligned} \tag{4.2}$$

Moreover, we need the value of (4.2) as α increases to $+\infty$, i.e., $\int_0^{+\infty}e^{-x}\log^3(x)dx$. This integral is $\Gamma^{(3)}(1)$, the third derivative of gamma function evaluated at 1, where the gamma function is defined by $\Gamma(p)=\int_0^{+\infty}t^{p-1}e^{-t}dt$, for a real number $p>0$. To know the value of $\Gamma^{(3)}(1)$ we can use the digamma function, define by $\psi(p) = \Gamma'(p)/\Gamma(p)$ and polygamma functions, $\psi'(p)$, $\psi^{(2)}(p)$, $\psi^{(3)}(p)$ .... These functions are derivatives of the logarithm of the gamma function. In particular, we have $\psi(1) = -\gamma$ and $\psi^{(n)}(1) = (-1)^{n+1}n!\zeta(n+1)$, for $n = 1, 2, 3...$ (see, e.g. [9, 5.15.2]). Using this relation we have $\psi'(1) = \zeta(2)$ and $\psi^{(2)}(1) = -2\zeta(3)$. So, the value of $\Gamma^{(3)}(1)$ is

$$\Gamma^{(3)}(1) = (\psi(1))^3 + 3\psi(1)\psi'(1) + \psi^{(2)}(1) = -\gamma^3 - 3\gamma\zeta(2) - 2\zeta(3), \tag{4.3}$$

where $\zeta(3) \approx 1.20205$ is a real number known as Apéry's constant. In the next theorem, we prove Conjecture 2.

Theorem 4.1. *The limit of the coefficient of skewness, γ_1, of the shifted Gompertz distribution X as the shape parameter α increases to $+\infty$ is finite and its value is*

$$\lim_{\alpha\to+\infty} \gamma_1 = \frac{12\sqrt{6}\zeta(3)}{\pi^3}. \tag{4.4}$$

Proof: The coefficient of skewness of X is

$$\gamma_1 = E[(X-\mu)^3]/\sigma^3 = \big(E[X^3] - 3\mu E[X^2] + 2\mu^3\big)/\sigma^3. \tag{4.5}$$

We can study every term of this equation. The first term of (4.5) is $E[X^3]$. According to (3.6), the moment of order 3 of X is

$$E[X^3] = \frac{3!}{\beta^3}\left(1 + \sum_{k=1}^{+\infty}\left(\frac{1}{(k+1)^3} - \frac{1}{k^3}\right)\frac{(-\alpha)^k}{k!}\right). \tag{4.6}$$

From (4.2), we have

$$\begin{aligned}\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^4} &= -\frac{1}{6\alpha}\Bigg(\int_0^\alpha e^{-x}\log^3(x)dx - 6\alpha\log(\alpha)\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^3}\\ &\quad +3\log^2(\alpha)\big(\gamma+\log(\alpha)+E_1(\alpha) - \frac{1}{3}\log(\alpha)(1-e^{-\alpha})\big)\Bigg).\end{aligned} \tag{4.7}$$

Given that $\sum_{k=1}^{+\infty}\frac{(-\alpha)^k}{k!k^i} = -\alpha\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^{i+1}}, i = 0, 1, 2, ...$, from (2.4), (4.6) and (4.7)

$$\begin{aligned}E[X^3] &= -\frac{1}{\beta^3}\Bigg[6\big(\alpha^{-1}+\log(\alpha)\big)\left(\int_\alpha^{+\infty}\frac{E_1(x)}{x}dx - \frac{1}{2}\big((\gamma+\log(\alpha))^2+\zeta(2)\big)\right)\\ &\quad + \int_0^\alpha e^{-x}\log^3(x)dx + 3\log^2(\alpha)\big(\gamma+\log(\alpha)+E_1(\alpha)\big) - \log^3(\alpha)(1-e^{-\alpha})\Bigg].\end{aligned} \tag{4.8}$$

Now, we study $-3\mu E[X^2]$, the second term of (4.5). From (2.1) and (2.2), it is

$$\begin{aligned}-3\mu E[X^2] &= -\frac{6}{\alpha\beta^3}\left(\gamma+\log(\alpha)+E_1(\alpha)+\frac{1-e^{-\alpha}}{\alpha}\right)\\ &\quad\times\left(\gamma+\log(\alpha)+E_1(\alpha) - \alpha\sum_{k=1}^{+\infty}\frac{(-\alpha)^k}{k!k^2}\right),\end{aligned} \tag{4.9}$$

and from (2.4), we have

$$\begin{aligned}-3\mu E[X^2] &= -\frac{6}{\alpha\beta^3}\left(\gamma+\log(\alpha)+E_1(\alpha)+\frac{1-e^{-\alpha}}{\alpha}\right)\\ &\times\left(\gamma+\log(\alpha)+E_1(\alpha)-\alpha\int_\alpha^{+\infty}\frac{E_1(x)}{x}dx+\frac{\alpha}{2}\left((\gamma+\log(\alpha))^2+\zeta(2)\right)\right).\end{aligned} \tag{4.10}$$

The third term of (4.5) is $2\mu^3$. From (2.1), it is

$$2\mu^3=\frac{2}{\beta^3}\left(\gamma+\log(\alpha)+E_1(\alpha)+\frac{1-e^{-\alpha}}{\alpha}\right)^3. \tag{4.11}$$

Finally, taking into account the three terms of (4.5), i.e., (4.8), (4.10) and (4.11), that $\lim_{\alpha\to+\infty}\int_0^\alpha e^{-x}\log^3(x)dx=\Gamma^{(3)}(1)$ given in (4.3) and the limits (2.8), we have

$$\lim_{\alpha\to+\infty}E[(X-\mu)^3]=\frac{2\zeta(3)}{\beta^3}. \tag{4.12}$$

According to Theorem 2.1, $\lim\limits_{\alpha\to+\infty}\sigma^3=\dfrac{\pi^3}{6\sqrt{6}\beta^3}$, and Conjecture 2 is proved. □

5. PROOF OF CONJECTURE 3

To prove Conjecture 3 we need the next expression (see Geller [4]), valid for $a>0$, $\rho>0$, $p>-1$ and $n=0,1,2,3,\ldots$

$$\int_0^\rho x^pe^{-ax}\log^n(x)dx=(-1)^nn!\rho^{p+1}\sum_{k=0}^n\frac{(-1)^k\log^k(\rho)}{k!}\sum_{l=0}^{+\infty}\frac{(-a\rho)^l}{l!(p+l+1)^{n-k+1}}. \tag{5.1}$$

In particular, we need (5.1) for $a=1$, $\rho=\alpha$, $p=0$ and $n=4$, i.e.,

$$\begin{aligned}\int_0^\alpha e^{-x}\log^4(x)dx &= 4!\alpha\sum_{k=0}^4\frac{(-1)^k\log^k(\alpha)}{k!}\sum_{l=0}^{+\infty}\frac{(-\alpha)^l}{l!(l+1)^{5-k}}\\ &= 4!\alpha\left[\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^5}-\log(\alpha)\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^4}\right.\\ &\quad+\frac{\log^2(\alpha)}{2}\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^3}-\frac{\log^3(\alpha)}{3!}\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^2}\\ &\quad\left.+\frac{\log^4(\alpha)}{4!}\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)}\right].\end{aligned} \tag{5.2}$$

Moreover, we need the value of (5.2) as α increases to $+\infty$, i.e., $\int_0^{+\infty}e^{-x}\log^4(x)dx$, the 4-th Euler–Mascheroni integral. This integral is $\Gamma^{(4)}(1)$, the fourth derivative of $\Gamma(p)$, evaluated at $p=1$. Given that $\psi^{(3)}(1)=6\zeta(4)$, $\big(\zeta(2)\big)^2=5\zeta(4)/2$ and $\zeta(4)=\pi^4/90$, the value of $\Gamma^{(4)}(1)$ is

$$\begin{aligned}\Gamma^{(4)}(1) &= \big(\psi(1)\big)^4+6\psi'(1)\big(\psi(1)\big)^2+4\psi^{(2)}(1)\psi(1)+\psi^{(3)}(1)+3\big(\psi'(1)\big)^2\\ &= \gamma^4+6\gamma^2\zeta(2)+8\gamma\zeta(3)+\frac{27}{2}\zeta(4).\end{aligned} \tag{5.3}$$

In the next theorem, we prove Conjecture 3.

Theorem 5.1. *The limit of the excess kurtosis, γ_2, of the shifted Gompertz distribution X as the shape parameter α increases to $+\infty$ is finite and its value is*

$$\lim_{\alpha\to+\infty}\gamma_2 = 2.4. \tag{5.4}$$

Proof: The excess kurtosis of X is

$$\gamma_2 = E[(X-\mu)^4]/\sigma^4 - 3 = (E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4)/\sigma^4 - 3. \tag{5.5}$$

We can study every term of this equation. The first term of (5.5) is $E[X^4]$. According to (3.6), the fourth moment of X is

$$\begin{aligned} E[X^4] &= \frac{4!}{\beta^4}\Big(1+\sum_{k=1}^{+\infty}\Big(\frac{1}{(k+1)^4}-\frac{1}{k^4}\Big)\frac{(-\alpha)^k}{k!}\Big)\\ &= \frac{4!}{\alpha\beta^4}\Big(\alpha\sum_{k=1}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^4}+\alpha^2\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^5}\Big). \end{aligned} \tag{5.6}$$

From (5.2), we have

$$\begin{aligned} \sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^5} &= \frac{1}{24\alpha}\Big[\int_0^\alpha e^{-x}\log^4(x)dx + 24\alpha\log(\alpha)\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^4}\\ &\quad -12\alpha\log^2(\alpha)\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^3}+4\alpha\log^3(\alpha)\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)^2}\\ &\quad -\alpha\log^4(\alpha)\sum_{k=0}^{+\infty}\frac{(-\alpha)^k}{k!(k+1)}\Big]. \end{aligned} \tag{5.7}$$

From (4.7) and (5.7),

$$\begin{aligned} E[X^4] &= \frac{24}{\alpha\beta^4}\Big[-\frac{1}{6}\int_0^\alpha e^{-x}\log^3(x)dx - \frac{1}{2}\log^2(\alpha)\big(\gamma+\log(\alpha)+E_1(\alpha)\big)\\ &\quad -\frac{1}{6}\log^3(\alpha)(1-e^{-\alpha}) - \log(\alpha)\Big(\int_\alpha^{+\infty}\frac{E_1(x)}{x}dx - \frac{1}{2}\big((\gamma+\log(\alpha)\big)^2\\ &\quad +\zeta(2)\big)\Big) + \frac{\alpha}{24}\Big(\int_0^\alpha e^{-x}\log^4(x)dx - 4\log(\alpha)\int_0^\alpha e^{-x}\log^3(x)dx\\ &\quad -12\log^2(\alpha)\int_\alpha^{+\infty}\frac{E_1(x)}{x}dx + 6\log^2(\alpha)\big((\gamma+\log(\alpha))^2+\zeta(2)\big)\Big)\\ &\quad -8\log^3(\alpha)\big(\gamma+\log(\alpha)+E_1(\alpha)\big)+3\log^4(\alpha)(1-e^{-\alpha})\Big]. \end{aligned} \tag{5.8}$$

Now, we study $-4\mu E[X^3]$, the second term of (5.5). From (2.1) and (4.8), it is

$$\begin{aligned} -4\mu E[X^3] &= -\frac{4}{\beta^4}\Big(\gamma+\log(\alpha)+E_1(\alpha)+\frac{1-e^{-\alpha}}{\alpha}\Big)\\ &\quad\times\Big[-6\big(\alpha^{-1}+\log(\alpha)\big)\Big(\int_\alpha^{+\infty}\frac{E_1(x)}{x}dx-\frac{1}{2}\big((\gamma+\log(\alpha))^2+\zeta(2)\big)\Big)\\ &\quad -\int_0^\alpha e^{-x}\log^3(x)dx - 3\log^2(\alpha)\big(\gamma+\log(\alpha)+E_1(\alpha)\big)+\log^3(\alpha)(1-e^{-\alpha})\Big)\Big]. \end{aligned} \tag{5.9}$$

The third term of (5.5) is $6\mu^2E[X^2]$. From (2.1), (2.2) and (2.4), it is

$$\begin{aligned} 6\mu^2E[X^2] &= \frac{12}{\beta^4}\left(\gamma+\log(\alpha)+E_1(\alpha)+\frac{1-e^{-\alpha}}{\alpha}\right)^2 \\ &\times\left[\frac{\gamma+\log(\alpha)+E_1(\alpha)}{\alpha}-\int_\alpha^{+\infty}\frac{E_1(x)}{x}dx+\frac{1}{2}\left((\gamma+\log(\alpha))^2+\zeta(2)\right)\right]. \end{aligned} \tag{5.10}$$

The fourth and last term of (5.5) is $-3\mu^4$. From (2.1), it is

$$-3\mu^4=-\frac{3}{\beta^4}\left(\gamma+\log(\alpha)+E_1(\alpha)+\frac{1-e^{-\alpha}}{\alpha}\right)^4. \tag{5.11}$$

Finally, taking into account that $\lim_{\alpha\to+\infty}\int_0^\alpha e^{-x}\log^4(x)dx=\Gamma^{(4)}(1)$ given in (5.3), the four terms of (5.5), i.e., (5.8), (5.9), (5.10) and (5.11), and the limits (2.8), we have

$$\lim_{\alpha\to+\infty}E[(X-\mu)^4]=\frac{27\zeta(4)}{2\beta^4}. \tag{5.12}$$

According to Theorem 2.1, $\lim\limits_{\alpha\to+\infty}\sigma^4=\dfrac{\pi^4}{36\beta^4}$. Given the value of $\zeta(4)=\dfrac{\pi^4}{90}$, Conjecture 3 is proved. □

6. REAL DATA APPLICATION

One of the human malaria parasites with the widest geographic distribution in the world is *plasmodium vivax*. If a patient was not fully cured or insufficiently treated, he can relapse in a few weeks after the initial infection, i.e., new clinical symptoms begin after the disease disappeared from the blood following the primary infection. In this section, we have considered an application with periods of time to first relapse or recurrence in 38 patients located at Brazil. We have chosen Brazil since it is located geographically in the New World tropical region, where following Lover *et al.* [8], the shifted Gompertz distribution is suitable for modeling times to first relapse. Tropical region is delimited by the ±23.5° latitude lines. Table 1 shows times (days) to first relapse observed, reported in Battle *et al.* [2].

Table 1: Real data set: Times (days) to first relapse observed (malaria parasite *plasmodium vivax*) in 38 patients located at Brazil.

31	32	32	33	34	35	37	37	44	45	48	53	57	57	58	62	63	64	68
69	70	70	70	71	75	78	80	82	83	86	91	97	97	112	124	132	158	185

According to Theorem 4.1, the values of γ_1 of the shifted Gompertz distribution are greater than $12\sqrt{6}\zeta(3)/\pi^3\approx 1.1395$, i.e., are always positive and possibly this can be a good model to fit a data set with positive asymmetry. Similarly, according to Theorem 5.1, the values of γ_2 of the shifted Gompertz distribution are greater than 2.4, i.e., are always positive. This means that the shifted Gompertz distribution is a fat-tailed probability distribution, and possibly it can be a good model to fit a data set with positive excess kurtosis.

The results proved in this paper allow to place the shifted Gompertz distribution in the Skewness–Kurtosis diagram (see Vargo *et al.* [11]). This moment-ratio diagram (see Figure 1) is a plot containing the (γ_1, γ_2) values for probability distributions. When a probabilistic model has no shape parameter (for example, normal, logistic, Gompertz, exponential or Gumbell distribution, among other), its locus in this diagram corresponds to a point. When a probabilistic model has one shape parameter (for example, log-logistic, gamma, Weibull, Lindley, Lomax or shifted Gompertz distribution, among other), its locus in this diagram corresponds to a curve. In this diagram, the shifted Gompertz distribution starts at the locus of the exponential distribution and ends at the locus of Gompertz distribution. Also, in Figure 1 there is a curve representing the frontier $\gamma_2 \geq \gamma_1^2 - 2$ for all distributions (see Stuart and Ord [10]).

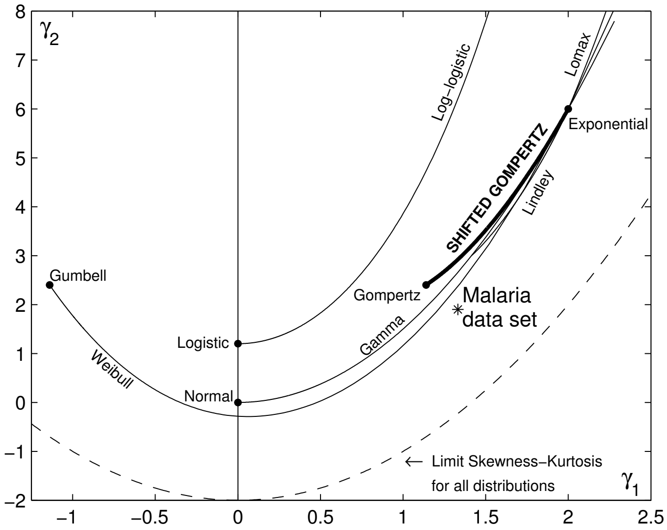


Figure 1: Skewness (γ_1) versus excess kurtosis (γ_2) for some probabilistic models and the locus of the malaria data set.

Given the observed values of skewness and excess kurtosis for malaria data set ($\gamma_1 = 1.3317$, $\gamma_2 = 1.9009$), we can place it in this diagram (see Figure 1) and use it as valuable help in model selection (see chosen models in Table 2).

Table 2: Models and their cumulative distribution functions $F(x)$.

Model	Shape parameter	$F(x)$
Exponential: $\mathrm{E}(\lambda)$	—	$1 - e^{-\lambda x}$
Gamma: $\mathrm{G}(\alpha, \beta)$	α	$\gamma(\alpha, x/\beta)/\Gamma(\alpha)$
Gompertz: $\mathrm{GO}(\alpha, \beta)$	—	$e^{-\alpha e^{-\beta x}}$
Gumbell: $\mathrm{GU}(\alpha, \beta)$	—	$1 - e^{-\alpha e^{\beta x}}$
Lindley: $\mathrm{LD}(\theta)$	θ	$1 - (1 + \theta + \theta x)e^{-\theta x}/(1 + \theta)$
Logistic: $\mathrm{LG}(\mu, s)$	—	$(1 + e^{-(x-\mu)/s})^{-1}$
Log-logistic: $\mathrm{LL}(\lambda, p)$	p	$1 - (1 + (\lambda x)^p)^{-1}$
Lomax: $\mathrm{LO}(\alpha, \beta)$	α	$1 - (1 + \beta x)^{-\alpha}$
Normal: $\mathrm{N}(\mu, \sigma)$	—	$\Phi((x - \mu)/\sigma)$
Weibull: $\mathrm{W}(\alpha, \beta)$	α	$1 - e^{-(x/\beta)^\alpha}$
Shifted Gompertz: $\mathrm{SG}(\alpha, \beta)$	α	$e^{-\alpha e^{-\beta x}}(1 - e^{-\beta x})$

It is reasonable to think that models located relatively near the locus of malaria data set (for example, Weibull, gamma, Gompertz or shifted Gompertz distribution) can provide a better fit than models located farther away (for example, Gumbell, logistic, normal or Lomax, among other). To accept or rejected this surmise, we estimate the parameters of the shifted Gompertz distribution and of all models represented in Figure 1 by the maximum likelihood method. We obtain the performance of each model based on the following goodness-of-fit measures: log-likelihood function (LogL), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Kolmogorov–Smirnov (K-S) statistic with the corresponding p-value, Cramer von Mises (W*) and Anderson–Darling (A*).

The results obtained (see Table 3) show that the shifted Gompertz distribution presents the best fit in almost all goodness-of-fit measures. The smallest values of −LogL, AIC, BIC, W* and A* correspond to the shifted Gompertz distributions. The best values of K-S and its p-value are obtained by gamma and shifted Gompertz distribution. In addition, Weibull, gamma or Gompertz distribution present, in general, better fit than Gumbell, logistic, normal, Lindley, exponential, log-logistic or Lomax distribution.

Table 3: The MLEs of the parameters and goodness-of-fit tests.

Malaria data set									
Model	MLE parameter		−LogL	AIC	BIC	K-S	p-val(K-S)	W*	A*
E	0.0139	—	200.290	402.580	404.218	0.351	10^{-4}	1.113	5.749
G	4.965	14.414	183.079	370.158	373.434	**0.089**	**0.923**	0.054	0.448
GO	10.024	0.040	182.983	369.967	373.242	0.102	0.823	0.047	0.425
GU	0.126	0.022	198.302	400.605	403.880	0.226	0.041	0.492	2.899
LD	0.0275	—	189.961	381.923	383.561	0.220	0.049	0.386	2.359
LG	67.520	18.091	186.779	377.558	380.834	0.117	0.673	0.056	0.667
LL	0.015	3.821	221.181	446.363	449.638	0.103	0.811	1.606	8.397
LO	0.114	99.634	278.846	561.693	564.968	0.600	10^{-12}	3.477	16.062
N	71.578	34.505	188.481	380.963	384.238	0.138	0.461	0.169	1.156
W	2.202	81.129	185.522	375.045	378.320	0.113	0.714	0.107	0.779
SG	8.709	0.040	**182.759**	**369.518**	**372.793**	0.101	0.831	**0.046**	**0.419**

Best fitting model is shown in bold.

7. CONCLUSIONS

Three conjectures on the standard deviation, skewness and kurtosis of the shifted Gompertz distribution, as the shape parameter α increases to $+\infty$, have been proved, solving the asymptotic problems found in Jiménez Torres and Jodrá [7]. In addition, an explicit expression for the i-th moment of the shifted Gompertz distribution has been obtained. These results allow to place the shifted Gompertz distribution in the Skewness–Kurtosis diagram, starting at the locus of the exponential distribution and ending at the locus of Gompertz distribution. To check their usefulness, a real malaria data set has been fitted, estimating the parameters by maximum likelihood. The results obtained show that the shifted Gompertz distribution presents a very good fit among the analyzed models, suggesting that the results proved in this paper can play an important rule in the decision to choose this model to fit data.

ACKNOWLEDGMENTS

The author would like to thank to the editors and an anonymous referee for their valuable comments, corrections and suggestions which improved the quality of the manuscript.

REFERENCES

[1] Bass, F.M. (1969). A new product growth model for consumer durables, *Management Science*, **15**(5), 215–227.

[2] Battle, K.E.; Karhunen, M.S.; Bhatt, S.; Gething, P.W.; Howes, R.E.; Golding, N.; Van Boeckel, T.P.; Messina, J.P.; Shanks, G.D.; Smith, D.L.; Baird, J.K. and Hay, S.I. (2014). Geographical variation in *plasmodium vivax* relapse, *Malaria Journal*, **13**, 144.

[3] Bemmaor, A.C. (1994). *Modeling the diffusion of new durable goods: word-of-mouth effect versus consumer heterogeneity.* In "Research Traditions in Marketing" (G. Laurent, G.L. Lilien and B. Pras, Eds.), Boston, MA, Kluwer, 201–223.

[4] Geller, M. (1963). A table of integrals involving powers, exponentials, logarithms, and the exponential integral, *Jet Propulsion Laboratory*, Technical Report No. 32-469.

[5] Geller, M. and Ng, E.W. (1969). A table of integrals of the exponential integral, *Journal of Research of the National Bureau of Standards – B. Mathematics and Mathematical Science*, **73B**(3), July–September.

[6] Jiménez Torres, F. (2014). Estimation of parameters of the shifted Gompertz distribution using least squares, maximum likelihood and moments methods, *Journal of Computational and Applied Mathematics*, **256**, 867–877.

[7] Jiménez Torres, F. and Jodrá, P. (2009). A note on the moments and computer generation of the shifted Gompertz distribution, *Communications in Statistics – Theory and Methods*, **38**, 75–89.

[8] Lover, A.A.; Zhao, X.; Gao, Z.; Coker, R.L. and Cook, A.R. (2014). The distribution of incubation and relapse times in experimental human infections with the malaria parasite *plasmodium vivax*, *BMC Infectious Diseases*, **14**, 539.

[9] Olver, F.W.J.; Lozier, D.W.; Boisvert, R.F. and Clark, C.W. (Eds.) (2010). *NIST Handbook of Mathematical Functions*, Cambridge University Press.

[10] Stuart, A. and Ord, K. (1994). *Kendall's Advanced Theory of Statistics*, Volume I, NY, Hodder Arnold, New York.

[11] Vargo, E.; Pasupathy, R. and Leemis, L.M. (2010). Moment-ratio diagrams for univariate distributions, *Journal of Quality Technology*, **42**(3), 1–11.

REVSTAT – Statistical Journal
Volume 19, Number 1, January 2021, 145–160

MIXED DOUBLE-RANKED SET SAMPLING: A MORE EFFICIENT AND PRACTICAL APPROACH

Authors: Monjed H. Samuh
– Mathematics & Statistics Department, King Fahd University of Petroleum & Minerals, Saudi Arabia
monjedsamuh@ppu.edu

M. Hafidz Omar
– Mathematics & Statistics Department, King Fahd University of Petroleum & Minerals, Saudi Arabia
omarmh@kfupm.edu.sa

M. Pear Hossain
– Mathematics & Statistics Department, King Fahd University of Petroleum & Minerals, Saudi Arabia
pearhossain35@bsmrstu.edu.bd

Received: April 2018 Revised: July 2018 Accepted: December 2018

Abstract:

- A new modification of ranked set sampling (RSS) is investigated to estimate the mean of the study population. This modified approach is a double-stage approach and a kind of combination between RSS and median RSS (MRSS). It is shown that this new modification is more efficient than of RSS, MRSS, and simple random sampling. The Hellinger distance is used to show that the new approach is more practical than any other double-stage RSS.

Keywords:

- *efficiency; Hellinger distance; median; practicality; ranked set sampling.*

AMS Subject Classification:

- 62D05, 62G08.

1. INTRODUCTION

Ranked set sampling (RSS), a data collection scheme, was first implemented by [9] as a good competitor to simple random sampling (SRS) scheme to estimate the mean of Australian pasture yields in agricultural experimentation. Due to its importance to other situations and for a variety of applications in statistics [9] is reprinted in [10]. RSS scheme has recently been getting some attention from researchers working in statistical process control. [11] and [12] for example, proposed different run rules for control charts under different RSS schemes. [19] studied the EWMA control chart for monitoring linear profiles under various RSS schemes. For discussions of some other situations where RSS found applications, see [17], [4], [18], [14], [5], and [13].

[9, 10] claimed that the RSS mean is an unbiased estimator of the population mean and the variance of the RSS mean is smaller than in simple random sampling (SRS) with equal measurement elements. This sampling scheme is useful when it is difficult to measure large number of elements but visually (without inspection) ranking some of them is easier. It involves randomly selecting m sets (each of size m elements) from the study population. The elements of each set are ordered with regards to the study variable, say $\mathbf{X}$, by any negligible cost method or visually without measurements. Finally, the i^{th} minimum from the i^{th} set, $i = 1, 2, ..., m$, are identified for measurement. The obtained sample is called a ranked set sample of set size m. It is worth to observe that visual ranking with large set size is prone to ranking errors. In practice, the set size should be small ($m = 2$, 3, or 4). For more details see [1], [8], and [21].

[25] provided the mathematical theory behind the claims of [9, 10]. They proved the following identities:

1. $$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_{X_{(i)}}(x),$$

2. $$\mu = \frac{1}{m}\sum_{i=1}^{m}\mu_i,$$

3. $$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\sigma_i^2 + \frac{1}{m}\sum_{i=1}^{m}(\mu_i - \mu)^2,$$

where μ is the mean and σ^2 is the variance of the study population $f(x)$ and μ_i and σ_i^2 are the mean and the variance of the i^{th} ordered statistic. They also showed that the efficiency of the RSS mean *with respect to* (w.r.t.) SRS, defined by the ratio of the variances of the two sample means, is bounded by 1 and $\frac{m+1}{2}$. In particular, when the study population is degenerate then the efficiency is 1, and when the study population is uniform then the efficiency is $\frac{m+1}{2}$.

As claimed by [9, 10] it is later shown in the literature that estimators calculated based on RSS are more efficient than their counterpart in SRS. For example, [24] showed that the empirical distribution function based on RSS is more efficient than its counterpart in SRS. Some authors estimate the parameters of a specific distribution using RSS, see for example [2] and [22].

For improving the efficiency of estimators, some variations of RSS were proposed. [1] suggested double RSS (DRSS), as a method that improves efficiency of the RSS estimators while keeping m fixed. They reported that the RSS mean estimator is less efficient than that based on DRSS. Median RSS (MRSS) is a modification of RSS proposed by [15] to decrease ranking error and to improve the efficiency of the estimators being estimated. The procedure of MRSS is similar to RSS but in lieu of identifying the i^{th} minimum from the i^{th} set only the median of each set is identified. Given odd set size m, the $\left(\frac{m+1}{2}\right)^{\text{th}}$ smallest element is identified from each set for measurement. When m is even, from the first $\frac{m}{2}$ sets the $\left(\frac{m}{2}\right)^{\text{th}}$ smallest element is identified for measurement and from the second $\frac{m}{2}$ sets the $\left(\frac{m}{2}+1\right)^{\text{th}}$ smallest element is identified for measurement. [20] suggested a double MRSS (DMRSS) as an alternative procedure to improve the efficiency of the sample mean. They compared the DMRSS with SRS, RSS, DRSS, and some other sampling schemes and found that DMRSS is the most efficient scheme.

In the process of DMRSS, the data points are identified based on the data points of MRSS. For example, if m is odd, the data points of the DMRSS are just the medians of the data points of MRSS; that is, the data points of DMRSS are the medians of the medians of the SRS. It is clear that identifying median of the medians is a hard process, and this contradict the nature of RSS schemes which require visual comparison without inspection (a rationale originally mentioned by [9]). On the other hand, in the process of DRSS, the data points are identified based on the data points of the RSS. For example, the first data point of DRSS is the minimum of the RSS data points, which is easy to be identified visually without inspection. [1] have shown by the degree of distinguishability and the probability of perfect ranking that ranking an independent and identically (iid) data points is harder than ranking ordered (but independent) data points. Thus, getting a DMRSS is harder than a DRSS. In other words, DRSS is more practical than DMRSS.

To improve the efficiency of RSS estimators, we suggest to combine MRSS scheme with RSS scheme; that is, to apply the method of MRSS on the obtained RSS data points. We shall call this method by mixed double-ranked set sampling (MxDRSS).

Section 2 introduces notations and some basic results. MxDRSS is clarified in Section 3. The practicality of this method is discussed and compared with other methods in Section 4. Estimation of the population mean based on MxDRSS is investigated in Section 5. Numerical results for specific distributions are presented in Section 6. Finally, Section 8 concludes the paper.

2. NOTATION AND SOME BASIC RESULTS

Let X be a continuous random variable with cumulative distribution function (cdf) $F(x)$, probability density function (pdf) $f(x)$, mean μ, and variance σ^2. Let $X_1, X_2, ..., X_m$ be a SRS from $f(x)$, then X_i are iid as $f(x)$. Note that when $f(x)$ is infinite, SRS and random sample are used synonymly.

Suppose $Y_1^{(1)}, Y_2^{(1)}, ..., Y_m^{(1)}$ be a RSS; that is $Y_i^{(1)}$ is the i^{th} order statistic of the random sample $X_1, X_2, ..., X_m$, where the superscript (1) represents stage 1. The cdf of Y_i (see for example [3]) is given by

$$F_{Y_i}(y) = F_{X_{(i)}}(y) = \sum_{k=i}^{m} \binom{m}{k} F^k(y)\,(1-F(y))^{m-k}\,,\; i = 1, 2, ..., m, \tag{2.1}$$

and the pdf of Y_i is

$$f_{Y_i}(y) = m\binom{m-1}{i-1} F^{i-1}(y)\,(1-F(y))^{m-i}\, f(y),\; i = 1, 2, ..., m. \tag{2.2}$$

Let $Y_1^{(2)}, Y_2^{(2)}, ..., Y_m^{(2)}$ be a DRSS; that is $Y_i^{(2)}$ is the i^{th} order statistic of the RSS $Y_1^{(1)}, Y_2^{(1)}, ..., Y_m^{(1)}$ and each of $Y_i^{(2)}$ are obtained from independent ranked set samples of size m. Apparently, $Y_1^{(2)}, Y_2^{(2)}, ..., Y_m^{(2)}$ are the order statistics of the independent (not identical) random variables $Y_1^{(1)}, Y_2^{(1)}, ..., Y_m^{(1)}$. Hence, the cdf of $Y_i^{(2)}$ (see for example [6]) is given by

$$F_{Y_i^{(2)}}(y) = \sum_{l=i}^{m} \sum_{S_l} \left(\prod_{k=1}^{l} F_{Y_{j_k}^{(1)}}(y) \prod_{k=l+1}^{m} \left(1 - F_{Y_{j_k}^{(1)}}(y)\right)\right), \tag{2.3}$$

where S_l is the set of the entire permutations $(j_1, j_2, ..., j_m)$, of the integers $(1, 2, ..., m)$ for which $j_1 < j_2 < \cdots < j_l$, and $j_{l+1} < j_{l+2} < \cdots < j_m$ ([6]). The pdf of $Y_i^{(2)}$ is the derivative of $F_{Y_i^{(2)}}(y)$.

Let $W_1^{(1)}, W_2^{(1)}, ..., W_m^{(1)}$ be a MRSS; that is

$$W_i^{(1)} = \begin{cases} X_{(\frac{m+1}{2})} & \text{if } m \text{ is odd } \&\, i = 1, ..., m, \\ X_{(\frac{m}{2})} & \text{if } m \text{ is even } \&\, i = 1, ..., \frac{m}{2}, \\ X_{(\frac{m+2}{2})} & \text{if } m \text{ is even } \&\, i = \frac{m+2}{2}, ..., m. \end{cases} \tag{2.4}$$

The pdf of $W_i^{(1)}$ is

$$f_{W_i^{(1)}}(x) = \begin{cases} f_{X_{(\frac{m+1}{2})}}(x) & \text{if } m \text{ is odd } \&\, i = 1, ..., m, \\ f_{X_{(\frac{m}{2})}}(x) & \text{if } m \text{ is even } \&\, i = 1, ..., \frac{m}{2}, \\ f_{X_{(\frac{m+2}{2})}}(x) & \text{if } m \text{ is even } \&\, i = \frac{m+2}{2}, ..., m. \end{cases} \tag{2.5}$$

Let $W_1^{(2)}, W_2^{(2)}, ..., W_m^{(2)}$ be a DMRSS; that is

$$W_i^{(2)} = \begin{cases} W^{(1)}_{(\frac{m+1}{2})} & \text{if } m \text{ is odd } \&\, i = 1, ..., m, \\ W^{(1)}_{(\frac{m}{2})} & \text{if } m \text{ is even } \&\, i = 1, ..., \frac{m}{2}, \\ W^{(1)}_{(\frac{m+2}{2})} & \text{if } m \text{ is even } \&\, i = \frac{m+2}{2}, ..., m. \end{cases}$$

The pdf of $W_i^{(2)}$ is

$$f_{W_i^{(2)}}(x) = \begin{cases} f_{W^{(1)}_{(\frac{m+1}{2})}}(x) & \text{if } m \text{ is odd } \&\, i = 1, ..., m, \\ f_{W^{(1)}_{(\frac{m}{2})}}(x) & \text{if } m \text{ is even } \&\, i = 1, ..., \frac{m}{2}, \\ f_{W^{(1)}_{(\frac{m+2}{2})}}(x) & \text{if } m \text{ is even } \&\, i = \frac{m+2}{2}, ..., m. \end{cases}$$

Referring to the procedures of MRSS and DMRSS, it is worth observing that both $W_i^{(1)}$ and $W_i^{(2)}$ are independent over i.

3. MIXED DOUBLE-RANKED SET SAMPLNG

MxDRSS scheme is similar to DRSS but in stage 2 MRSS is applied in lieu of RSS. The following steps describe the procedure of MxDRSS:

1. Choose m sets randomly of size m^2 elements each from the study population.
2. Apply the procedure of RSS on each set of Step 1 to acquire a RSS of size m. This produces m ranked sets (each of size m).
3. Apply the procedure of MRSS on each ranked set in Step 2 to acquire a second stage sample, which we call it a MxDRSS of size m.
4. Repeat Steps 1–3 independently h cycles, if needed, to acquire an MxDRSS of size $n = mh$.

In order to clarify this procedure, it is helpful to refer to some illustrations. First let us denote X_{ijk}, $i, j, k = 1, 2, ..., m$ for the units obtained by Step 1, where i is for the number of sets and $j \times k$ is the size of the i^{th} set. X_{ijk} are iid with common distribution function $F(x)$ and density $f(x)$. Second, let $Y_{ij} = X_{(ijj)}$, $i, j = 1, 2, ..., m$ be the units obtained by Step 2 (Y_{ij} denote the j^{th} order statistic from the i^{th} set). Finally, the units obtained in Step 3 are denoted by Z_i, $i = 1, 2, ..., m$. Tables 1 and 2 explain the procedure when $m = 3$ and 4, respectively.

Table 1: Mixed double-ranked set sampling: $m = 3$.

Step 1	Step 2	Step 3
X_{111}, X_{112}, X_{113} X_{121}, X_{122}, X_{123} X_{131}, X_{132}, X_{133}	$Y_{11} = X_{(111)}$ $Y_{12} = X_{(122)}$ $Y_{13} = X_{(133)}$	$Z_1 = Y_{(12)}$
X_{211}, X_{212}, X_{213} X_{221}, X_{222}, X_{223} X_{231}, X_{232}, X_{233}	$Y_{21} = X_{(211)}$ $Y_{22} = X_{(222)}$ $Y_{23} = X_{(233)}$	$Z_2 = Y_{(22)}$
X_{311}, X_{312}, X_{313} X_{321}, X_{322}, X_{323} X_{331}, X_{332}, X_{333}	$Y_{31} = X_{(311)}$ $Y_{32} = X_{(322)}$ $Y_{33} = X_{(333)}$	$Z_3 = Y_{(32)}$

Table 2: Mixed double-ranked set sampling: $m = 4$.

Step 1	Step 2	Step 3
X_{111}, X_{112}, X_{113}, X_{114} X_{121}, X_{122}, X_{123}, X_{124} X_{131}, X_{132}, X_{133}, X_{134} X_{141}, X_{142}, X_{143}, X_{144}	$Y_{11} = X_{(111)}$ $Y_{12} = X_{(122)}$ $Y_{13} = X_{(133)}$ $Y_{14} = X_{(144)}$	$Z_1 = Y_{(12)}$
X_{211}, X_{212}, X_{213}, X_{214} X_{221}, X_{222}, X_{223}, X_{224} X_{231}, X_{232}, X_{233}, X_{234} X_{241}, X_{242}, X_{243}, X_{244}	$Y_{21} = X_{(211)}$ $Y_{22} = X_{(222)}$ $Y_{23} = X_{(233)}$ $Y_{24} = X_{(244)}$	$Z_2 = Y_{(22)}$
X_{311}, X_{312}, X_{313}, X_{314} X_{321}, X_{322}, X_{323}, X_{324} X_{331}, X_{332}, X_{333}, X_{334} X_{341}, X_{342}, X_{343}, X_{344}	$Y_{31} = X_{(311)}$ $Y_{32} = X_{(322)}$ $Y_{33} = X_{(333)}$ $Y_{34} = X_{(344)}$	$Z_3 = Y_{(33)}$
X_{411}, X_{412}, X_{413}, X_{414} X_{421}, X_{422}, X_{423}, X_{424} X_{431}, X_{432}, X_{433}, X_{434} X_{441}, X_{442}, X_{443}, X_{444}	$Y_{41} = X_{(411)}$ $Y_{42} = X_{(422)}$ $Y_{43} = X_{(433)}$ $Y_{44} = X_{(444)}$	$Z_4 = Y_{(43)}$

4. PRACTICALITY OF MxDRSS

In this section, Hellinger distance is defined and used as a *measure of added practicality* and applied to some variations of RSS.

Suppose Y and X are two random variables with density functions $f_Y(x)$ and $f_X(x)$, respectively. The Hellinger distance (see for example [16]) between Y and X is defined by

$$H(X,Y) = \left(1 - \int_{-\infty}^{\infty} \sqrt{f_Y(x) f_X(x)} dx\right)^{\frac{1}{2}} .$$

Obviously, for iid random variables, $H(X,Y) = 0$. So the Hellinger distance between any two data points of the SRS $X_1, X_2, ..., X_m$ is zero. Therefore, identifying the ordered data points (for getting either RSS or MRSS) based on the SRS is difficult. That is, obtaining MRSS and RSS are equivalent in terms of practicality.

Now, given the data points of the RSS $(Y_1^{(1)}, Y_2^{(1)}, ..., Y_m^{(1)})$, and using the pdf's of the order statistics, it can be shown after simple calculation that the Hellinger distances between any pair of RSS data points are given in the third column of Table 3. Note that the Hellinger distances in this case are not zeros; that is, the additional work of identifying the ordered data points of DRSS (i.e., for stage 2) based on the RSS data points (stage 1) is simpler now than using SRS data points.

Table 3: Hellinger distances, $m = 2, 3, 4$; 1st and 2nd stage.

m	(k,l)	stage 1	stage 2
2	$(1,2)$	0.4633	0.5920
3	$(1,2)$	0.4086	0.5473
	$(1,3)$	0.7071	0.8625
	$(2,3)$	0.4086	0.5473
4	$(1,2)$	0.3870	0.5306
	$(1,3)$	0.6501	0.8304
	$(1,4)$	0.8399	0.9628
	$(2,3)$	0.3412	0.4889
	$(2,4)$	0.6501	0.8304
	$(3,4)$	0.3870	0.5306

Now, given the data points of the MRSS $(W_1^{(1)}, W_2^{(1)}, ..., W_m^{(1)})$, and suppose m is odd. Due to the iid case, $H\left(W_k^{(1)}, W_l^{(1)}\right) = 0$ for each $k, l = 1, 2, ..., m$. Therefore, getting a DMRSS based on the MRSS practically is the same as obtaining a MRSS based on the SRS. When m is even, the Hellinger distance is given by

$$H\left(W_k^{(1)}, W_l^{(1)}\right) = \begin{cases} H(W_{\frac{m}{2}}^{(1)}, W_{\frac{m+2}{2}}^{(1)}) > 0 & \text{if } k \le \frac{m}{2} \,\&\, l > \frac{m}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Now suppose $Y_1^{(2)}, Y_2^{(2)}, ..., Y_m^{(2)}$ be a DRSS, then the Hellinger distance between any pairs of DRSS data points are shown in the last column of Table 3. It is clear that Hellinger distances are higher in stage 2 than in stage 1.

Similarly, for the DMRSS $W_1^{(2)}, W_2^{(2)}, ..., W_m^{(2)}$, the Hellinger distance is zero when m is odd. When m is even, the Hellinger distance is given by

$$H\left(W_k^{(2)}, W_l^{(2)}\right) = \begin{cases} H(W_{\frac{m}{2}}^{(2)}, W_{\frac{m+2}{2}}^{(2)}) > H(W_{\frac{m}{2}}^{(1)}, W_{\frac{m+2}{2}}^{(1)}) > 0 & \text{if } k \le \frac{m}{2} \,\&\, l > \frac{m}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

To sum up, for a single stage sampling scheme, MRSS and RSS have same practicality, and since it is shown in the literature that MRSS is more efficient than RSS, we recommend to use MRSS. For a double stage sampling scheme, DRSS is more practical than DMRSS. But, it is shown in the literature DMRSS is more efficient. So, to gain the efficiency provided by applying MRSS, we suggest to mix MRSS with RSS by applying the procedure of MRSS on the data points of RSS. That is, in the first stage we apply RSS and in the second stage we apply MRSS. So, the obtained sample is just a combination between RSS and MRSS and it is a double stage approach, and we call it MxDRSS. The practicality of this new MxDRSS scheme is same as DRSS but in Section 6 we show it is more efficient.

Due to the properties of order statistics $V_1, ..., V_m$, it can be seen that $H(V_1, V_m)$ is the largest distance and $H(V_{\frac{m}{2}}, V_{\frac{m+2}{2}})$ is the minimum distance. Also note that $H(V_1, V_{1+r}) = H(V_{m-r}, V_m)$, $r = 2, ..., m-1$. Apparently increasing m decreases the Hellinger distances for the same pair of order statistics; which is reasonable in the sense that identifying the ordered data points from a small m is easier than in a large m. It can also be concluded from Table 3 that identifying the ordered data points for stage 2 (DRSS) based on the ordered data points of stage 1 (RSS) is consistently easier than identifying the ordered data points for stage 1 (RSS) based on the identical data points of SRS. This result is consistent with the findings of [1].

5. ESTIMATION OF THE POPULATION MEAN

In this section estimation of the population mean is studied. Particularly, in Section 5.1 the population mean estimation is reviewed under the SRS, RSS, and DRSS schemes. In Sections 5.2 and 5.3 the population mean estimation is reviewed respectively under the MRSS and DMRSS schemes and also the results given in the literature about these schemes are enhanced and some new closed form expressions for the variances of the sample means and efficiencies are provided. Finally, in Section 5.4 the population mean estimation is investigated under the proposed MxDRSS scheme.

5.1. Population mean estimation based on SRS, RSS, and DRSS

Let $X_1, X_2, ..., X_m$ be a SRS from $f(x)$. The mean of the sample $\bar{X} = \sum_{i=1}^m X_i/m$ is an unbiased estimator of μ with variance σ^2/m.

Let $Y_1^{(1)}, Y_2^{(1)}, ..., Y_m^{(1)}$ be a RSS. It is shown by [25] (see also [26]) that $\bar{Y}^{(1)} = \sum_{i=1}^m Y_i^{(1)}/m$ is an unbiased estimator of μ and $\text{Var}(\bar{Y}^{(1)}) \le \text{Var}(\bar{X})$. [7] reported that $\text{Var}(\bar{Y}^{(1)}) = \sigma^2/m - \sum_{i=1}^m \left(\mu_i^{(1)} - \mu\right)^2/m^2$, where $\mu_i^{(1)}$ is the i^{th} order statistic's mean.

Let $Y_1^{(2)}, Y_2^{(2)}, ..., Y_m^{(2)}$ be a DRSS. [1] reported that the mean $\bar{Y}^{(2)} = \sum_{i=1}^m Y_i^{(2)}/m$ is an unbiased estimator of μ with $\mathrm{Var}(\bar{Y}^{(2)}) = \sigma^2/m - \sum_{i=1}^m \left(\mu_i^{(2)} - \mu\right)^2/m^2$, where $\mu_i^{(2)}$ is the i^{th} order statistic's mean of the RSS $Y_1^{(1)}, Y_2^{(1)}, ..., Y_m^{(1)}$. They also showed that $\mathrm{Var}(\bar{Y}^{(2)}) \leq \mathrm{Var}(\bar{Y}^{(1)})$.

5.2. Population mean estimation based on MRSS

Let $W_1^{(1)}, W_2^{(1)}, ..., W_m^{(1)}$ be a MRSS. Let $\bar{W}^{(1)} = \frac{1}{m}\sum_{i=1}^m W_i^{(1)}$ be the sample mean of MRSS. Then

$$E\left(\bar{W}^{(1)}\right) = \begin{cases} \mu^{(1)}_{\frac{m+1}{2}} & \text{if } m \text{ is odd,} \\ \frac{1}{2}\left(\mu^{(1)}_{\frac{m}{2}} + \mu^{(1)}_{\frac{m+2}{2}}\right) & \text{if } m \text{ is even,} \end{cases}$$

where $\mu_k^{(1)} = E\left(X_{(k)}\right)$. [15] reported that, for symmetric distribution, $\bar{W}^{(1)}$ is an unbiased estimator of μ.

The variance of $\bar{W}^{(1)}$ can be derived as follows:

$$\mathrm{Var}\left(\bar{W}^{(1)}\right) = \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^m W_i^{(1)}\right).$$

Since the data points of MRSS are independent, then

$$\mathrm{Var}\left(\bar{W}^{(1)}\right) = \frac{1}{m^2}\sum_{i=1}^m \mathrm{Var}\left(W_i^{(1)}\right).$$

Now, from Eq (2.4) and Eq (2.5), we have

$$\mathrm{Var}\left(\bar{W}^{(1)}\right) = \begin{cases} \frac{1}{m}\sigma^{2(1)}_{\frac{m+1}{2}} & \text{if } m \text{ is odd,} \\ \frac{1}{2m}\left(\sigma^{2(1)}_{\frac{m}{2}} + \sigma^{2(1)}_{\frac{m+2}{2}}\right) & \text{if } m \text{ is even,} \end{cases}$$

where $\sigma_k^{2(1)} = \mathrm{Var}\left(X_{(k)}\right)$. Using the result of [7],

$$\mathrm{Var}\left(\bar{W}^{(1)}\right) = \begin{cases} \sigma^2 - \frac{1}{m}\sum_{i=1}^m \left(\mu_i^{(1)} - \mu\right)^2 - \frac{1}{m}\sum_{i:i\neq\frac{m+1}{2}}^m \sigma_i^{2(1)} & \text{if } m \text{ is odd,} \\ \frac{1}{2}\sigma^2 - \frac{1}{2m}\sum_{i=1}^m \left(\mu_i^{(1)} - \mu\right)^2 - \frac{1}{2m}\sum_{i:i\neq\frac{m}{2},\frac{m+2}{2}}^m \sigma_i^{2(1)} & \text{if } m \text{ is even.} \end{cases}$$

5.3. Population mean estimation based on DMRSS

Let $W_1^{(2)}, W_2^{(2)}, ..., W_m^{(2)}$ be a DMRSS. Let $\bar{W}^{(2)} = \frac{1}{m}\sum_{i=1}^{m} W_i^{(2)}$ be the sample mean of DMRSS. Then

$$E\left(\bar{W}^{(2)}\right) = \begin{cases} \mu^{(2)}_{\frac{m+1}{2}} & \text{if } m \text{ is odd,} \\ \dfrac{1}{2}\left(\mu^{(2)}_{\frac{m}{2}} + \mu^{(2)}_{\frac{m+2}{2}}\right) & \text{if } m \text{ is even,} \end{cases}$$

where $\mu_k^{(2)} = E\left(W_{(k)}^{(1)}\right)$. Using the properties of order statistics and for symmetric distribution it can be shown that $E\left(\bar{W}^{(2)}\right) = \mu$ and the variance of $\bar{W}^{(2)}$ is

$$\text{Var}\left(\bar{W}^{(2)}\right) = \begin{cases} \dfrac{1}{m}\sigma^{2(2)}_{\frac{m+1}{2}} & \text{if } m \text{ is odd,} \\ \dfrac{1}{2m}\left(\sigma^{2(2)}_{\frac{m}{2}} + \sigma^{2(2)}_{\frac{m+2}{2}}\right) & \text{if } m \text{ is even,} \end{cases}$$

where $\sigma_k^{2(2)} = \text{Var}\left(W_{(k)}^{(1)}\right)$. Using the result of [1],

$$\text{Var}\left(\bar{W}^{(2)}\right) = \begin{cases} \sigma^2 - \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\left(\mu_i^{(2)} - \mu\right)^2 - \dfrac{1}{m}\sum_{i:i\neq\frac{m+1}{2}}^{m} \sigma_i^{2(2)} & \text{if } m \text{ is odd,} \\ \dfrac{1}{2}\sigma^2 - \dfrac{1}{2m}\displaystyle\sum_{i=1}^{m}\left(\mu_i^{(2)} - \mu\right)^2 - \dfrac{1}{2m}\sum_{i:i\neq\frac{m}{2},\frac{m+2}{2}}^{m} \sigma_i^{2(2)} & \text{if } m \text{ is even.} \end{cases}$$

5.4. Population mean estimation based on MxDRSS

Let $Z_1, Z_2, ..., Z_m$ be a MxDRSS; that is

$$Z_i = \begin{cases} Y^{(1)}_{(\frac{m+1}{2})} & \text{if } m \text{ is odd } \&\, i = 1, ..., m, \\ Y^{(1)}_{(\frac{m}{2})} & \text{if } m \text{ is even } \&\, i = 1, ..., \frac{m}{2}, \\ Y^{(1)}_{(\frac{m+2}{2})} & \text{if } m \text{ is even } \&\, i = \frac{m+2}{2}, ..., m. \end{cases}$$

Referring to the procedure of MxDRSS, one may conclude that Z_i are independent over i, and it is worth observing that they are not identical. The pdf of Z_i is

$$f_{Z_i}(x) = \begin{cases} f_{Y^{(1)}_{(\frac{m+1}{2})}}(x) & \text{if } m \text{ is odd } \&\, i = 1, ..., m, \\ f_{Y^{(1)}_{(\frac{m}{2})}}(x) & \text{if } m \text{ is even } \&\, i = 1, ..., \frac{m}{2}, \\ f_{Y^{(1)}_{(\frac{m+2}{2})}}(x) & \text{if } m \text{ is even } \&\, i = \frac{m+2}{2}, ..., m. \end{cases}$$

Let $\bar{Z} = \frac{1}{m}\sum_{i=1}^{m} Z_i$ be the sample mean of MxDRSS. Then

$$E\left(\bar{Z}\right) = \begin{cases} \mu_{Y^{(1)}_{\left(\frac{m+1}{2}\right)}} & \text{if } m \text{ is odd,} \\ \dfrac{1}{2}\left(\mu_{Y^{(1)}_{\left(\frac{m}{2}\right)}} + \mu_{Y^{(1)}_{\left(\frac{m+2}{2}\right)}}\right) & \text{if } m \text{ is even,} \end{cases}$$

where $\mu_{Y^{(1)}_{(k)}} = E\left(Y^{(1)}_{(k)}\right)$. Using the properties of order statistics and for symmetric distribution it can be shown that $E\left(\bar{Z}\right) = \mu$ and the variance of $\bar{Z}$ is

$$\operatorname{Var}\left(\bar{Z}\right) = \begin{cases} \dfrac{1}{m}\sigma^2_{Y^{(1)}_{\left(\frac{m+1}{2}\right)}} & \text{if } m \text{ is odd,} \\ \dfrac{1}{2m}\left(\sigma^2_{Y^{(1)}_{\left(\frac{m}{2}\right)}} + \sigma^2_{Y^{(1)}_{\left(\frac{m+2}{2}\right)}}\right) & \text{if } m \text{ is even,} \end{cases}$$

where $\sigma^2_{Y^{(1)}_{(k)}} = \operatorname{Var}\left(Y^{(1)}_{(k)}\right)$.

6. NUMERICAL RESULTS FOR SPECIFIC DISTRIBUTIONS

6.1. Results from a uniform distribution

Suppose that the underlying population is uniform U$(0,1)$, then the sample means using SRS, RSS, MRSS, DRSS, DMRSS and MxDRSS of size m are unbiased estimators of μ, while the variances depend on the sampling scheme.

1. For a SRS, $\operatorname{Var}(\bar{X}) = 1/12m$.

2. For a RSS, $\operatorname{Var}(\bar{Y}^{(1)}) = 1/6m(m+1)$, and the relative efficiency (see [25]) w.r.t. SRS is $\operatorname{Eff}(\bar{Y}^{(1)}; \bar{X}) = \operatorname{Var}(\bar{X}) / \operatorname{Var}(\bar{Y}^{(1)}) = (m+1)/2$.

3. For a MRSS, the variance of the sample mean and the relative efficiency have not been provided in the literature in closed form. However, we find that the following expressions can be obtained for this situation:

$$\operatorname{Var}\left(\bar{W}^{(1)}\right) = \begin{cases} \dfrac{1}{4m(m+2)} & \text{if } m \text{ is odd,} \\ \dfrac{1}{4(m+1)^2} & \text{if } m \text{ is even.} \end{cases}$$

   Thus, the relative efficiency w.r.t. SRS is given by

$$\operatorname{Eff}(\bar{W}^{(1)}; \bar{X}) = \frac{\operatorname{Var}(\bar{X})}{\operatorname{Var}(\bar{W}^{(1)})} = \begin{cases} \dfrac{m+2}{3} & \text{if } m \text{ is odd,} \\ \dfrac{(m+1)^2}{3m} & \text{if } m \text{ is even.} \end{cases}$$

4. For a DRSS, when $m = 3$, $\mathrm{Var}(\bar{Y}^{(2)}) \approx 0.0092$, and the relative efficiency is $\mathrm{Eff}(\bar{Y}^{(2)}; \bar{X}) = 3.026$. When $m = 4$, $\mathrm{Var}(\bar{Y}^{(2)}) \approx 0.0049$, and the relative efficiency is $\mathrm{Eff}(\bar{Y}^{(2)}; \bar{X}) = 4.281$.

5. For a DMRSS, when $m = 3$, $\mathrm{Var}(\bar{W}^{(2)}) = \sigma_2^{2(2)}/3 \approx 0.0089$, and the relative efficiency is $\mathrm{Eff}(\bar{W}^{(2)}; \bar{X}) = 3.130$. For $m = 4$, $\mathrm{Var}(\bar{W}^{(2)}) = (\sigma_2^{2(2)} + \sigma_3^{2(2)})/8 \approx 0.0047$, and the relative efficiency is $\mathrm{Eff}(\bar{W}^{(2)}; \bar{X}) = 4.422$.

6. For a MxDRSS, when $m = 3$, $\mathrm{Var}(\bar{Z}) = \sigma^2_{Y^{(1)}_{(2)}} \approx 0.0115$, and the relative efficiency is $\mathrm{Eff}(\bar{Z}; \bar{X}) = 2.406$. When $m = 4$, $\mathrm{Var}(\bar{Z}) = (\sigma^2_{Y^{(1)}_{(2)}} + \sigma^2_{Y^{(1)}_{(3)}})/2 \approx 0.0060$, and the relative efficiency is $\mathrm{Eff}(\bar{Z}; \bar{X}) = 3.470$.

So far, we have discussed results for symmetric but rectangular distribution. In the next subsection, we will discuss results for other types of well known distributions.

6.2. Results for the normal, exponential, and skew normal distributions

The relative efficiencies of the sample means obtained by RSS, MRSS, DRSS, DMRSS, and MxDRSS w.r.t. SRS for the normal distribution $\mathrm{N}(0, 1)$, skew normal distribution $\mathrm{SN}(0, 1, 1)$, and exponential distribution $\mathrm{Exp}(1)$ are summarized in Table 4. Also the results of the uniform distribution $\mathrm{U}(0, 1)$ are provided. Table 5 shows the bias and variance of the obtained estimators from the skewed distributions. Moreover, to examine the effect of the kurtosis and skewness on the biasedness and relative efficiency of the considered sampling schemes the gamma distribution $\mathrm{Gamma}(\alpha, 1)$ is used, where α is changed from 1 to 6 (note that increasing α decreases the kurtosis and the skewness) and the results are shown in Figures 1 and 2 for $m = 3$ and $m = 4$, respectively. So, from Figures 1 and 2 one may conclude that:

(**a**) bias is a bit higher for skewed distributions than non-skewed distributions;

(**b**) the efficiency is low for highly skewed distributions.

From the results of Tables 4, 5, and Figures 1 and 2 the remarks below can be observed:

1. In terms of efficiency, the best sampling scheme among those studied in this paper is the DMRSS except for highly skewed distribution like the exponential distribution.

2. As m increases, the efficiency also increases except for the $\mathrm{Exp}(1)$ under DMRSS (it decreases when $m > 2$ as shown by [20]) and MxDRSS (it decreases when $m > 3$). Our MxDRSS scheme shows better performance than DMRSS when $m > 3$.

3. The efficiency is lower for those distributions with large skewness and large kurtosis.

4. In terms of biasedness, the MRSS has the smallest bias.

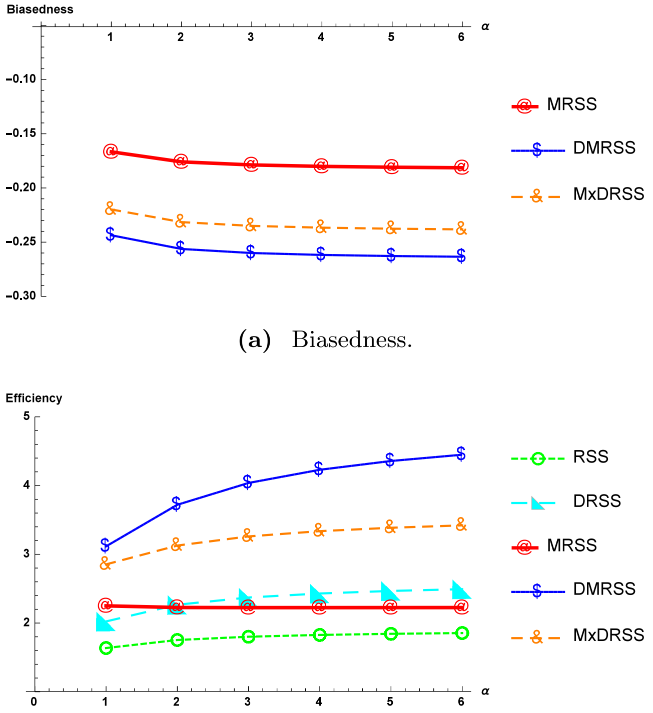
5. The bias is small when the skewness is small.

Table 4: The efficiency in the population mean estimation under the considered sampling schemes w.r.t. SRS.

Distribution	(Skewness, kurtosis)	Method	m			
			2	3	4	5
U(0, 1)	(0, −1.2)	RSS	1.500	2.000	2.500	3.000
		MRSS	1.500	1.667	2.083	2.333
		DRSS	1.923	3.026	4.281	5.670
		DMRSS	1.923	3.130	4.422	6.925
		MxDRSS	1.923	2.406	3.470	4.350
N(0, 1)	(0, 0)	RSS	1.467	1.914	2.347	2.770
		MRSS	1.467	2.229	2.774	3.486
		DRSS	1.785	2.633	3.526	4.456
		DMRSS	1.785	4.992	7.091	12.226
		MxDRSS	1.785	3.615	5.046	7.318
SN(0, 1, 1)	(0.137, 0.062)	RSS	1.465	1.909	2.339	2.759
		MRSS	1.465	2.241	2.786	3.500
		DRSS	1.780	2.620	3.503	4.419
		DMRSS	1.780	5.016	7.089	12.030
		MxDRSS	1.780	3.635	5.059	7.290
Exp(1)	(2, 6)	RSS	1.333	1.636	1.920	2.190
		MRSS	1.333	2.250	2.441	2.230
		DRSS	1.516	2.024	2.523	3.016
		DMRSS	1.516	3.116	2.867	2.226
		MxDRSS	1.516	2.854	2.988	2.265

Table 5: The (bias, variance) of the sample mean obtained by MRSS, DMRSS, and MxDRSS for skewed distributions.

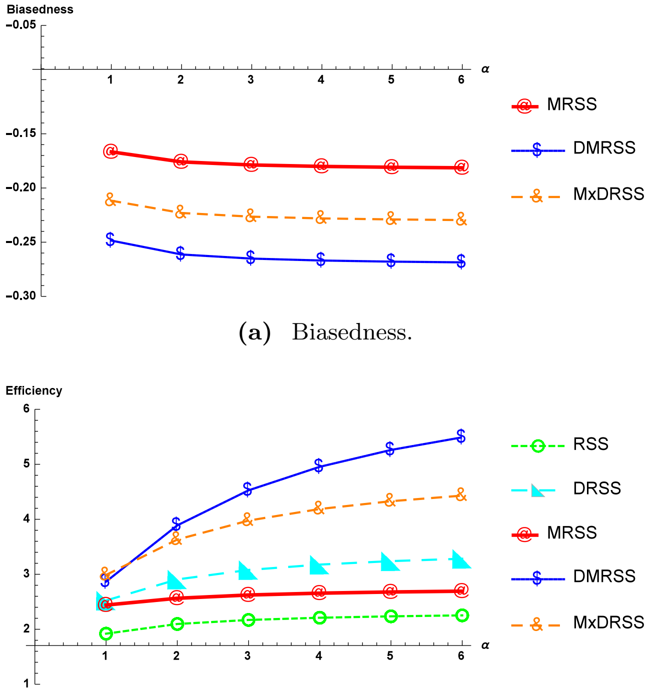
Distribution (Skewness, kurtosis)	Method	m		
		3	4	5
SN(0, 1, 1) (0.137, 0.062)	MRSS	(−0.010, 0.101)	(−0.010, 0.061)	(−0.014, 0.039)
	DMRSS	(−0.015, 0.045)	(−0.016, 0.024)	(−0.018, 0.011)
	MxDRSS	(−0.014, 0.062)	(−0.013, 0.034)	(−0.017, 0.018)
Exp(1) (2, 6)	MRSS	(−0.167, 0.120)	(−0.167, 0.075)	(−0.217, 0.043)
	DMRSS	(−0.244, 0.048)	(−0.249, 0.025)	(−0.281, 0.011)
	MxDRSS	(−0.220, 0.068)	(−0.212, 0.039)	(−0.264, 0.019)



(a) Biasedness.

(b) Efficiency.

Figure 1: The effectiveness of the skewness parameter of the Gamma$(\alpha, 1)$ on the biasedness and efficiency of the estimates when $m = 3$ under the considered sampling schemes.



(a) Biasedness.

(b) Efficiency.

Figure 2: The effectiveness of the skewness parameter of the Gamma$(\alpha, 1)$ on the biasedness and efficiency of the estimates when $m = 4$ under the considered sampling schemes.

7. A REAL DATA EXAMPLE

In this section, a real data set is analyzed to illustrate the usefulness of our proposed methodology.

The body mass index (BMI) is a measure of relative size based on the mass and height of an individual. It is commonly employed among children and adults to predict health outcomes. Commonly accepted BMI ranges are underweight: under 18.5, normal weight: 18.5 to 25, overweight: 25 to 30, obese: over 30. A data set that has a BMI for 2107 people is contained in R-package `mixsmsn`. Six types of samples (obtained by using SRS, RSS, MRSS, DRSS, DMRSS, and MxRSS) of size 5 each are presented in Table 6 and the question of interest is to estimate the mean of the BMI. The estimated BMI mean and the standard error of the mean under SRS, RSS, MRSS, DRSS, DMRSS, and MxDRSS are obtained and reported in Table 6.

Table 6: Body mass index example.

	SRS	RSS	MRSS	DRSS	DMRSS	MxDRSS
	20.00	22.00	25.91	22.00	26.36	26.36
	22.62	20.25	21.97	26.89	23.08	28.30
	23.70	26.36	31.63	22.09	28.68	22.09
	32.79	31.96	26.51	30.78	24.86	26.30
	35.18	33.46	34.63	32.64	26.51	23.32
estimated mean	26.858	26.806	28.130	26.880	25.898	25.274
estimated standard error	2.9951	2.6184	2.2361	2.1811	0.9313	1.1257

As suggested by [23] the estimated variance of the sample mean obtained by RSS is given by

$$S^2_{\text{RSS}} = \frac{\sum_{i=1}^{m}(Y_i^{(1)} - \bar{Y}^{(1)})^2}{m-1}.$$

Accordingly, one may define the estimated variances of the sample means obtained by MRSS, DRSS, DMRSS, and MxDRSS in the same way. For example, in case of MxDRSS,

$$S^2_{\text{MxDRSS}} = \frac{\sum_{i=1}^{m}(Z_i - \bar{Z})^2}{m-1},$$

and hence the estimated standard error is given by

$$\text{SE}(\bar{Z}) = \sqrt{\frac{S^2_{\text{MxDRSS}}}{m}}.$$

8. CONCLUSION

Practically, given an RSS in stage 1, applying RSS or MRSS in stage 2 is the same because identifying the sample observations is done after the ranking process. But as discussed in Section 6 it is shown that efficiency is higher if we apply MRSS in stage 2. It is also found that efficiency decreases by increases in the kurtosis and skewness. To sum up, DRSS and MxDRSS will behave the same in practicality, but in terms of efficiency MxDRSS is better than DRSS (except for the uniform distribution, which is fatter tailed).

ACKNOWLEDGMENTS

The authors wish to thank the associated editor and referees for their constructive comments which improve the final version of this paper. The authors also would like to acknowledge excellent research support from KFUPM grant number SR151009.

REFERENCES

[1] Al-Saleh, M.F. and Al-Kadiri, M.A. (2000). Double-ranked set sampling, *Statistics & Probability Letters*, **48**(2), 205–212.

[2] Al-Saleh, M.F. and Diab, Y.A. (2009). Estimation of the parameters of Downton's bivariate exponential distribution using ranked set sampling scheme, *Journal of Statistical Planning and Inference*, **139**(2), 277–286.

[3] Arnold, B.; Balakrishnan, N. and Nagaraja, H. (2008). *A First Course in Order Statistics*, Society for Industrial and Applied Mathematics.

[4] Barnett, V. and Moore, K. (1997). Best linear unbiased estimates in ranked-set sampling with particular reference to imperfect ordering, *Journal of Applied Statistics*, **24**(6), 697–710.

[5] Chen, Z.; Bai, Z. and Sinha, B. (2004). *Ranked Set Sampling: Theory and Applications*, Vol. 176, Springer Science & Business Media.

[6] David, H.A. and Nagaraja, H.N. (2003). *Order Statistics*, 3rd ed., Wiley, New York.

[7] Dell, T.R. and Clutter, J.L. (1972). Ranked set sampling theory with order statistics background, *Biometrics*, **28**(2), 545–555.

[8] Jemain, A.A.; Al-Omari, A. and Ibrahim, K. (2008). Some variations of ranked set sampling, *Electronic Journal of Applied Statistical Analysis*, **1**(1), 1–15.

[9] McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets, *Australian Journal of Agricultural Research*, **3**(4), 385–390.

[10] McIntyre, G. (2005). A method for unbiased selective sampling, using ranked sets, *The American Statistician*, **59**(3), 230–232.

[11] Mehmood, R.; Riaz, M. and Does, R.J.M.M. (2013). Control charts for location based on different sampling schemes, *Journal of Applied Statistics*, **40**(3), 483–494.

[12] Mehmood, R.; Riaz, M. and Does, R.J.M.M. (2014). Quality quandaries: On the application of different ranked set sampling schemes, *Quality Engineering*, **26**(3), 370–378.

[13] Murff, E.J.T. and Sager, T.W. (2006). The relative efficiency of ranked set sampling in ordinary least squares regression, *Environmental and Ecological Statistics*, **13**(1), 41–51.

[14] Murray, R.A.; Ridout, M.S. and Cross, J.V. (2000). The use of ranked set sampling in spray deposit assessment, *Aspects of Applied Biology*, **57**, 141–146.

[15] Muttlak, H. (1997). Median ranked set sampling, *Journal of Applied Statistical Science*, **6**, 245–255.

[16] Nikulin, M.S. (2001). Hellinger Distance, *Encyclopedia of Mathematics.*

[17] Patil, G.P. (1995). Editorial: ranked set sampling, *Environmental and Ecological Statistics*, **2**(4), 271–285.

[18] Philip, L.H. and Lam, K. (1997). Regression estimator in ranked set sampling, *Biometrics*, **53**(3), 1070–1080.

[19] Riaz, M.; Mahmood, T.; Abbasi, S.A. and Abbas, N. (2017). Linear profile monitoring using EWMA structure under ranked set schemes, *The International Journal of Advanced Manufacturing Technology*, **91**(3), 2751–2775.

[20] Samawi, H.M. and Tawalbeh, E.M. (2002). Double median ranked set sample: comparing to other double ranked samples for mean and ratio estimators, *Journal of Modern Applied Statistical Methods*, **1**(2), 428–442.

[21] Samuh, M.H. and Al-Saleh, M.F. (2011). The effectiveness of multistage ranked set sampling in stratifying the population, *Communications in Statistics — Theory and Methods*, **40**(6), 1063–1080.

[22] Sarikavanij, S.; Kasala, S.; Sinha, B.K. and Tiensuwan, M. (2014). Estimation of location and scale parameters in two-parameter exponential distribution based on ranked set sample, *Communications in Statistics – Simulation and Computation*, **43**(1), 132–141.

[23] Stokes, S.L. (1980). Estimation of variance using judgment order ranked set samples, *Biometrics*, **36**, 35–42.

[24] Stokes, S.L. and Sager, T.W. (1988). Characterization of a ranked-set sample with application to estimating distribution functions, *Journal of the American Statistical Association*, **83**(402), 374–381.

[25] Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering, *Annals of the Institute of Statistical Mathematics*, **20**(1), 1–31.

[26] Wolfe, D.A. (2010). Ranked set sampling, *Wiley Interdiscip. Rev. Comput. Stat.*, **2**, 460–466.

REVSTAT – STATISTICAL JOURNAL

Background

Statistics Portugal (INE, I.P.), well aware of how vital a statistical culture is in understanding most phenomena in the present-day world, and of its responsibility in disseminating statistical knowledge, started the publication of a scientific statistical journal called Revista de Estatística. The original language used in this publication was Portuguese and the idea behind it was to publish it, three times a year, containing original research results, and application studies, namely in the economic, social and demographic fields.

In 1998 it was decided that the publication should also include papers in English. This step was taken to achieve a broader dissemination, and to encourage foreign contributors to submit their work for publication.

At the time, the Editorial Board was mainly comprised of Portuguese university professors. It is now comprised of international university faculties and this has been the first step aimed at changing the character of Revista de Estatística from a national to an international scientific journal.

We have also initiated a policy of publishing special volumes that may be thematic highlighting areas of interest or associated with scientific events in Statistics. For example, in 2001, a special issue of Revista de Estatística was published containing three volumes of extended abstracts of the invited contributed papers presented at the 23rd European Meeting of Statisticians.

In 2003, the name of the Journal has been changed to REVSTAT - STATISTICAL JOURNAL, now fully published in English, with a prestigious international editorial board, aiming to become a reference scientific journal that promotes the dissemination of relevant research results in Statistics.

The editorial policy of REVSTAT Statistical Journal is mainly placed on the originality and importance of the research.

All articles consistent with REVSTAT aims and scope will undergo scientific evaluation by at least two reviewers, one from the Editorial Board and another external.

The only working language allowed is English.

Abstract and Indexing Services

The REVSTAT is covered by the following abstracting/indexing services:

- Current Index to Statistics
- Google Scholar
- Mathematical Reviews® (MathSciNet®)
- Science Citation Index Expanded
- Zentralblatt für Mathematic
- Scimago Journal & Country Rank
- Scopus

Instructions to Authors

Articles must be written in English and will be submitted according to the following guidelines:

The corresponding author sends the manuscript in PDF format to the Executive Editor (revstat@ine.pt) with the Subject "New Submission to REVSTAT"; a MS#REVSTAT reference will be assigned later.

Optionally, in a mail cover letter, authors are welcome to suggest one of the Editors or Associate Editors, whose opinion may be considered suitable to be taken into account.

The submitted manuscript should be original and not have been previously published nor about to be published elsewhere in any form or language, avoiding concerns about self-plagiarism'.

Content published in this journal is peer-reviewed (Single Blind).

All research articles will be refereed by at least two researchers, including one from the Editorial Board unless the submitted manuscript is judged unsuitable for REVSTAT or does not contain substantial methodological novelty, in which case is desk rejected.

Manuscripts should be typed only in black, in double-spacing, with a left margin of at least 3 cm, with numbered lines, and with less than 25 pages. Figures (minimum of 300dpi) will be reproduced online in colours, if produced this way; however, authors should take into account that the printed version is always in black and grey tones.

The first page should include the name, ORCID iD (optional), Institution, country, and mail-address of the author(s) and a summary of fewer than one hundred words, followed by a maximum of six keywords and the AMS 2000 subject classification.

Authors are encouraged to submit articles using LaTeX, in the REVSTAT style, which is available at the LaTeX2e MACROS webpage.

References about the format and other useful information on the submission are available in the LaTeX2e Templates page.

Acknowledgments of people, grants or funds should be placed in a short section before the References title page. Note that religious beliefs, ethnic background, citizenship and political orientations of the author(s) are not allowed in the text.

Supplementary files (in REVSTAT style) may be published online along with an article, containing data, programming code, extra figures, or extra proofs, etc; however, REVSTAT is not responsible for any supporting information supplied by the author(s).

Any contact with REVSTAT must always contain the assigned REVSTAT reference number.

Accepted papers

Authors of accepted papers are requested to provide the LaTex files to the Secretary of the REVSTAT revstat@ine.pt. The authors should also mention if figure files were included, and submit electronic figures separately in .gif, .jpg, .png or .pdf format. Figures must be a minimum of 300dpi.

Copyright and reprints

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, Statistics Portugal, in order to ensure the widest possible dissemination of information, namely through the Statistics Portugal website (http://www.ine.pt).

After assigning copyright, authors may use their own material in other publications provided that REVSTAT is acknowledged as the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.

Editorial Board