SHIVAJI UNIVERSITY KOLHAPUR (M.S.)

CRITERIA-III

EVIDENCE(S) AS PER SOP

METRIC NO. 3.4.7	Bibliometric of the publications during the last five years based on average citation index in Scopus/ Web of Science or Pub Med/ Indian Citation Index

Note: The evidences for the said matrix are partly uploaded on NAAC portal and partly on university web portal. The web link is provided in the said matrix.

On the Interval Estimation of Stress-Strength Reliability for Exponentiated Scale Family of Distributions: ON THE INTERVAL ESTIMATION OF STRESS-STRENGTH RELIABILITY

Article *in* Quality and Reliability Engineering · February 2017

DOI: 10.1002/qre.2117

CITATION
1

READS
42

2 authors:

Kiran Patil
Shivaji University, Kolhapur
9 PUBLICATIONS **53** CITATIONS
SEE PROFILE

Hemangi V. Kulkarni
Shivaji University, Kolhapur
27 PUBLICATIONS **66** CITATIONS
SEE PROFILE

On the Interval Estimation of Stress-Strength Reliability for Exponentiated Scale Family of Distributions

K. P. Patil[a] and H. V. Kulkarni*[a]

[a]Department of Statistics, Shivaji University, Kolhapur.

*kulkarni.hemangi@gmail.com

Abstract

The Stress-Strength reliability $R = P(X_1 < X_2)$ where X_1 and X_2 represent the stress applied and strength of an equipment respectively, plays a crucial role in setting warranty periods while launching new brands of a product. The paper addresses the issue of estimating R when X_1 and X_2 belong to the exponentiated scale family which includes the popular exponentiated exponential distribution that has proven to be an excellent model for life time distributions. The cases of known/unknown and equal/unequal scale parameters are handled separately. For known scale parameter, a generalized pivot quantity (GPQ) for the shape parameter and R are developed. The interval estimates of R based on the proposed GPQ exhibited uniformly best performance. For an unknown scale parameter a maximum scale invariant likelihood estimator (MSILE) of the shape and an allied estimator of the scale are introduced. The parametric bootstrap interval estimates of R based on a proposed MSILE of the shape parameter exhibited best performance among others. An application in setting warranty periods is illustrated based on two real data sets.

Keyword: Exponentiated Exponential Distribution; Generalized Pivot Quantity; Maximal Scale Invariant Likelihood Estimator; Warranty Period.

1 Introduction

The Stress-Strength reliability of an equipment defined by R=P(X_1<X_2) quantifies the probability that the strength X_2 is larger than the stress X_1. This probability can be used to assess if the stress exceeds strength, when there is a high chance of instant failure and vice versa, and has elegant applications in the field of setting warranty periods for products to be launched in the market,

customer usage data, reliability engineering among other applications.

A thorough review on various inferential procedures for stress-strength reliability analysis with illustrative applications can be found in Kotz et. al.[7]. In the recent years, numerous articles have addressed the problem of inference related to R, see for example Zhou[15], Raqab et. al. [11], Surles and Padgett [13], Ahmad et. al. [1], Kundu and Gupta [8] among others. Under independence of X_1 and X_2,

$$R = P(X_1 < X_2) = \int_y G_1(y) g_2(y) dy, \tag{1}$$

where $G_i(.)$ and $g_i(.)$ are the cumulative distribution function (CDF) and probability distribution function (PDF) of $X_i, i = 1, 2$. The inferential procedures addressed in the current literature for R include maximum likelihood inference, some asymptotic methods, Bayesian methods among others which are constrained by stringent assumptions and complexity in estimation. Most often, existence of one or more nuisance parameters disturbs the quality of the underlying non Bayesian inference.

Recently the inference based on generalized pivotal quantity (GPQ) introduced by Tsui and Weerahandi [14] has received a wide attention in almost every discipline. GPQs have been observed to handle the nuisance parameters efficiently and yield accurate simple inference procedures even under small to moderate sample sizes in almost all cases were they have been used. Asymptotic properties of the CI based on the GPQs have been discussed by Hanning et. al.[6] and Roy and Bose [12]. The present article exploits this technique for interval estimation of R, when X_1 and X_2 are independently distributed members of the exponentiated scale family, also known as resilience or frailty parameter family (Marshall and Olkin[9]):

$$G(\frac{x}{\lambda}, \alpha) = F^{\alpha}(\frac{x}{\lambda}) \quad \text{Resilience family} \quad or \tag{2}$$

$$\bar{G}(\frac{x}{\lambda}, \alpha) = \bar{F}^{\alpha}(\frac{x}{\lambda}), \quad \text{Frailty family} \quad x \, \epsilon \, \mathbb{R}, \, \lambda, \, \alpha > 0, \tag{3}$$

where, λ and α, are the scale and resilience (frailty) parameters respectively, and F is a ~~given~~ known distribution function.

The exponentiated scale family encompasses many popular distributions, see for example Nadarajah and Kotz [10]. Our main emphasis is on the widely applicable and recently most popular exponentiated exponential distribution (EED) developed by Gupta and Kundu [2], (see for example Gupta and Kundu [3–5]) obtained by introducing a resilience parameter in the exponential distribution :

$$\begin{aligned} G(\frac{x}{\lambda}, \alpha) &= \left(1 - e^{-\frac{x}{\lambda}}\right)^{\alpha} \\ & x > 0, \alpha > 0, \lambda > 0. \end{aligned}$$

In the sequel, section 2 outlines a unified procedure for obtaining GPQs for a resilience (frailty) parameter and the allied interval estimation for the stress-strength reliability, when the scale parameter is known. The proposed interval estimation of stress-strength reliability is based on these GPQs. It is notable that the form of the GPQ for the resilience/frailty parameter remains same within the entire scale family $(F(\frac{x}{\lambda}),\ \lambda > 0)$ used in equation (2) and (3). For the case of an unknown scale parameter the performance of the approximate interval estimates obtained by replacing unknown scale parameters by their GPQs was not found up to the mark. In this case, a better procedure is given in section 3 based on a proposed maximal scale invariant likelihood estimator (MSILE) of α. Section 4 reports the findings of an empirical assessment of the procedures proposed in section 2 and section 3. The procedures are illustrated with the real-life data in section 5 in the context of setting warranty periods.

2 Confidence Interval for R Under Known Scale Parameter

2.1 A GPQ for the resilience (frailty) parameter

Let, $\boldsymbol{X}$ be a random variable with CDF $F_{\boldsymbol{\zeta}}(.)$, where $\boldsymbol{\zeta}$=($\boldsymbol{\theta}$, $\boldsymbol{\delta}$) is the unknown parameter vector. The interest lies in the parameter $\boldsymbol{\theta}$ while $\boldsymbol{\delta}$ is the nuisance parameter. A GPQ for $\boldsymbol{\theta}$ is defined below:

Definition 1: Generalized Pivot Quantity

A random quantity $\mathcal{G}_{\boldsymbol{\theta}} = \psi(\boldsymbol{X};\ \boldsymbol{x},\ \boldsymbol{\zeta})$ is said to be a generalized pivotal quantity for the parameter of interest $\boldsymbol{\theta}$ if it satisfies following two properties:

1. The probability distribution of $\mathcal{G}_{\boldsymbol{\theta}}$ is free from any unknown parameters.
2. The value of $\mathcal{G}_{\boldsymbol{\theta}}$=$\psi(\boldsymbol{X};\ \boldsymbol{x},\ \boldsymbol{\zeta})$ at $\boldsymbol{X} = \boldsymbol{x}$ does not depend on the nuisance parameter $\boldsymbol{\delta}$. For most of the cases $\mathcal{G}_{\boldsymbol{\theta}}$=$\boldsymbol{\theta}$.

Let, $\mathbf{X} = (X_1,\ X_2,\ ...,\ X_n)$ be n independent observations on a random variable X from the distribution function defined in equation (2) or (3). In the following theorem we develop the main result used for constructing the GPQs for a resilience(frailty) parameter assuming that the scale parameter is known.

Theorem 1: Let, $\hat{\alpha}$ be the maximum likelihood estimator (MLE) of α based on the exponentiated scale family (2) or (3). Then the distribution of $\alpha/\hat{\alpha}$ is $Gamma(n,\ 1)/n$ or equivalently $\chi^2_{2n}/2n$, where χ^2_{2n} is the Chi-square random variable with $2n$ degrees of freedom.

Proof: By probability integral transform, we have,

$$F^{\alpha}\left(\frac{X_i}{\lambda}\right) \overset{d}{\sim} U_i(0,1),$$

where U_i, $i = 1, 2, ..., n$ are independent standard uniform variates. Using standard results and independence of $X_1, X_2, ..., X_n$ we have,

$$-\alpha \sum_{i=1}^{n} \log(F(\frac{X_i}{\lambda})) \overset{d}{\sim} Gamma(n, 1). \tag{4}$$

Furthermore, the log-likelihood function for the exponentiated scale family is,

$$l(\alpha|\mathbf{X}) = nlog(\alpha) + (\alpha - 1)\sum_{i=1}^{n} \log(F(\frac{X_i}{\lambda})) + \sum_{i=1}^{n} \log(f(\frac{X_i}{\lambda})). \tag{5}$$

Equating the derivative of l with respect to α to zero, one gets

$$\hat{\alpha} = -\frac{n}{\sum_{i=1}^{n} \log(F(\frac{X_i}{\lambda}))}. \tag{6}$$

Noting the result in (4),

$$\frac{\alpha}{\hat{\alpha}} = \left(\frac{n}{-\alpha \sum_{i=1}^{n} \log(F(\frac{X_i}{\lambda}))}\right)^{-1} \overset{d}{\sim} \frac{Gamma(n,\, 1)}{n}, \text{which is same as} \quad \frac{\chi^2_{2n}}{2n}.$$

It is now clear that, the GPQ for a resilience parameter is $\mathcal{G}_{\alpha} = \frac{\hat{\alpha} W}{2n}$, where $W \sim \chi^2_{2n}$. Exactly similar arguments hold for a frailty parameter.

2.2 Confidence interval for R under common known scale parameter

It is easily verifiable that under common scale parameters, the reliability R for the exponentiated scale family is:

$$R(\alpha_1,\, \alpha_2) = \begin{cases} \dfrac{\alpha_2}{\alpha_1 + \alpha_2} & \text{resilience parameter family} \\ \dfrac{\alpha_1}{\alpha_1 + \alpha_2} & \text{frailty parameter family.} \end{cases} \tag{7}$$

Furthermore, it is easily deduced from the definition of GPQ that, if $\mathcal{G}_{\boldsymbol{\theta}}$ is a GPQ for $\boldsymbol{\theta}$, then GPQ for any function $\pi(\boldsymbol{\theta})$ is $\pi(\mathcal{G}_{\boldsymbol{\theta}})$. As such the GPQ for $R(\alpha_1,\, \alpha_2)$ is $\mathcal{G}_R = R(\mathcal{G}_{\alpha_1},\, \mathcal{G}_{\alpha_2})$, where $\mathcal{G}_{\alpha_1}$ and $\mathcal{G}_{\alpha_2}$ are GPQs of α_1 and α_2 respectively.

Let, X_1 and X_2 be independent but not identical random variables from distribution functions $G(\frac{x}{\lambda},\, \alpha_i), i = 1,\, 2$ respectively. By implementing the *algorithm* 1 given below an observation on $\mathcal{G}_R$ can be easily generated for the case of known common scale parameter.

Algorithm 1:

Step 1. For observed data $\mathbf{x}_i = \{x_{i1}, x_{i2}, ..., x_{in_i}\}$, compute MLEs for resilience / frailty parameters $\hat{\alpha}_i$, $i = 1, 2$.

Step 2. Generate independent random numbers W_1 and W_2 from $\chi^2_{2n_1}$ and $\chi^2_{2n_2}$ respectively.

Step 3. Compute GPQ for α_i, $G_{\alpha_i} = \frac{\hat{\alpha}_i W_i}{2n_i}, i = 1, 2$.

Step 4. For resilience parameter, compute $\mathcal{G}_R = \frac{\mathcal{G}_{\alpha_2}}{\mathcal{G}_{\alpha_1} + \mathcal{G}_{\alpha_2}}$ and for frailty parameter $\mathcal{G}_R = \frac{\mathcal{G}_{\alpha_1}}{\mathcal{G}_{\alpha_1} + \mathcal{G}_{\alpha_2}}$.

The *Algorithm* 1 can be repeated B times where B is a sufficiently large number, (say 10000) to generate B independent copies of $\mathcal{G}_R$. At the level of significance γ, the sample $(\gamma/2)^{th}$ and $(1-\gamma/2)^{th}$ quantiles $\xi_{\gamma/2}$ =L and $\xi_{1-\gamma/2}$ =U of the generated sample give the proposed interval estimate [L, U] for R.

2.3 Confidence interval for R under unequal known scale parameters

When scale parameters are known and unequal, that is $\lambda_1 \neq \lambda_2$, it is easily seen that

$$R = P(X_1 < X_2) = \int_v G_1(\eta v, \alpha_1) g_2(v, \alpha_2) dv, \tag{8}$$

where $\eta = \frac{\lambda_2}{\lambda_1}$, $G_i(.)$ and $g_i(.)$ correspond to the standard (with scale parameter equal to 1) CDF and PDF of the scale family under consideration. This integral most often may not exist in closed form and its numerical computations in standard packages like MATLAB and R often gave absurd results. An easier but very closely accurate computation can be attempted noting that,

$$\begin{aligned} R &= \int_0^\infty \{G_1(\eta v, \alpha_1) g_2(v, \alpha_2) e^v\} e^{-v} dv, \\ &= E_V[H(V, \alpha_1, \alpha_2, \eta)], \end{aligned} \tag{9}$$

where $H(V, \alpha_1, \alpha_2, \eta) = G_1(\eta V, \alpha_1) g_2(V, \alpha_2) e^V$ and $V \sim exp(1)$. This expectation can then be evaluated empirically by simulating a large number of standard exponential random numbers v_i, $i = 1, 2, ..., M$ and estimating R by $\bar{R}(\alpha_1, \alpha_2, \eta) = \{\sum_{i=1}^M H(v_i, \alpha_1, \alpha_2, \eta)\}/M$. For M larger than 10000 most often $\bar{R}(.)$ was found close to R up to $O(10^{-2})$. In *algorithm* 1, M independent copies of $\mathcal{G}_R$ can be generated using this computational procedure to produce $100(1-\gamma)\%$ CI for R.

Remark 1.

When the support of X_2 is the entire $\mathbb{R}$, the integral in (9) will be from $-\infty$ to ∞. Here use of the standard normal distribution for V is recommended instead of standard exponential distribution in the above procedure.

3 Confidence Intervals for R Under Unknown Scale Parameter

In case of unknown scale parameters the MLE $\hat{\alpha}_i$ of α_i being a function of MLE $\hat{\lambda}_i$ of λ_i, $i = 1, 2$ the distributional result proved in *Theorem* 1 does not hold exactly and the resulting GPQs and hence the above inferential procedures are approximate. As an alternative, four bootstrapping procedures, namely parametric and nonparametric bootstrap techniques employed with regular MLEs and MSILEs of the parameters under consideration were compared empirically. The MSILE is invariant under the nuisance scale parameter and is obtained by maximizing the likelihood $L^*(\alpha_i|\boldsymbol{y_i})$ of the transformed data $\mathbf{y}_1 = (y_{11}, y_{12}, ..., y_{1\,n_1-1})$ and $\mathbf{y}_2 = (y_{21}, y_{22}, ..., y_{2\,n_2-1})$ obtained by the following transformation and integrating over $y_{i\,n_i}$, $i = 1, 2$:

$$y_{ij} = \begin{cases} x_{ij}/x_{i\,n_i} & \text{for } j = 1, 2, ..., n_i - 1 \\ x_{i\,n_i} & \text{for } j = n_i; \qquad i = 1, 2. \end{cases}$$

Often computation of $L^*(\alpha_i|\boldsymbol{y_i})$ needs numerical integration which can be circumvented by the technique suggested in section 2.3, equation(9).

3.1 Bootstrap confidence interval for R under common unknown scale parameter

Following algorithm is used for computing a bootstrap CI for R. In the sequel, $\tilde{\alpha}_i$ denotes the MSILE of α_i and $\tilde{\lambda}$ is the maximizer of $L^*(\lambda|\tilde{\alpha}_1, \tilde{\alpha}_2, \mathbf{x}_1, \mathbf{x}_2)$.

Algorithm 2:

For $i = 1, 2$ follow the following steps:

Step 1. For observed data $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{in_i})$ compute $\tilde{\alpha}_i$, note that $\tilde{\alpha}_i$ does not depend on the unknown λ_i, $i = 1, 2$.

Step 2. Compute $\tilde{\lambda}$ for given $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ by maximizing the likelihood $L^*$for the combined sample.

Step 3. Generate bootstrap samples $\mathbf{x}_i^* = (x_{i1}^*, x_{i2}^*, ..., x_{in_i}^*)$. Note that for the parametric bootstrap $\mathbf{x}_i^*$ is generated from $G_i(\tilde{\lambda}, \tilde{\alpha}_i)$ while for the nonparametric bootstrap $\mathbf{x}_i^*$ is a random sample with replacement from $\mathbf{x}_i$.

Step 4. Obtain $\tilde{\alpha_i}^*$ based on $\mathbf{x}_i^*$.

Step 5. Next compute $\tilde{\lambda}^*$ for given $\tilde{\alpha_i}^*$ based on the combined sample $(\mathbf{x}_1^*, \mathbf{x}_2^*)$ as in step 2.

Step 6. Compute bootstrap estimate R^* by replacing α_i by $\tilde{\alpha_i}^*$ in (7).

Step 7. By repeating steps 2 to 6 generate sufficiently large number (say B=1000) of copies of bootstrap estimates, $\tilde{\mathbf{R}}_{boot} = (R_1^*, R_2^*, ..., R_B^*)$. The $(\gamma/2)^{th}$ and $(1-\gamma/2)^{th}$ sample quantiles of $\tilde{\mathbf{R}}_{boot}$ say L and U are the bootstrap confidence limits of R.

3.2 Bootstrap confidence interval for R under unequal unknown scale parameters

Confidence interval of R for unequal and unknown scale parameters $\lambda_1 \neq \lambda_2$, can also be obtained based on parametric and non-parametric bootstrap technique. In this case the estimates $\tilde{\lambda}_i$ in step 2 of *Algorithm* 2 are to be independently computed from the respective sample for given $\tilde{\alpha}_i$, $i = 1, 2$. The procedure to compute bootstrap estimates $\tilde{\alpha}_i^*, \tilde{\lambda}_i^*$, $i = 1, 2$ is the same as in *Algorithm* 2, while the reliability estimate in Step 6 should be computed based on equation (9) by replacing η by $\tilde{\eta}^* = \frac{\tilde{\lambda}_2^*}{\tilde{\lambda}_1^*}$ and α_i by their bootstrap estimates $\tilde{\alpha}_i^*$, $i = 1, 2$. Rest of the procedure is same as above.

4 Empirical Assessment

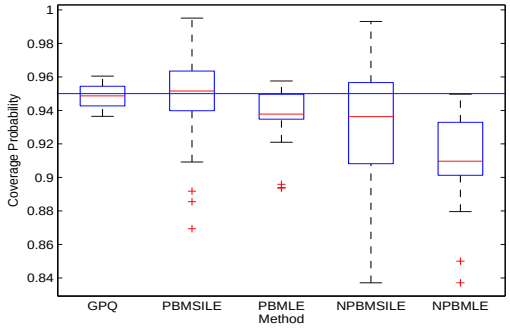
EED being a widely used distribution from the exponentiated scale family, is employed for the comparative empirical study. A comparative study is attempted among the methods discussed in sections (2) and (3). Performances of all the methods are assessed based on the estimates of coverage probability and average widths on 2500 simulations. The parametric combinations considered are: sample sizes (n_1, n_2)=(10, 10), (10, 30), (10, 40), (30, 30), (30, 40), (40, 40). Since the procedure is invariant under common scale parameter, the value of λ=10 is fixed while the resilience parameter is chosen to be α_1=0.5, 1, 2, 5. α_2 is adjusted such that R=0.1, 0.4, 0.7, 0.9. Under unequal scale parameters, the parameters are set to α_1 = 0.5, 5, α_2 = 1, 6 and for each λ_1=0.5, 5 the values of η are set to 0.5, 2, 5, 10 by adjusting $\lambda_2 = \eta\lambda_1$. The level of significance used is $\gamma = 0.05$.

The following five methods are compared:

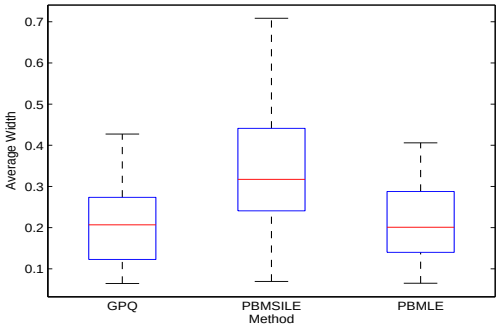
GPQ: Generalized pivotal quantity
PBMSILE: A parametric bootstrap technique employed on MSILE
PBMLE: A Parametric bootstrap technique employed on MLE
NPBMSILE: A non-parametric bootstrap technique employed on MSILE
NPBMLE: A non-parametric bootstrap technique employed on MLE

Figure 1 displays the results for known scale parameters. Figure 1 (a) and (c) display the box plots of coverage probabilities of CI's under equal and unequal scale parameters respectively

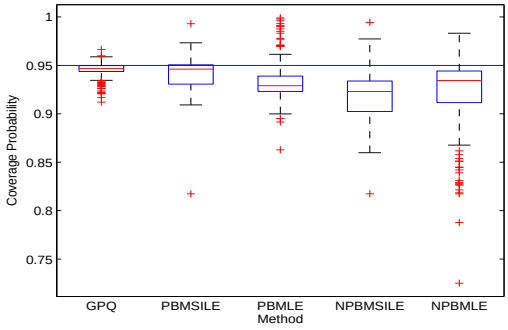
while figure 1 (b) and (d) display corresponding average widths of those procedures exhibiting satisfactory coverage performance. Figure 2 displays the counterparts under unknown scale parameters. Here the performance of GPQ based CI was not recommendable and is omitted.



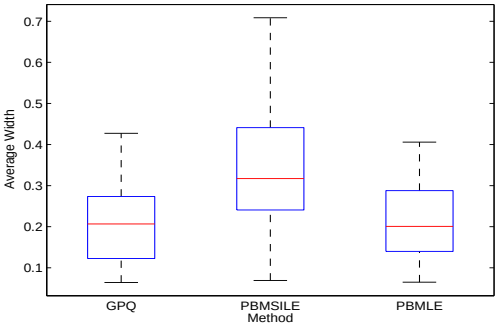
(a) *Coverage probabilities under common scale parameter.*



(b) *Average widths of CI with sizes close to the nominal level*

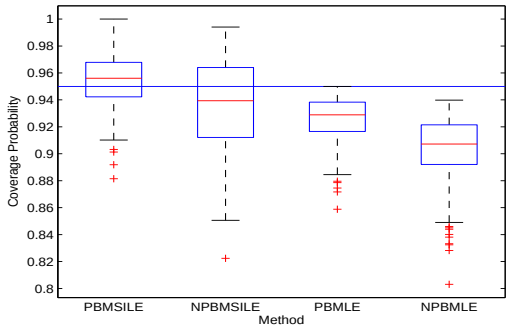


(c) *Coverage probabilities under unequal scale parameters.*

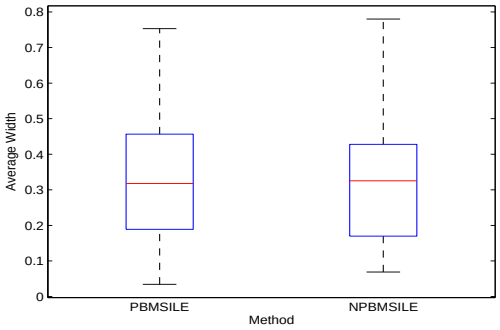


(d) *Average widths of CI with sizes close to the nominal level*

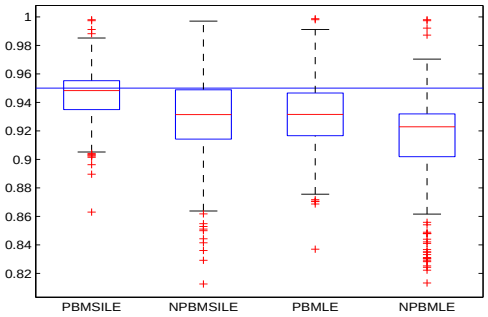
Figure 1: *Box plots of simulated coverage probabilities and average widths of CI's (conforming the size performance) for R under known scale parameters.*



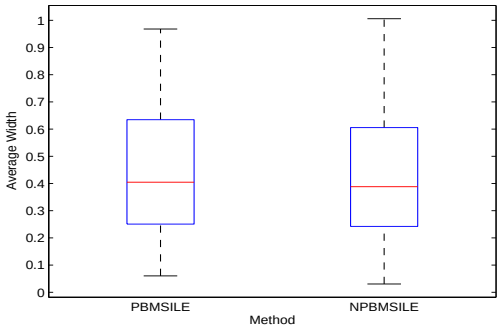
(a) *Coverage probabilities under common scale parameter.*



(b) *Average widths of CI with sizes close to the nominal level*



(c) *Coverage probabilities under unequal scale parameters.*



(d) *Average widths of CI with sizes close to the nominal level*

Figure 2: *Box plots of simulated coverage probabilities and average widths of CI's (conforming the size performance) for R under unknown scale parameters.*

For the known scale parameters, observation of Figure 1 reveals that GPQ based CI are clearly outperforming the rest with respect to both the criteria. Here the coverages are very well concentrated around the nominal level with shorter widths among others. For the case of unknown scales, observation of Figure 2 reveals that PBMSILE outperforms the rest and is recommended.

5 Real Life Applications

The data given in Table 1 (http://reliawiki.org/index.php/Stress-Strength_Analysis) reports the miles traveled by 20 sold vehicles in a year (stress variable: X_1) and miles traveled before failure of an independent sample of 50 new vehicles of the same type (strength variable: X_2). The probability of vehicle failure during a period of one year can be estimated by using the stress-strength reliability analysis, discussed in section 2.

Table 1: *Miles traveled by vehicle per year and miles traveled before failure (In 100 miles).*

Stress (X_1)	100.96	111.83	115.34	121.41	125.36	137.77	125.95	138.62	125.27	109.55
(miles traveled	104.69	113.91	119.19	124.05	140.32	141.38	126.57	139.71	121.05	114.86
in a year)										
Strength (X_2)	135.07	161.25	148.10	167.49	155.22	174.30	139.43	163.27	149.51	168.62
(miles traveled	137.93	163.20	149.40	167.93	155.47	178.05	140.17	163.49	151.04	169.30
before failure)	141.47	164.06	152.18	169.48	160.03	185.75	143.76	166.11	153.11	170.41
	143.51	165.01	153.03	170.24	160.18	188.13	145.95	166.25	154.80	172.63
	155.70	178.84	159.75	185.49	160.52	189.44	147.46	166.70	154.96	173.47

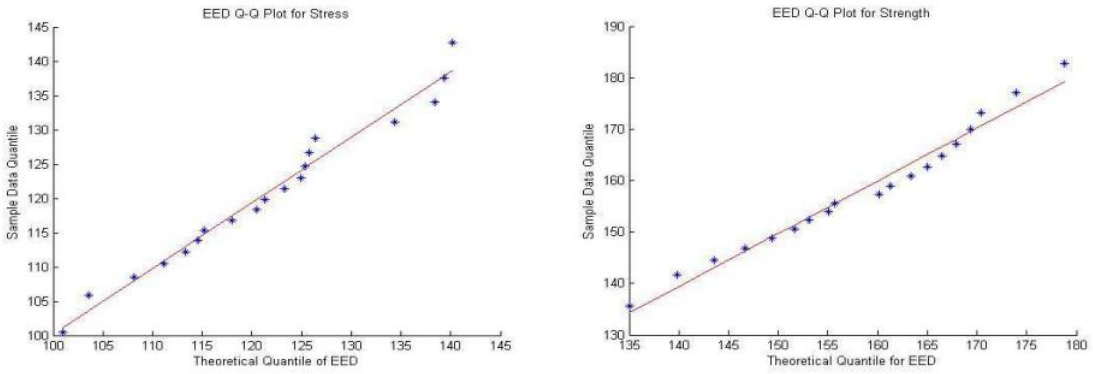
The Table 2 below reports the Akaike information criterion (AIC), Bayesian information criterion (BIC) and P-value corrosponding to Kolmogrov-Smirnov test statistics (KS test) for fitting a best life distribution to the data presented in Table 1.

Table 2: *Goodness of fit to the data given in Table 1.*

Distribution	Stress			Strength		
	P-Value	AIC	BIC	P-Value	AIC	BIC
Weibull	**0.5186**	**160.3475**	**162.3390**	**0.7497**	**409.8429**	**413.6669**
Exponential	0.0000	234.4878	235.4835	0.0000	610.0452	611.9572
Gamma	0.7520	8233.1490	8235.1400	0.9781	28032.1200	28035.9400
Log-Normal	0.8076	2612.3390	2614.331	0.9611	8981.6030	8985.4270
Normal	0.7264	163.7162	162.7076	0.9646	416.6732	420.4973
Pareto	0.0956	170.2461	172.2376	0.0037	434.7542	438.5783
EED	**0.9405**	**160.8999**	**162.8913**	**0.5319**	**407.5343**	**411.3584**

The three criteria together indicate that the Weibull and EED are two almost equally best fitted

distributions to both the variables. To illustrate the estimation of R, using the results of previous sections it is reasonable to assume that the data is coming from EED.



(a) **(b)**

Figure 3: *EED Q-Q plot for Stress* (X_1) *and Strength* (X_2) *variables.*

The Q-Q plots for the two data sets given in Figure 3 also support in favor of the EED to both X_1 and X_2. MLEs of the scale parameters for (X_1) and (X_2) are respectively 10.8707 and 12.1455. The CI and point estimates for $R = P(X_1 < X_2)$ computed by the four methods are presented in Table 3. Based on the likelihood ratio test for equality of scales (P-value = 0.5834) it seems reasonable to assume that the samples are coming from populations with common scale parameters, so that based on the recommendations, the CI obtained by PBMSILE are most reliable.

The point estimates obtained by mean and median of bootstrap samples and confidence intervals of stress-strength reliability for four bootstrap methods are depicted in the following Table 3:

Table 3: *Point Estimates and CI for R.*

Method	Point Estimate		Confidence Interval
	Mean	Median	(L, U)
PBMSILE	**0.9833**	**0.9964**	**(0.9570, 0.9986)**
PBMLE	0.958	0.9597	(0.9219, 0.9824)
NPBMSILE	0.9844	0.9965	(0.9570, 0.9986)
NPBMLE	0.9567	0.959	(0.9253, 0.9803)

The confidence interval of reliability computed with PBMSILE (as well as NPBMSILE) indicates that the probability of instantaneous failure of a vehicle within a year lies in (0.0014, 0.043), which

is very low. It thus follows that setting one year warranty period for a vehicle of this brand is almost risk free.

A similar analysis for the data on number of pages printed by printers (http://www.weibull.com/hotwire/issue163/ hottopics163.htm) also best fitted with EED. Here, the number of pages printed by users in one year is the stress variable X_1 and the number of pages printed before the component failed during in-house testing is the strength variable X_2. The resulting CI for R is (0.9859, 0.9999), indicating that the probability of failure within a year lies between $(0.0001, 0.0141)$ which is very small and here also one year warranty period for the printers can be set with almost no risk.

6 Concluding Remarks

Efficient estimation of the stress-strength reliability is of prime importance in reliability applications, particularly in setting warranty period for products to be launched in the market. We have addressed this issue when the distribution of stress and strength belongs to exponentiated scale family. When the scale parameters are known, interval estimation of R based on GPQ is recommended while the case of unknown scale parameters, is recommended to be handled through the parametric bootstrap approach based on the maximum scale invariant likelihood estimator of the shape parameter.

Acknowledgment

The Authors are very much thankful to Council of Scientific and Industrial Research (CSIR), India, for financial support for this research work, under the grant sanction Order No. 25(0211)/13/EMR-II. We also thank to the referees and the editor(s) for helpful comments leading to the improved version of the work.

References

1. Ahmad K. E., Fakhry M.E., Jaheen Z.F. Empirical Bayes estimation of $P(Y < X)$ and characterizations of Burr-type X model, *Journal of Statistical Planning and Inference* 1997; **67**: 297-308. DOI: 10.1016/S0378-3758(97)00038-4

2. Gupta R. D., Kundu D. Theory and Methods: Generalized exponential distributions, *Australian and New Zealand Journal of Statistics* 1999; **41(2)**: 173-188. DOI: 10.1111/1467-842X.00072

3. Gupta R. D., Kundu D. Exponentiated Exponential Family: An Alternative to Gamma and Weibull Distribution, *Biometrical Journal*,2001; **43(1)**: 117-130.

4. Gupta R. D., Kundu D. Generalized exponential distribution: different method of estimations, *Journal of Statistical Computation and Simulation* 2001; **69(4)**: 315-337. DOI: 10.1080/00949650108812098

5. Gupta R. D., Kundu D. Generalized exponential distribution: Existing results and some recent developments, *Journal of Statistical Planning and Inference* 2007; **137(11)**: 3537-3547.

6. Hannig J., Iyer H., Patterson P. Fiducial generalized confidence intervals, *Journal of the American Statistical Association* 2006; **101**: 245-269. DOI:10.1198/016214505000000736

7. Kotz S., Lumelskii Y., Pensky M. *The Stress-Strength Model and its Generalizations Theory and Applications*, World Scientific, New York 2003; ISBN: 978-981-4488-19-8.

8. Kundu D., Gupta R. D. Estimation of $P[Y < X]$ for generalized exponential distribution, *Metrika* 2005; **61(3)**: 291-308. DOI: 10.1007/s001840400345

9. Marshall A. W., Olkin I. *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families* , Springer Series in Statistics 2007; ISBN 978-0-387-68477-2.

10. Nadarajah S., Kotz S. The Exponentiated Type Distributions, *Acta Applicandae Mathematica* 2006; **92**: 97-111. DOI 10.1007/s10440-006-9055-0

11. Raqab M. Z., Madi M. T., kundu D. Estimation of $P(Y < X)$ for the Three-Parameter Generalized Exponential Distribution, *Communications in Statistics - Theory and Methods* 2008;**37(18)**: 2854-2864. DOI: 10.1080/03610920802162664

12. Roy A., Bose A. Coverage of generalized confidence intervals, *Journal of Multivariate Analysis* 2009;**100(7)**: 1384-1387, DOI:10.1016/j.jmva.2008.12.001

13. Surles, J.G., Padgett, W. J. Inference for Reliability and Stress-Strength for a Scaled Burr-Type X Distribution, *Lifetime Data Analysis* 2001; **7(2)**: 187-200. DOI: 10.1023/A:1011352923990

14. Tsui, K.W., Weerahandi, S. Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters, *Journal of the American Statistical Association* 1989; **84**: 602-607. DOI: 10.2307/2289949

15. Zhou W. Statistical inference for $P(X < Y)$, *Statistics in Medicine* , 2008; **27**: 257-279. DOI: 10.1002/sim.2838

Journal

Communications in Statistics - Theory and Methods ›

Volume 46, 2017 - Issue 18

Original Articles

Economic design of non parametric sign control chart

S. H. Patil & D. T. Shirke

Pages 8987-8998 | Received 30 Jul 2015, Accepted 02 Jun 2016, Accepted author version posted online: 23 Aug 2016, Published online: 14 Jun 2017

Download citation https://doi.org/10.1080/03610926.2016.1200095

ABSTRACT

An economic design of sign chart to control the median is proposed. Since the sign chart is distribution free, it can easily be applied to any process without prior knowledge of process quality distribution. The effect on loss cost observed for different shifts in location shows that the sign chart performs better for large shifts. The economic statistical performance study reveals that statistical performance of sign chart can be improved sufficiently for moderate shifts in the process. Sensitivity study shows that design is more sensitive for change in values of penalty loss cost and time required for search and repair of an assignable cause.

KEYWORDS: Economic design, economic statistical design, non parametric control chart, non normal process, optimal design, sign control chart.

Economic design of moving average control chart for non-normally distributed data: a comparative study

S.H. Patil*

Doodhsakhar Mahavidyalaya, Bidri,
Tal: Kagal, Dist: Kolhapur,
Maharashtra 416208, India
Email: psubhash_2007@rediffmail.com
*Corresponding author

D.T. Shirke

Department of Statistics,
Shivaji University,
Kolhapur, Maharashtra 416004, India
Email: dts_stats@unishivaji.ac.in

Abstract: An economic cost function for moving average control chart under non-normal quality characteristic is obtained using continuous and ceased production modes. This cost function can be used to obtain optimal design parameters which minimise the loss cost per hour of the production. Burr distribution has been used to transform the non-normal data to normal. Sensitivity analysis of the economic design is also discussed. The ceased process is observed to be dominant with significant saving in the cost over the continuous process.

Keywords: moving average; economic design; burr distribution; continuous process; ceased process; non-normal data.

Reference to this paper should be made as follows: Patil, S.H. and Shirke, D.T. (2017) 'Economic design of moving average control chart for non-normally distributed data: a comparative study', *Int. J. Operational Research*, Vol. 28, No. 1, pp.98–120.

Biographical notes: S.H. Patil received his Masters degree in Statistics from the Shivaji University, Kolhapur, Maharashtra (India). He is completing his PhD at the Department of Statistics, Shivaji University, Kolhapur. He is an Assistant Professor at the Department of Statistics, Doodhsakhar Mahavidyalaya, Bidri, Kolhapur. His research interests include statistical quality control and optimisation.

D.T. Shirke received his PhD in Statistics from the Shivaji University, Kolhapur, Maharashtra (India). He is a Professor in the Department of Statistics at the Shivaji University, Kolhapur, Maharashtra. His research interests include statistical process control and asymptotic inference. He has published several research papers in national and international journals.

This paper is a revised and expanded version of a paper entitled 'Economic design of moving average control chart for non-normal data using ceased production process' presented at National Seminar on Stochastic Modelling and Analysis organised by Department of Statistics, Cochin University of Science and Technology, Cochin, 25 March 2011.

1 Introduction

Control chart technique was first developed by Shewhart (1931), as an online process control technique to control the variability in a production process. Since then, there has been lot of developments taken place in the construction of control charts. This includes development of control chart for normal as well as non-normal process distributions. In the design of control chart, the decision about the sample size (n), sampling interval (h) and the control limit multiplier (k) of the control chart is made to monitor the manufacturing process in control according to statistical and/or an economic criteria. These three factors are generally known as control chart parameters. In statistical control chart, the design parameters are chosen in such a way that the chart meets some statistical requirements. In the economic design, the overall focus is given on the minimisation of the total loss cost from the process (that is maximisation of the profit). In economic-statistical design, while minimising the production cost, some statistical constraints are applied to the process.

Duncan (1956) was the first to propose an economic design for $\bar{x}$ control chart. Since then, based on his foundation, many other researchers have been working on the economic design for different type of control charts. Montgomery (1980), Collani (1986), McWilliams (1989), Saniga (1989), Rahim and Banerjee (1993), Yu and Chen (2005) and many others have worked on the different types of economic designs. Saniga (1989) first considered the economic-statistical design for $\bar{x}$ and R chart by applying statistical constraints on type I and type II error probabilities and concluded that the design gives better performance as compared to the fully economic model at slight increase in the cost. Al-oraini and Rahim (2003) also concluded in the same way. Zhang and Berardi (1997), Chou et al. (2000), Chen and Cheng (2007) and Yeh and Chen (2010) have also worked on this type of design of $\bar{x}$ control chart. Alkhedher and Darwish (2013), Sing et al. (2014), Hashemi et al. (2014) have also reported optimisation procedures in various applications.

Wu et al. (2008) have developed an optimum design of combined $\bar{x}$ and cumulative sum (CUSUM) chart based on extra quadratic loss and compared performance of the chart with single charts. Trovato et al. (2010) have economically compared several control strategies including Shewhart, Exponentially Weighted Moving Average (EWMA) and CUSUM to monitor the process dispersion in short run. They conclude that while monitoring short run process, production and inspection rates are to be estimated accurately. Wu et al. (2010) have compared performance of two CUSUM schemes for shift in mean and variance using optimum design. Nenes (2011) has provided unified approach for the economic optimisation of different variable parameter (VP) control charts. He has developed a single cost function to optimise the chart parameters. Mahadik and Shirke (2011) have compared performance of variable sampling interval (VSI), variable sample size and sampling interval (VSSI) and special variable sample size and

sampling interval (SVSSI) T^2 control charts in terms of steady state average time to signal (SSATS). Kolli and Limam (2011) have developed an economic design for np control chart. Zhang Wu et al. (2011) have developed the optimisation design treating sampling inspection cost and in control average time to signal (ATS) as adjustable parameters for $\bar{x}$ and CUSUM chart.

Ou et al. (2012) have proposed an optimal sequential probability ratio test (SPRT) control chart and compared its performance with basic SPRT chart. Lee et al. (2012) designed a control chart with double sampling and VSI. Khoo et al. (2013) and Tu (2013) have developed economic and economic-statistical designs for two unit series system $\bar{x}$ chart and synthetic $\bar{x}$ chart. Amiri et al. (2013) provided economic statistical design of modified exponentially weighted moving average (MEWMA) control chart. Franco et al. (2014) have investigated economic-statistical design of $\bar{x}$ charts using skip sampling strategies for autocorrelated processes. A&L switching rule is provided to reduce the switching between sampling intervals in VSI control charts. Guo et al. (2014) have studied the economic design of variable parameter $\bar{x}$ chart with a correlated A&L switching rule. Rostami and Ali (2014) have provided approximation algorithm for minimum cost flows. Basically, in most of these studies, the quality characteristic is assumed to be normally distributed and hence for mean charts, the distribution of mean becomes normal. But, sometimes the quality characteristic may not have normal distribution. Considering this situation, Rahim (1985), Chou et al. (2001), Chen (2004), Chen and Yeh (2006) have developed economic design for $\bar{x}$ control chart for non-normal data, under different situations.

In the literature, there are good number of research papers on economic design of $\bar{x}$ control chart, but relatively less on the economic design of moving average (MA) control chart. Chen and Yang (2002), Chen and Yu (2003), Yu and Chen (2005) and Yu and Wu (2004) have reported economic design of MA control charts. These designs are also based on normal quality characteristics and most of them are for the continuous process. Patil and Rattihalli (2009) have proposed the economic design for continuous as well as ceased MA process control chart. In this paper, considering non-normal input quality characteristics, we have developed the combined loss cost function for MA control chart under continuous, ceased and semi-ceased process model using unified approach by Lorenzen and Vance (1986). While developing the cost function, the approach by Yu and Chen (2005) is found to be useful. The cost function is optimised with respect to design parameters n, h and k, and the effect on loss cost is observed particularly for continuous and ceased process. The sensitivity of the design is carried out by applying change in the input parameters. While developing the cost function we have used different input cost and time parameters, which may be estimated by original sampled data from trial production.

The present paper consists of seven sections excluding the present one. In Section 2, we have introduced a process model and given definitions of notations used in the paper. A short description about Burr distribution is given in Section 3. Expressions for expected cycle length and expected loss cost are obtained in Section 4 and Section 5 respectively. An example is given in Section 6 and the sensitivity analysis is carried out in Section 7. Final conclusions are presented in Section 8.

2 Process model and notations

Consider the production process monitored by drawing a single unit sample at the interval of every 'h' hour. It is assumed that the time to take the sample, inspect it and to draw the conclusion is negligible. Let x_1, x_2 … ... be the sample observations collected on the process quality characteristic of the process and are assumed to have distribution with mean μ and known variance σ^2. The moving average of the observations at time t is given by,

$$\begin{aligned} \bar{x}_m &= \frac{1}{t}\sum_{i=0}^{t-1} X_{t-i}; \qquad \text{if } t < n \\ &= \frac{1}{n}\sum_{i=0}^{n-1} X_{t-i}; \qquad \text{if } t \geq n \end{aligned} \tag{1}$$

The process is monitored by a single assignable cause with the process target value as the first moment E(x) of the distribution and is denoted by μ_0. The shift in the process target value is instantaneous and whenever, the process shifts to out of control state (assignable cause occurs), the target mean shifts from μ_0 to $\mu_0 + \delta\sigma$, and otherwise it remains at μ_0.

The process is assumed to be start in control state and the period up to next in control state through an out of control state is termed to be one production cycle. That is, the time between start of two successive in control states is termed as a production cycle. This production cycle consists of in control period as well as out of control period. The model targets to find expected cycle length of this cycle and the total expected loss during the cycle, so that, expected loss per unit time from the process is obtained and optimised. The notations used are as follows.

n	sample size
h	sampling interval
k	control limits coefficient
δ	magnitude of shift in the process
λ	parameter of the exponential life time distribution for in control state
c_1, k_1	parameters of the burr distribution
$\delta_1 = 0$	if the process is ceased during the search of an assignable cause
$= 1$	if the process is continued during the search of an assignable cause
$\delta_2 = 0$	if the process is ceased for the rectification (repair) of an assignable cause
$= 1$	if the process is continued during the rectification of an assignable cause
α	the probability of false alarm
T_0	expected time to search for an assignable cause during the false alarm
T_1	expected time to search for an assignable cause during the true alarm
T_2	expected repair time of an assignable cause during the true alarm

C_0	expected loss cost per unit time due to the nonconformities produced when the process is in control
C_1	expected loss cost per unit time due to the nonconformities produced when the process is out of control
C_2	expected production loss cost per unit time due to ceasing of the process
ASN	average sample number to detect the shift
N	no. of samples during the cycle
g	time to sample one unit
a, b	fixed and variable costs of sampling
V	loss cost due to single false alarm
W	loss cost due to search and repair of an assignable cause.

3 The Burr distribution

The Burr distribution is discussed in detail by Burr (1942) to represent various types of non-normal distributions. The probability density function of Burr distribution with parameters c_1 and k_1 is given by,

$$\begin{aligned} f(y) &= \frac{c_1 k_1 y^{c_1-1}}{(1+y^{c_1})^{k_1+1}}; \quad y \geq 0,\ c_1 \geq 1,\ k_1 \geq 1. \\ &= 0; \quad y < 0 \end{aligned} \tag{2}$$

and its cumulative distribution function is given by,

$$\begin{aligned} F(y) &= 1 - \frac{1}{(1+y^{c_1})^{k_1}}; \quad y \geq 0,\ c_1 \geq 1,\ k_1 \geq 1. \\ &= 0; \quad y < 0 \end{aligned} \tag{3}$$

One can apply first four moments or 3rd and 4th moments of the underlying distribution, as the case may be, to approximate the parameters (c_1, k_1) of the Burr distribution. The resulting coefficient of skewness and kurtosis from the function cover the broad range within which many empirical and theoretical distribution lies. Further, the Burr distribution can be approximated by the normal or gamma distribution (Chen and Yeh, 2006).

Burr (1942) has provided two tables. In Table 2 values of the mean and standard deviation (S.D.) of the Burr distribution are given and in Table 3, values of skewness and kurtosis coefficients for different values of parameters c_1 and k_1 are provided. From the original sampled data mean ($\bar{x}$), , variance (S_x^2), coefficient of skewness (α_3) and coefficient of kurtosis (α_4) can be obtained. Using Table 3 and the values of coefficient of skewness and kurtosis one can estimate the values of parameters (c_1, k_1) of the family of Burr distribution. Using these values of c_1 and k_1 and with the help of Table 2, mean (M) and S.D. (S) for the Burr distribution can be estimated. Using these estimated values of

mean and variance, the standardised transformation between Burr variable (y) and any other variable (x) is made by,

$$\frac{Y-M}{S}=\frac{X-\bar{x}}{S_x} \tag{4}$$

This transformation is useful to detect type I and type II error probabilities of the control chart, when the incoming data is non-normal.

4 Expected cycle length

The expected cycle length consists of the in control period, time for search due to false alarm, the out of control period, time for sampling and testing and the time for search and repair of an assignable cause. Assuming that, an assignable cause occurs according to the Poisson process of an intensity of λ occurrences per unit time (That is the distribution of an in control time has an exponential distribution with parameter λ), the expected in control time becomes $1/\lambda$.

The probability of false alarm is the probability that, the test statistic falls outside the control limits, when the process is in control. Our test statistic is based on the moving average and has mean μ_0 and variance σ^2/n. Hence,

$$\begin{aligned}\alpha &= P(\bar{x}_m < \mu_0 - k\sigma/\sqrt{n},\ \bar{x}_m > \mu_0 + k\sigma/\sqrt{n});\ \text{at mean} = \mu_0,\\ &= 1 + P(\bar{x}_m < \mu_0 - k\sigma/\sqrt{n}) - P(\bar{x}_m < \mu_0 + k\sigma/\sqrt{n}):\ \text{at mean} = \mu_0,\end{aligned} \tag{5}$$

According to standardised transformation between Burr variable (Y) and r.v. ($\bar{x}_m$), we get,

$$\frac{Y-M}{S}=\frac{\bar{x}m-\mu_0}{\sigma/\sqrt{n}}.$$

This gives,

$$\bar{x}_m=\mu_0+\frac{(Y-M)\sigma}{S\sqrt{n}},$$

Using this in (5), we get,

$$\begin{aligned}\alpha &= 1 + P\left(\mu_0+\frac{(Y-M)\sigma}{S\sqrt{n}} < \mu_0 - k\sigma/\sqrt{n}\right)\\ &\quad - P\left(\mu_0+\frac{(Y-M)\sigma}{S\sqrt{n}} < \mu_0 + k\sigma/\sqrt{n}\right)\\ &= 1 - P(Y < M - kS) + P(Y < M + kS)\\ &= 1+\frac{1}{\left[1+(M+kS)^{c_1}\right]^{k_1}}-\frac{1}{\left[1+(M-kS)^{c_1}\right]^{k_1}}\end{aligned} \tag{6}$$

Let, s denotes expected number of samples taken during in control period, then

$$s = \sum_{i=0}^{\infty} i\left(e^{-\lambda ih} - e^{-\lambda(i+1)h}\right) = \frac{e^{-\lambda h}}{1-e^{-\lambda h}} \tag{7}$$

and, the expected time elapsed due to false alarm = $T_0 \alpha s$.

Therefore,

$$\text{In control time} = IT = \frac{1}{\lambda} + (1-\delta_1)T_0 \alpha s. \tag{8}$$

Let, T be expected time of occurrence of an assignable cause in i^{th} and $(i+1)^{th}$ sample, then

$$T = \frac{e^{\lambda h} - (1+\lambda h)}{\lambda(e^{\lambda h} - 1)} \tag{9}$$

Let, $\bar{u}_{ij}$, i = 0, 1, 2 … … n–1; j = 1, 2 … ... be the mean of moving group, when an assignable cause occurs in i^{th} and $(i+1)^{th}$ sample and is detected in j subsequent moving subgroup and p_{ij} be the probability that an assignable cause occurs in i^{th} and $(i+1)^{th}$ sample and is detected in j subsequent moving subgroup, then

$$\bar{u}_{ij} = \begin{cases} \mu_0 + \dfrac{\delta j}{i+j}\sigma & ; i+j < n \\ \mu_0 + \dfrac{\delta j}{n}\sigma & ; i+j \geq n, j < n \\ \mu_0 + \delta\sigma & ; i+j, j \geq n \end{cases} \tag{10}$$

Further,

$$P_{ij} = P(\bar{x}_m < \mu_0 - k\sigma/\sqrt{n},\ \bar{x}_m > \mu_0 + k\sigma/\sqrt{n});\ \text{at mean} = \bar{u}_{ij}.$$

Hence by standardised transformation of $\bar{x}_m$ to Burr variable, we get

$$\frac{Y-M}{S} = \begin{cases} \dfrac{\bar{x}_m - \mu_0 - \dfrac{\delta j}{i+j}\sigma}{\sigma/\sqrt{i+j}} & ; i+j < n \\ \dfrac{\bar{x}_m - \mu_0 - \dfrac{\delta j}{n}\sigma}{\sigma/\sqrt{n}} & ; i+j \geq n, j < n \\ \dfrac{\bar{x}_m - \mu_0 - \delta\sigma}{\sigma/\sqrt{n}} & ; i+j, j \geq n. \end{cases}$$

This gives,

$$\bar{x}_m = \begin{cases} \mu_0 + \dfrac{\delta j}{i+j}\sigma + \dfrac{(Y-M)\sigma}{S\sqrt{i+j}} & ; i+j < n \\ \mu_0 + \dfrac{\delta j}{n}\sigma + \dfrac{(Y-M)\sigma}{S\sqrt{n}} & ; i+j \geq n, j < n \\ \mu_0 + \delta\sigma + \dfrac{(Y-M)\sigma}{S\sqrt{n}} & ; i+j, j \geq n. \end{cases}$$

Hence,

$$p_{ij} = \begin{cases} P\left(Y < M - kS\sqrt{\dfrac{i+j}{n}} - \dfrac{\delta jS}{\sqrt{i+j}}, \right. \\ \quad \left. Y > M + kS\sqrt{\dfrac{i+j}{n}} - \dfrac{\delta jS}{\sqrt{i+j}} \right) & ; i+j < n \\ P\left(Y < M - kS - \dfrac{\delta jS}{\sqrt{n}}, Y > M + kS - \dfrac{\delta jS}{\sqrt{n}} \right) & ; i+j \geq n, j < n \\ P\left(Y < M - kS - \delta S\sqrt{n}, Y > M + kS - \delta S\sqrt{n} \right) & ; i+j, j \geq n. \end{cases}$$

$$= \begin{cases} 1 + \dfrac{1}{\left[1 + \left(M + kS\sqrt{\dfrac{i+j}{n}} - \dfrac{\delta jS}{\sqrt{i+j}} \right)^{c_1} \right]^{k_1}} \\ \quad - \dfrac{1}{\left[1 + \left(M - kS\sqrt{\dfrac{i+j}{n}} - \dfrac{\delta jS}{\sqrt{i+j}} \right)^{c_1} \right]^{k_1}} & ; i+j < n \\ 1 + \dfrac{1}{\left[1 + \left(M + kS - \dfrac{\delta jS}{\sqrt{n}} \right)^{c_1} \right]^{k_1}} - \dfrac{1}{\left[1 + \left(M - kS - \dfrac{\delta jS}{\sqrt{n}} \right)^{c_1} \right]^{k_1}} & ; i+j \geq n, j < n \\ 1 + \dfrac{1}{\left[1 + \left(M + kS - \delta S\sqrt{n} \right)^{c_1} \right]^{k_1}} - \dfrac{1}{\left[1 + \left(M - kS - \delta S\sqrt{n} \right)^{c_1} \right]^{k_1}} & ; i+j, j \geq n. \end{cases} \quad (11)$$

We have,

$$q_{ij} = 1 - p_{ij}.$$

Also, the probability in the last case is independent of i and j, and can be taken as constant (P), so that Q = 1-P.

If an assignable cause occurs in i^0 and $(i+1)^{th}$ sample and is detected in next j samples, the expected number of units to detect the shift (e_i) are given by,

$$e_i = p_{i1} + 2q_{i1}p_{i2} + 3q_{i1}q_{i2}p_{i3} + \ldots..$$
$$= \begin{cases} p_{i1} + \sum_{j=2}^{n-1} jp_{ij} \prod_{d=1}^{j-1} q_{id} + \prod_{d=1}^{n-1} q_{id}(n + \frac{Q}{P}) & ;i < n \\ e_{n-1} & ;i \geq n \end{cases} \quad (12)$$

Hence the ASN is given by,

$$ASN = \left(1 - e^{-\lambda h}\right) \sum_{i=0}^{n-2} e_i + e^{-(n-1)\lambda h} e_{n-1}. \quad (13)$$

Therefore, average out of control time (AOT) is given by,

$$AOT = h * ASN - T \quad (14)$$

and expected cycle length is given by,

$$E(T) = IT + AOT + g + T_1 + T_2. \quad (15)$$

5 Expected loss cost function during the cycle

The loss cost consists of loss during in control and out of control period, loss cost due to false alarm, cost for search and repair and cost of sampling and testing.

The loss cost due to non-conformities produced is given by,

$$L_1 = C_0 \left[\frac{1}{\lambda}\right] + C_1 \left[(h * ASN - T) + g + \delta_1 T_1 + \delta_2 T_2\right]$$

If, $C = C_1 - C_0$, is the overhead cost due to nonconformities produced in the out of control state, then

$$L_1 = C_0 \left[\frac{1}{\lambda} + (h * ASN - T) + g + \delta_1 T_1 + \delta_2 T_2\right] + C\left[(h * ASN - T) + g + \delta_1 T_1 + \delta_2 T_2\right], \quad (16)$$

The production loss cost due to ceasing of process during false alarm is,

$$L_2 = C_2 (1 - \delta_1) T_0 \alpha s \quad (17)$$

If V and W are the cost for search and repair when there is false alarm and true alarm respectively, then cost for search and repair is,

$$L_3 = V\alpha s + W \quad (18)$$

The expected numbers of samples during the cycle are,

$$E(N) = \frac{\frac{1}{\lambda} + (h * ASN - T) + g + \delta_1 T_1 + \delta_2 T_2}{h} \quad (19)$$

Therefore, cost of sampling and testing is,

$$L_4 = (a + b) * E(N) \tag{20}$$

Therefore, expected total loss cost per unit time during the cycle is,

$$E(L) = \frac{L_1 + L_2 + L_3 + L_4}{E(T)} \tag{21}$$

The above equation represents a general loss cost function derived for continued as well as for ceased non-normal production processes and is a function of design parameters n, h and k. Here, 'n' is integer whereas 'h' and 'k' may be positive real numbers. If we put $\delta_1 = \delta_2 = 1$, the cost function represents loss cost function for continuous process, whereas if we put $\delta_1 = \delta_2 = 0$, it represents loss cost function for ceased process. If we put $\delta_1 = 1$, $\delta_2 = 0$, the function represents loss cost function for (1-0) semi-ceased process, where the process is continued during the search of an assignable cause and ceased during repair of the cause, if occurred (Patil and Rattihalli, 2009).

The cost function developed is a complicated function of three design parameters (n, h, k), hence to get desired values of parameters and the expected loss, we use algorithmic procedure instead of direct derivative method. The optimisation objective function to develop an algorithm is as given below,

$$\text{Min } E(L) = \frac{L_1 + L_2 + L_3 + L_4}{E(T)}$$

$$\text{s. t.} \quad h \geq 0.01,$$

$$k \geq 0,$$

$$n \geq 2$$

Based on this optimisation function, a MATLAB program is written to find the optimal values of design parameters and loss. This program works with arbitrary initial values of the parameters.

6 An example

To illustrate the comparison between the normal and non-normal process as well between continued and ceased process, here we consider the example by Koo and Case (1990), which is used by Patil and Rattihalli (2009) with some modifications and also by Yu and Chen (2005). Following are the values of input parameters selected.

$$C_0 = 200,\ C = 4{,}000,\ C_2 = 1{,}500,\ V = 1{,}000,\ W = 1{,}000,\ a = b = 20,$$
$$\lambda = 0.02,\ \delta = 2,\ T_0 = T_1 = 1.25,\ T_2 = 2$$

6.1 Comparison between normal and non-normal (Burr) process

To investigate the accuracy of the program, here we consider the parameters of Burr distribution as $c_1 = 5$ and $k_1 = 6$, which have skewness parameter $\alpha_3 = -0.013$ and kurtosis parameter $\alpha_4 = 3.010$, which are very close to the normal distribution. The comparative values of the design parameters and the loss for Normal and Burr distributed process for different type of process design are given in Table 1.

Table 1 Loss cost for normal and Burr process for different type of process design

Process type	*Input distribution*	*n*	*h*	*k*	α	*loss*
Continuous	Normal	1.75	0.7263	2.3789	0.0174	620.0820
	Burr	1.75	0.7229	2.3896	0.0171	620.0097
Semi-ceased	Normal	1.75	0.6936	2.3924	0.0167	463.5946
	Burr	1.75	0.6903	2.4031	0.0164	463.5105
Ceased	Normal	1.8750	0.5846	2.7292	0.0063	386.4249
	Burr	1.8750	0.5827	2.7305	0.0062	386.0861

We observe that, almost all values of design parameters and expected loss for Normal and Burr distributed data are very close to each other. Moderate differences exist due to slight skewness and kurtosis in the burr data. It is also observed that, for both type of input data distributions, the loss for continuous process is larger as compared to ceased process. Also the type I error is large for continuous process and small for ceased process. The sampling interval is small and control limit parameter is large for ceased process as compared to continuous process.

6.2 Comparison of expected loss cost between continuous and ceased process designs for non-normal process

We consider some non-normal input data through coefficient of skewness (α_3) and Kurtosis (α_4). Let $\alpha_3 = 0.884$ and $\alpha_4 = 4.122$, corresponding to the parameters $c_1 = 2$ and $k_1 = 10$ of the Burr distribution with mean 0.29134 and Standard Deviation 0.16197. Let us consider, the two extreme procedures continued production and ceased production and compare the performance of the process for different values of 'n'. Table 2 gives findings.

Table 2 Comparison of loss between continuous and ceased process for different 'n'

n	*Continued process*			*Ceased process*			*% Saving in the loss*
	h	*k*	*loss*	*h*	*k*	*loss*	
2	1.3936	1.9356	587.0542	1.7820	1.9318	373.58	36.36
3	1.2318	1.9442	604.1059	1.5713	1.9378	395.52	34.53
4	1.1547	1.9509	614.1468	1.4708	1.9425	408.32	33.51
5	1.1005	1.9536	622.0929	1.4007	1.9443	418.37	32.75
6	1.0608	1.9559	628.4968	1.3492	1.9458	426.42	32.15
7	1.0289	1.9585	634.1562	1.3081	1.9476	433.48	31.64
8	1.0046	1.9599	638.7200	1.2768	1.9485	439.15	31.25
9	0.9837	1.9613	642.9034	1.2502	1.9493	444.32	30.89
10	0.9663	1.9624	646.5597	1.2281	1.9500	448.81	30.58

Table 2 shows that, the ceased process ensures near about 32% saving over the continuous process for a particular type of non-normal process data. Further, the loss cost increases with increase in 'n' for both the processes. The values of control limit

parameter (k) change in the direction of sample size (n) and that of sampling interval (h) changes in the opposite direction of 'n', this may cause increase in the loss for both the type of processes with increasing values of 'n'. The percentage saving due to ceasing of the process during search and repair is larger for smaller values of 'n' and smaller for larger values of 'n'. Thus, the changes in the values of sample size affects more on continuous process as compared to ceased process.

6.3 Effect of non-normality on the expected loss cost of the continuous and ceased process designs

To check the effect of non-normality on the design parameters and economy of the design we make the changes in the skewness (α_3) and Kurtosis (α_4) parameters of the input data and observe the changes in the output. We consider the three cases as follows.

1 α_4 is chosen near to 3 and α_3 is varied from –0.363 to 0.329

2 α_3 is chosen near to 0 and α_4 is varied from 2.866 to 3.646

3 α_3 is varied from –0.128 to 3.381 and α_4 is varied from 2.92 to 27.86.

Table 3 shows the outcomes for all the three cases for continued and ceased process. In the table c_1 and k_1 are the parameters of Burr distribution for corresponding values of skewness (α_3) and Kurtosis (α_4).

Following are some observations:

1 For case 1, the values of sampling interval (h) and of control limit parameter (k) slowly decrease and again increase for both the processes, for increase in skewness parameter.

2 For case 2 and case 3, values of sampling interval (h) increase with increase in the values of kurtosis parameter for continuous production. The values of control limit parameter (k) do not show any pattern for both the processes but it appears to be decreasing for continuous process.

3 The optimum sample size 'n' is fixed for continuous process but it changes between 2 and 3 for ceased process.

4 The last rows in case 3 show that, the process behaves abnormally for high increase in skewness and kurtosis. It is observed that for very high values of skewness and kurtosis the sampling interval and control limit multiplier reaches at extreme point.

5 For case 2 and case 3, the loss from both the processes increases with increase in values of kurtosis parameter.

6 The loss cost does not much affected for the change in skewness but is more affected by change in kurtosis. Hence, the process is much sensitive for change in kurtosis as compared to change in skewness. Also both the processes are more sensitive for extreme changes in skewness as well as kurtosis.

7 Both the processes (continuous and ceased) behave almost in similar fashion for change in skewness and kurtosis.

Table 3 Effect of non-normality on the design parameters and expected loss from the process

Case	α_3	α_4	c_1	k_1	*Continuous process*				*Ceased process*			
					n	*h*	*k*	*Loss*	*n*	*h*	*k*	*Loss*
	–0.363	3.143	7	8	2	0.6635	2.4925	626.495	3	0.4957	3.0549	391.1022
	–0.254	3.027	6	11	2	0.6529	2.5008	624.5557	3	0.4927	3.024	387.2004
	–0.147	3.065	6	6	2	0.6499	2.511	625.6379	3	0.4907	3.0344	387.5208
1	–0.013	3.01	5	6	2	0.6441	2.5021	624.9716	2	0.5554	2.777	387.1274
	0.04	3.07	5	5	2	0.6487	2.4942	625.871	2	0.5629	2.7667	389.2465
	0.136	2.979	4	7	2	0.5938	2.7804	614.155	2	0.5571	3.9966	353.0809
	0.238	3.154	4	5	2	0.6474	4.1243	600.0667	2	0.5804	3.8763	349.61
	0.329	3.006	3	11	2	1.0706	1.6877	693.0866	3	0.6282	3	337.2009
	0.05	2.866	4	11	2	0.5981	2.6889	614.9314	3	0.5095	3.9521	360.0047
	–0.051	2.975	5	7	2	0.6417	2.5058	624.3678	2	0.5515	2.782	385.8012
2	0.04	3.07	5	5	2	0.6487	2.4942	625.871	2	0.5629	2.7667	389.2465
	–0.019	3.169	6	4	2	0.6538	2.5072	627.2365	2	0.4935	3.0461	389.6302
	0.005	3.329	7	3	2	0.6648	2.4999	629.2016	3	0.4983	3.0683	393.5268
	0.044	3.646	10	2	2	0.6848	2.4826	632.2289	3	0.5063	3.1068	399.7197
	–0.128	2.92	5	11	2	0.6391	2.5082	623.2044	2	0.5469	2.7869	383.5494
	–0.098	2.938	5	9	2	0.6396	2.5082	623.6284	2	0.548	2.7859	384.3128
	–0.097	3.098	6	5	2	0.6506	2.5113	626.2446	3	0.4912	3.0394	388.1719
3	–0.019	3.169	6	4	2	0.6538	2.5072	627.2365	3	0.4935	3.0461	389.6302
	0.884	4.122	2	10	2	1.3936	1.9356	587.0542	2	1.782	1.9318	373.5834
	1.218	5.832	2	5	2	1.5308	1.7971	587.865	2	1.9949	1.7941	377.497
	1.909	12.46	2	3	2	1.7227	1.5955	586.0863	2	2.2865	1.5941	377.6384
	2.811	17.83	1	10	2	0.8509	1	512.4371	2	0.1	3	2307.4
	3.381	27.86	1	7	2	1.6127	0.9	567.0552	2	0.1	3	2303.1

Table 4 Effect of change in shift (δ) on the design parameters and expected loss from the process

δ	*Continued process*				*Ceased process*				*% Saving*
	n	*h*	*k*	*loss*	*n*	*h*	*k*	*loss*	
0.1	6	0.702	2.4431	908.5934	10	0.5141	2.1026	771.0529	15.1377
0.2	6	0.5218	2.1258	840.8781	7	0.5833	2.0391	686.1791	18.3973
0.3	5	0.5627	2.0675	793.1132	5	0.664	2.0188	629.5656	20.621
0.4	4	0.6161	2.0432	758.1593	4	0.7405	2.0043	587.7811	22.4726
0.5	3	0.6731	2.0356	731.2867	3	0.8162	2	555.5191	24.0354
0.6	3	0.7223	2.0107	708.9933	3	0.887	1.9834	527.9541	25.5347
0.7	3	0.7699	1.9936	691.6897	3	0.9532	1.9718	506.4796	26.7765
0.8	2	0.8346	2.0021	674.4798	2	1.0352	1.978	485.4187	28.0307
0.9	2	0.8891	1.9869	660.0417	2	1.11	1.9675	467.2062	29.2157
1	2	0.9426	1.9749	647.9481	2	1.1825	1.9592	451.8913	30.2581

6.4 Effect of change in shift (δ) on the design parameters and the expected loss cost of the continuous and ceased moving average process designs

To check the effect of change in shift (δ) on the parameters and the minimum loss cost of the design, we change the values of shift from 0.1 to 1 with increment of 0.1. We consider Burr process with parameters $c_1 = 2$ and $k_1 = 10$ corresponding to $\alpha_3 = 0.884$ and $\alpha_4 = 4.122$, we study the change in the values of design parameters (n, h, k) and minimum loss cost for continuous as well as for ceased processes. Table 4, gives the outcomes corresponding to the above settings.

From Table 4, we observe the following.

1 As the magnitude of the shift increases, the sample subgroup size (n) and control limit parameter (k) goes on decreasing, while sampling interval (h) goes on increasing for both the processes.

2 As the magnitude of the shift increases, the minimum loss cost decreases for both the processes and the percentage saving due to ceasing increases. That is, for larger shifts, ceased process is more economical than continuous process.

3 The values of sample size (n) and sampling interval (h) are larger, where as the values of control limit multiplier (k) are smaller, for ceased process as compared to the continuous process. This shows that ceased process works better than the continuous flow process for particular input values.

To check the sensitivity of the design for input parameters, in the following section, we fix the design parameters corresponding to the magnitude of the shift 0.4 and observe the change in the loss cost.

7 Sensitivity of the design

The illustrations in the above section of the example represents that, the ceased process seems better as compared to continuous process, for the given input parameters of the particular non-normal process quality characteristics. To check the sensitivity of the design with respect to the input parameters, we change one of the input parameter by 25, 50, 200 and 300 percent, by keeping other constant and observe the effect on the expected loss cost. In Table 5, we have chosen two optimum designs, one corresponds to optimum parameters of continuous process noted as Plan 1 and other for ceased process noted as Plan 2. Both the plans have design parameter (n, h, k) values corresponding to shift $\delta = 0.4$.

Table 5 Optimum design parameters corresponding to Plan 1 and Plan 2 used to check sensitivity of the design

δ	*Continued process (Plan 1)*				*Ceased process (Plan 2)*			
	n	*h*	*k*	*loss*	*n*	*h*	*k*	*loss*
0.4	4	0.6161	2.0432	758.159	4	0.7405	2.0043	587.781

Table 6 Effect on the expected loss from the continuous and ceased process due to change in the input parameter

Input parameter	*Value*	*LOSS*			*% Increase in loss*		*Saving by ceasing*	
		Cont (1)	*Ceas (2)*	*Ceas (1)*	*Cont (1)*	*Ceas (2)*	*Plan 1*	*Plan 1 and Plan 2*
C	1,000	459.96	448.53	484.461	–39.331	–23.69	–5.326	2.4864
	2,000	559.36	494.95	521.706	–26.221	–15.79	6.7321	11.516
	8,000	1155.8	773.45	745.177	52.4482	31.589	35.527	33.081
	12,000	1553.3	959.12	894.158	104.878	63.177	42.435	38.253
C_0	50	608.16	456.51	467.468	–19.785	–22.33	23.134	24.936
	100	658.16	500.26	510.378	–13.19	–14.89	22.454	23.99
	400	958.16	762.81	767.833	26.3797	29.779	19.864	20.388
	600	1158.2	937.85	939.471	52.7647	59.557	18.885	19.026
C_2	375	758.16	507.94	496.686	0	–13.58	34.488	33.003
	750	758.16	534.55	529.856	0	–9.056	30.113	29.493
	3,000	758.16	694.23	728.876	0	18.111	3.8624	8.4316
	4,500	758.16	800.69	861.556	0	36.222	–13.64	–5.6094
V	250	699.94	545.2	543.124	–7.6794	–7.244	22.404	22.107
	500	719.34	559.39	560.815	–5.1196	–4.83	22.038	22.236
	2,000	835.79	644.56	666.959	10.2392	9.6592	20.2	22.88
	3,000	913.42	701.33	737.721	20.4783	19.318	19.235	23.219
W	250	744.65	575.35	583.882	–1.7818	–2.115	21.59	22.736
	500	749.15	579.49	587.987	–1.1879	–1.41	21.513	22.647
	2,000	776.17	604.36	612.615	2.37575	2.8199	21.072	22.136
	3,000	794.18	620.93	629.034	4.75152	5.6398	20.795	21.815

Table 6 Effect on the expected loss from the continuous and ceased process due to change in the input parameter (continued)

Input parameter	*Value*	*LOSS*			*% Increase in loss*		*Saving by ceasing*	
		Cont (1)	*Ceas (2)*	*Ceas (1)*	*Cont (1)*	*Ceas (2)*	*Plan 1*	*Plan 1 and Plan 2*
a and b	5	709.47	552.33	554.408	–6.4226	–6.032	21.856	22.149
	10	725.7	564.14	568.338	–4.2817	–4.021	21.684	22.262
	40	823.08	635.06	651.914	8.56344	8.0428	20.796	22.844
	60	888.01	682.33	707.631	17.1269	16.086	20.313	23.162
λ	0.005	461.53	462.89	499.477	–39.125	–21.25	–8.223	–0.2959
	0.01	565.61	506.77	533.271	–25.397	–13.78	5.7179	10.403
	0.04	1090.7	727.13	706.117	43.8616	23.708	35.26	33.333
	0.06	1367.8	842.63	798.881	80.4106	43.357	41.594	38.395
T_0	0.313	758.16	536.53	532.015	0	–8.72	29.828	29.233
	0.625	758.16	554.22	554.375	0	–5.71	26.879	26.899
	2.5	758.16	648.23	669.644	0	10.284	11.675	14.499
	3.75	758.16	701.17	732.051	0	19.29	3.4436	7.5174
T_1	0.313	697.96	597.05	605.512	–7.9404	1.5776	13.245	14.457
	0.625	718.23	593.93	602.378	–5.2663	1.0468	16.131	17.306
	2.5	835.38	575.85	584.206	10.1847	–2.03	30.067	31.067
	3.75	909.26	564.39	572.689	19.9306	–3.979	37.016	37.929
T_2	0.5	660.78	602.77	611.25	–12.844	2.5496	7.496	8.7798
	1	693.84	597.69	606.149	–8.484	1.6854	12.638	13.858
	4	880.09	568.92	577.241	16.0831	–3.209	34.411	35.357
	6	993.84	551.23	559.454	31.0853	–6.218	43.708	44.535

Table 6 gives the sensitivity results obtained for continuous and ceased process according to design parameters in Table 5 and by varying the input parameter C = 4000; C_0 = 200; C_2 = 1500;V = 1000;W = 1000;a = 20;b = 20;λ = 0.02; g = 0; T_0 = 1.25; T_1 = 1.25; T_2 = 2. In Table 6, Cont (1), denotes results using continuous process and optimum design parameters of Plan 1. Ceas (1), denotes results using ceased process and optimum design parameters of Plan 1. Similarly, Ceas (2), denotes results using ceased process and optimum design parameters of Plan 2. In the column saving by ceasing, Plan 1, we means, percentage saving due to ceasing using Plan 1 parameters for both the process and Plan 1 and Plan 2, mean percentage saving due to ceasing using Plan 1 parameters for continuous process and Plan 2 parameters for ceased process.

For a quick review of the sensitivity of the design, we may refer to Figures 1 to 4. Figures 1 and 2, reveal percentage increase in loss with respect to standard value (according Plan 1 or Plan 2) of the loss due to change in input values for continuous and ceased process respectively. Figures 3 and 4, show effect on percentage saving by ceasing of process due to change in input parameters respectively by using unique Plan for both and their own optimal plans.

Figure 1 Increase in loss due to change in input parameters for continuous process

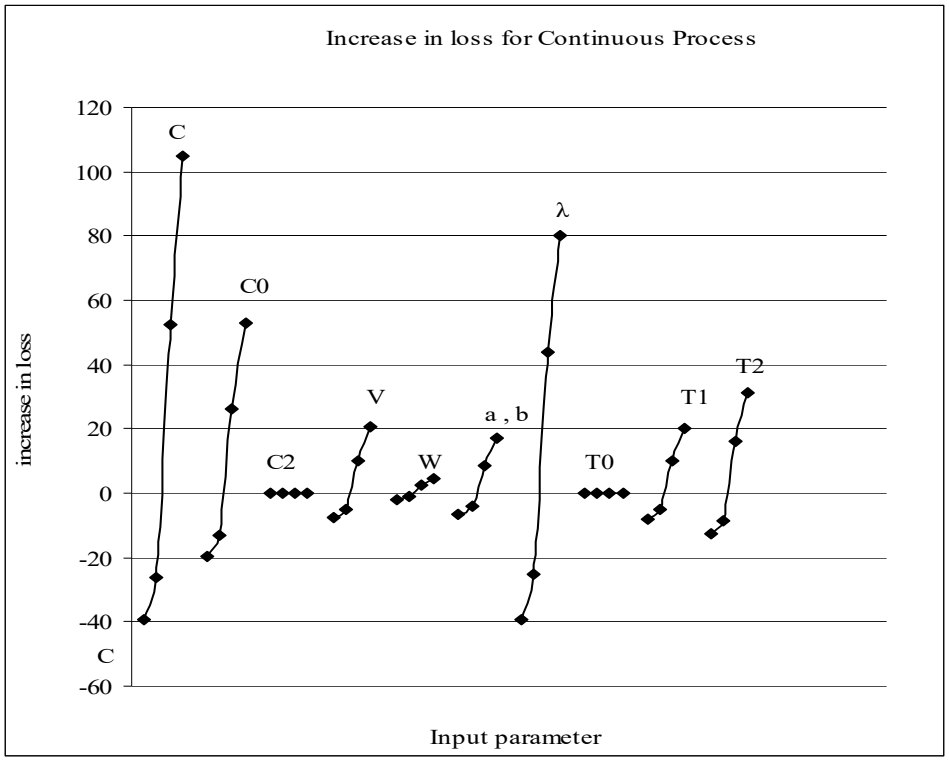


Figure 2 Increase in loss due to change in input parameters for ceased process

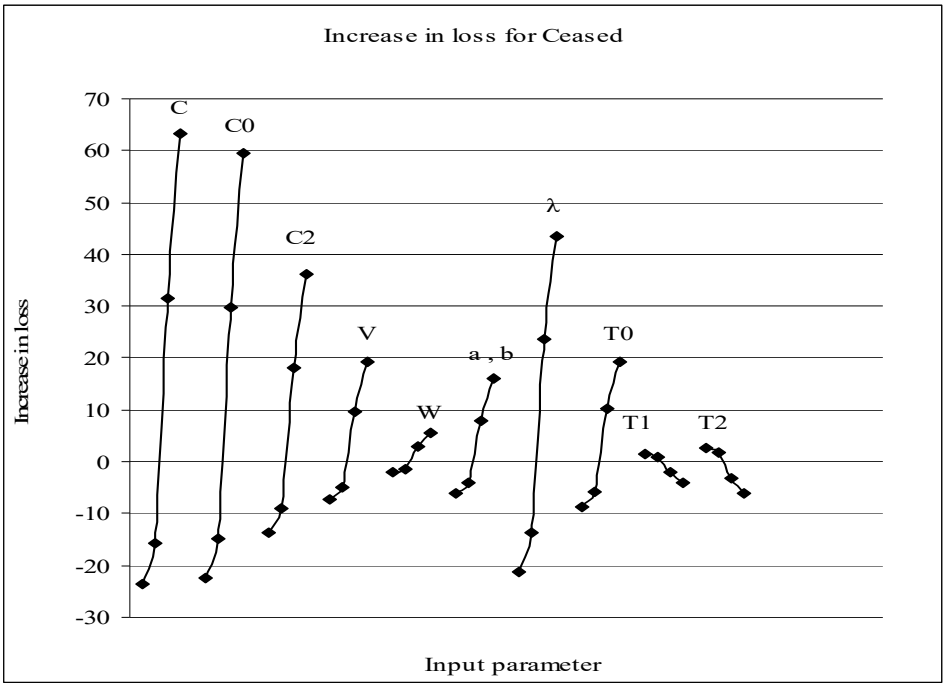


Figure 3 Percentage saving in the loss due to ceased process using Plan 1 parameters

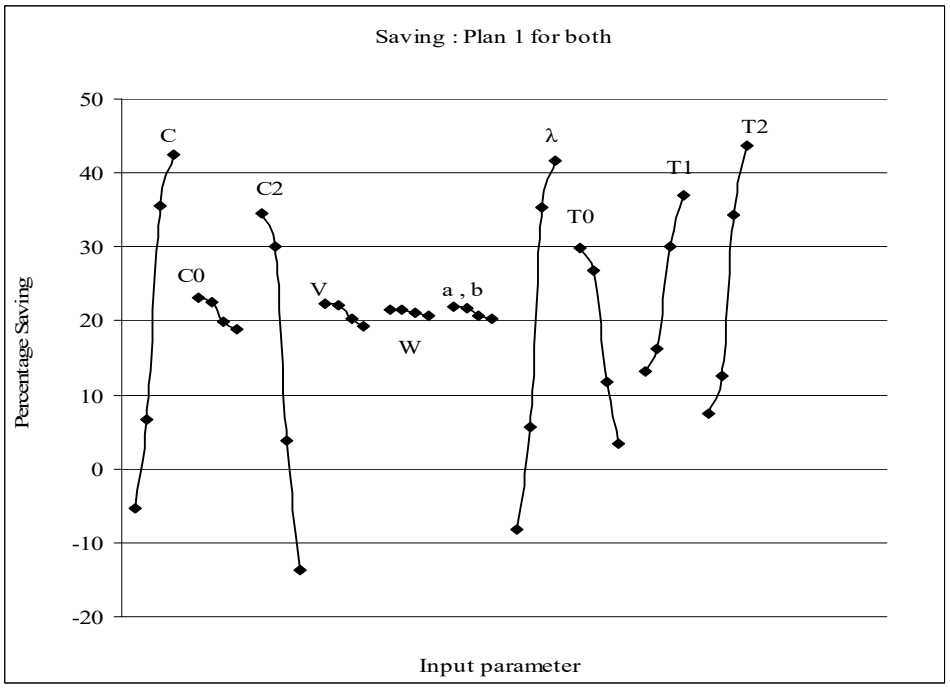
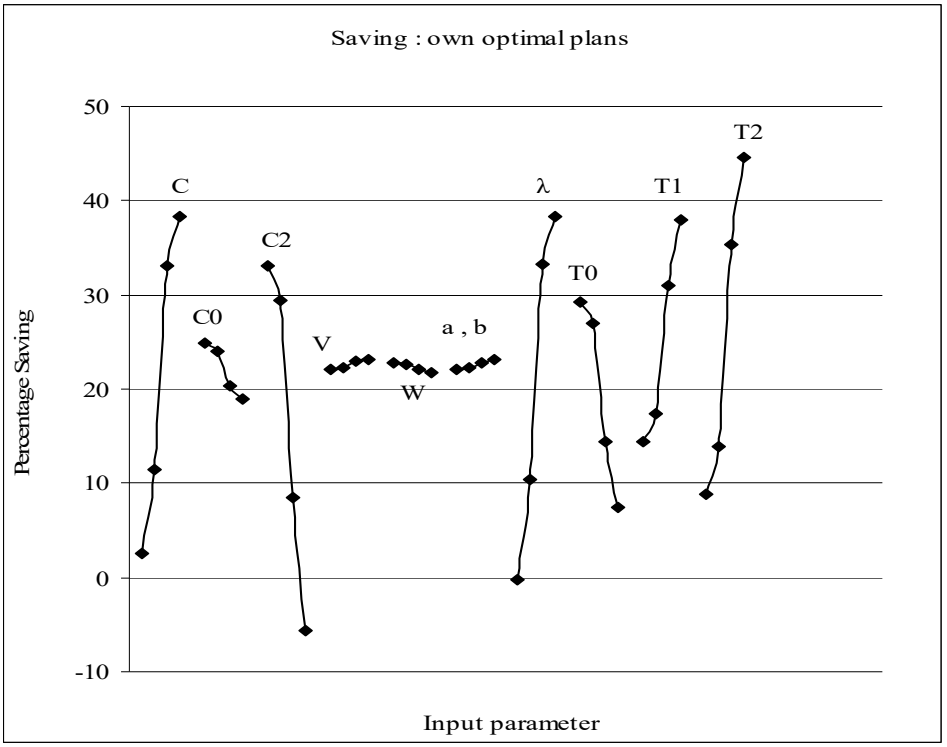


Figure 4 Percentage saving in the loss due to ceased process using individual optimal plan parameters



From Table 5 and Figures 1 to 4, we have following observations.

1 The continuous as well as ceased process is too much sensitive to the change in the parameters C, C_0 and λ. The parameters C_2 and T_0 are related to only ceased process and shows considerable effect on the expected loss from the process.

2 The parameter V has significant effect on both the process. The continuous process is sensitive with respect to T_1 and T_2, where as ceased process seems to be less sensitive. Both the processes are less sensitive to W.

3 On an average, the continuous process is much sensitive as compared to ceased one with respect to changes in all the factors.

4 There is significant effect on saving occurred by ceased process due to change in C, C_2, λ, T0, T_1 and T_2. Change in the values of C_0, V, W, a and b does not show much effect on saving due to ceasing.

8 Conclusions

In this paper, we have obtained an expression for expected loss cost function to control location for continuous and ceased moving average production processes depending on the non-normal input quality characteristic. The approximation based on Burr distribution is used to monitor with non-normality of quality characteristic. The cost function is

optimised with respect to the three design parameters n, h and k. The performance of two production modes is studied in different views and also the effect of non-normality is observed. Moreover, this article provides SPC practitioners the way to improve the process in economic view. The design is as simple as $\bar{x}$ chart and will promote the application of this effective chart. Mostly, the producers prefers for continuation of the process during the search and repair of the process. This article provides the benefit of ceasing of the process to such practitioners.

There is comparable difference in the loss due to continuous, ceased and semi-ceased process. The design parameters h and k are affected by non-normality also the loss cost is mostly affected by change in kurtosis. As the sample size 'n' increases, percentage saving due to ceasing decreases. It shows that, for very larger sample size both the process models become equally efficient. For the smaller values of shift (δ), the process model shows greater loss and also causes increase in percentage saving due to ceasing with increasing δ. That is, ceased process designs are more beneficial for larger shifts. Overall the ceased process appears to be dominant over continuous process and shows almost 20 to 30% saving in the loss cost. The process model is highly sensitive to C, C0, λ which are the major parameters of the process and is less sensitive to V, T1, T2 and W. The parameters C2 and T0 related to ceased production seems to be sensitive.

The novelty of the study is, if the production process is affected by small shifts, the ceasing of process during the repair will be advisable. At the same time, if we use MA procedure instead of usual mean chart, we are using past sample observations, which may turn to be long run control of the process, which will be beneficial to the producers. Further, we have used input parameters from the real life data. There is scope for investigators to estimate the input parameters from original sampled data. The study can be extended further with the use of CUSUM or EWMA charts with the non-normal input data.

Acknowledgements

The authors are grateful to the editor and both the referees for their valuable suggestions and comments. Second author would like to acknowledge the support from the Department of Science and Technology, New Delhi under the DST-PURSE scheme to carry out this work.

References

Alkhedher, M.J. and Darwish, M.A. (2013) 'Optimal process mean for a stochastic inventory model under service level constraint', *Int. J. Operational Research*, Vol. 18, No. 3, pp.346–363.

Al-Oraini, H.A. and Rahim, M.A. (2003) 'Economic statistical design of $\overline{X}$ charts for system with gamma (λ, 2) in control times', *Journal of Applied Statistics*, Vol. 30, No. 4, pp.397–409.

Amiri, A., Mogouie, H. and Doroudyan, M.H. (2013) 'Multi-objective economic-statistical design of MEWMA control chart', *Int. J. Productivity and Quality Management*, Vol. 11, No. 2, pp.131–149.

Burr, I.W. (1942) 'Cumulative frequency functions', Presented in the *Meeting of Institute of Mathematical Statistics*, *American Mathematical Society and Econometric Society* at Chieago III, September 2.

Chen, F.L. and Yeh, C.H. (2006) 'Economic design of control charts with Burr distribution for non-normal data, under Weibull failure mechanism', *Journal of the Chinese Institute of Industrial Engineering*, Vol. 23, No. 3, pp.200–206.

Chen, H and Cheng, Y. (2007) 'Non-normality effects on the economic-statistical design of $\overline{X}$ charts with Weibull in-control time', *European Journal of Operational Research*, Vol. 176, No. 2, pp.986–998.

Chen, Y.K. (2004) 'Economic design of $\overline{X}$ charts for non-normal data using variable sampling policy', *Int. Journal of Production Economics*, Vol. 92, No. 1, pp.61–74.

Chen, Y.S. and Yang, Y.M. (2002) 'An extension of Banerjee and Rahim's model for economic design of moving average control chart for a continuous flow process', *European Journal of Operational Research*, Vol. 143, No. 3, pp.600–610.

Chen, Y.S. and Yu, F.J. (2003) 'Determination of optimal design parameters of moving average control charts', *Int. Journal of Advanced Manufacturing Technology*, Vol. 21, No. 6, pp.397–402.

Chou, C.Y., Chen, C.H. and Liu, H.R. (2000) 'Economic-statistical design of $\overline{X}$ charts for non-normal data by considering quality loss', *Journal of Applied Statistics*, Vol. 27, No. 8, pp.939–951.

Chou, C-Y. Li, M-H.C. and Wang, P-H. (2001) 'Economic statistical design of averages control charts for monitoring a process under non-normality', *Int. Journal of Advanced Manufacturing Technology*, Vol. 17, No. 8, pp.603–609.

Duncan, A.J. (1956) 'The economic design of $\overline{X}$ charts to maintain current control of a process', *Journal of American Statistical Association*, Vol. 51, No. 274, pp.228–242.

Franco, B.C., Celano, G., Castagliola, P. and Costa, A. F.B. (2014) 'Economic design of Shewhart control charts for monitoring autocorrelated data with skip sampling strategies', *Int. J. Production Economics*, Vol. 151, pp.121–130.

Guo, Z-F, Cheng, L-S and Luc, Z-D (2014) 'Economic design of the variable parameters $\overline{X}$ control chart with a corrected A&L switching rule', *Quality and Reliability Engineering Int.*, Vol. 30, No. 2, pp.235–246.

Hashemi, M., Shahmorad-Moghanloo, S. and Behboudi, D. (2014) 'Review the allocation of production lines in shifts with minimising energy costs approach in Tehran Pegah Co.', *Int. J. Operational Research*, Vol. 19, No. 1, pp.68–77.

Khoo M.B.C., Yeong, W.C., Lee, M.H. and Rahim, M.A. (2013) 'Economic and economic statistical designs of the synthetic X chart using loss functions', *European Journal of Operational* Research, Vol. 228, No. 3, pp.571–581.

Koo, T.Y. and Case, C.E. (1990) 'Economic design of $\overline{X}$ control charts for use in monitoring continuous flow process', *Int. Journal of Production Research*, Vol. 28, No. 11, pp.2001–2011.

Kooli, I. and Limam, M. (2011) 'Economic design of an attribute np control chart using a variable sample size', *Sequential Analysis*, Vol. 30, No. 2, pp.145–159.

Lee, P-H., Chang, Y-C. and Torng, C-C. (2012) 'A design of s control charts with a combined double sampling and variable sampling interval scheme', *Communications in Statistics – Theory and Methods*, Vol. 41, No. 1, pp.153–165.

Lorenzen, T.J. and Vance, L.C. (1986) 'The economic design of control charts: a unified approach', *Technometrics*, Vol. 28, No. 1, pp.3–10.

Mahadik, S.B. and Shirke, D.T. (2011) 'A special variable sample size and sampling interval Hotelling's T2 chart', *Int. Journal of Advanced Manufacturing Technology*, Vol. 53, Nos. 1–4, pp.379–384.

Mc Williams, T.P. (1989) 'Economic control chart designs and the in-control time distribution: a sensitivity study', *Journal of Quality Technology*, Vol. 21, No. 2, pp.103–110.

Montgomery, D.C. (1980) 'The economic design of control charts: a review and literature survey', *Journal of Quality Technology*, Vol. 12, No. 2, pp.75–87.

Montgomery, D.C. (2001) *Introduction to Statistical Quality Control*, 4th ed., John Wiley and Sons, New York.

George N. (2011) 'A new approach for the economic design of fully adaptive control charts', *Int. J. Production Economics*, Vol. 131, No. 2, pp.631–642.

Ou, Y., Wu Zhang, Lee, K.M. and Chen, S.L. (2012) 'An optimal design algorithm of the SPRT chart for minimizing weighted ATS', *Int. J. Production Economics*, Vol. 139, No. 2, pp.564–574.

Patil, S.H. and Rattihalli, R.N. (2009) 'Economic design of moving average control chart for continued and ceased production process', *Economic Quality Control*, Vol. 24, No. 1, pp.391–397.

Rahim, M.A. (1985) 'Economic model of $\overline{X}$ chart under non-normality and measurement errors', *Computer and Operations Research*, Vol. 12, No. 3, pp.291–299.

Rahim, M.A. and Banerjee, P.K. (1993) 'A generalized model for the economic design of $\overline{X}$ control charts for production system with increasing failure rate and early replacement', *Naval Research Logistics*, Vol. 40, No. 6, pp.787–809.

Rostami, R. and Ebrahimnejad, A. (2014) 'An approximation algorithm for discrete minimum cost flows over time problem', *Int. J. Operational Research*, Vol. 20, No. 2, pp.226–239.

Saniga, E.M. (1989) 'Economic statistical control-chart designs with an application to $\overline{X}$ and R charts', *Technometrics*, Vol. 31, No. 3, pp.313–320.

Shewhart, W.A. (1931) *Economic Control of Quality of Manufactured Product*, p.501, D. Van Nostrand Company, New York.

Singh, K.P., Kansal, M.L. and Deep, K. (2014) 'GA-NR for optimal design of water distribution networks', *Int. J. Operational Research*, Vol. 20, No. 3, pp.241–261.

Trovato, E., Castagliola, P., Celano, G. and Fichera, S. (2010) 'Economic design of inspection strategies to monitor dispersion in short production runs', *Computers & Industrial Engineering*, Vol. 59, No. 4, pp.887–897.

Tu, Y., Liu, L., Yu, M. and Ma, Y. (2013) 'Economic and economic-statistical designs of an X control chart for two-unit series systems with condition-based maintenance', *European Journal of Operational* Research, Vol. 226, No. 3, pp.491–499.

Collani, V. (1986) 'A simple procedure to determine the economic design of an $\overline{X}$ control chart', *Journal of Quality Technology*, Vol. 18, No. 3, pp.145–151.

Woodall, W.H., Lorenzen, T.J. and Vance, L.C. (1986) 'Weaknesses of the economic design of control charts', *Technometrics*, Vol. 28, No. 4, pp.408–410.

Zhang, W., Yang, M., Jiang, W. and Khoo M.B.C. (2008) 'Optimization designs of the combined Shewhart-CUSUM control charts', *Computational Statistics and Data Analysis*, Vol. 53, No. 2, pp.496–506.

Zhang, W., Yang, M., Khoo M.B.C. and Yu, F.J. (2010) 'Optimization designs and performance comparison of two CUSUM schemes for monitoring process shifts in mean and variance', *European Journal of Operational Research*, Vol. 205, No. 1, pp.136–150.

Zhang, W., Yang, M., Khoo, M.B.C. and Castagliola, P. (2011) 'What are the best sample sizes for the Xbar and CUSUM charts?', *Int. J. Production Economics*, Vol. 131, No. 2, pp.650–662.

Yeh, L.L. and Chen, F.L. (2010) 'An extension of Banerjee and Rahim's model for an economic design of X-bar control chart for non-normally distributed data, under gamma failure model', *Communication in Statistics – Simulation and Computation*, Vol. 39, No. 5, pp.994–1015.

Yu, F.J. and Chen, Y.S. (2005) 'Economic design of moving average control charts', *Journal of Quality Engineering*, Vol. 17, No. 3, pp.391–397.

Yu, F.J. and Wu, H.H. (2004) 'An economic design for variable sampling interval MA control charts', *Int. Journal of Advanced Manufacturing Technology*, Vol. 24, Nos. 1–2, pp.41–47.

Zhang, G and Berardi, V. (1997) 'Economic statistical design of $\overline{X}$ control charts for system with Weibull in control times', *Computers and Industrial Engineering*, Vol. 32, No. 3, pp.575–586.

Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 19

5-1-2017

Confidence Intervals for the Scaled Half-Logistic Distribution under Progressive Type-II Censoring

Kiran Ganpati Potdar
Department of Statistics, Ajara Mahavidyalaya, Ajara, potdarkiran.stat@gmail.com

D. T. Shirke
Department of Statistics, Shivaji University, Kolhapur, dts_stats@unishivaji.ac.in

Recommended Citation
Potdar, K. G. & Shirke, D. T. (2017). Confidence intervals for the scaled half-logistic distribution under progressive Type-II censoring. Journal of Modern Applied Statistical Methods, 16(1), 324-349. doi: 10.22237/jmasm/1493597880

Confidence Intervals for the Scaled Half-Logistic Distribution under Progressive Type-II Censoring

Cover Page Footnote
The first author wishes to thank University Grants Commission, New Delhi, India for providing Fellowship under Faculty Improvement Programme to carry out this research.

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 324-349.
doi: 10.22237/jmasm/1493597880

ISSN 1538 − 9472

Confidence Intervals for the Scaled Half-Logistic Distribution under Progressive Type-II Censoring

Kiran G. Potdar
Ajara Mahavidyalaya
Ajara, India

D. T. Shirke
Shivaji University
Kolhapur, India

Confidence interval construction for the scale parameter of the half-logistic distribution is considered using four different methods. The first two are based on the asymptotic distribution of the maximum likelihood estimator (MLE) and log-transformed MLE. The last two are based on pivotal quantity and generalized pivotal quantity, respectively. The MLE for the scale parameter is obtained using the expectation-maximization (EM) algorithm. Performances are compared with the confidence intervals proposed by Balakrishnan and Asgharzadeh via coverage probabilities, length, and coverage-to-length ratio. Simulation results support the efficacy of the proposed approach.

Keywords: Progressively Type-II censoring, EM algorithm, MLE, pivotal quantity, confidence interval, generalized confidence interval, coverage probability, coverage to length ratio, half-logistic distribution

Introduction

In many life testing situations, an experiment has to be terminated before completion. Because of the various limitations of time and money, testing of life may need to be stopped for some of the units. In day-to-day experiments, incomplete information about the failure times is available, or some of the units must be removed before completion of the experiment. A plan is necessary for removal of the units before the termination of an experiment to save time and cost, which is called the censored data.

Type-I censoring depends on time, where the time is fixed for the termination of experiment. Suppose an observer continues an experiment up to time T; lifetimes of units will be known exactly only if these are less than T.

Dr. Potdar is an Assistant Professor in the Department of Statistics. Email them at: potdarkiran.stat@gmail.com. Dr. Shirke is a Professor in the Department of Statistics. Email them at: dts_stats@unishivaji.ac.in.

Failure times of units which have not failed by the time T are not observed. Suppose n units are being tested, but the decision is made to terminate the experiment at time T. In this experiment, lifetimes will be known exactly only for those units that fail before time T. In Type-I censoring, the number of exact lifetimes observed is random.

A Type-II censoring scheme is often used in life testing experiments where the number of units that can be observed before the termination of the experiment is fixed. In this scheme, only a pre-planned number m out of n units ($m < n$) are observed. In the case of Type-II censoring, the number of exact lifetimes observed is fixed, but the time required for the termination of the experiment is unknown. In conventional Type-I and Type-II censoring, units are removed from the experiment at the terminal stage, while in a progressive censoring scheme, units are removed at different stages. Progressive censoring schemes can be applied in both Type-I and Type-II censoring schemes. More details about various censoring schemes are available in Lawless (1982).

In an $(R_1, R_2,\ldots, R_m)$ progressive type-II censoring scheme, the number m and $R_1, R_2,\ldots, R_m$ are fixed before the start of the experiment and $\sum_{i=1}^{m} R_i = n-m$. At the first failure, R_1 units are randomly removed from the remaining $n-1$ units. At the second failure, R_2 units are randomly removed from the remaining $n-2-R_1$ units, etc. At the m^{th} failure, all the remaining R_m units are removed. Here, we observe failure times of m units and the remaining $n-m$ units are removed at different stages of the experiment. In a conventional Type-II censoring scheme, $R_m = n - m$ and the rest of the R_i are zero.

Consider the problem of interval estimation for the scale parameter of a half-logistic distribution under a progressive Type-II censoring scheme. Progressive Type-II censoring schemes for various lifetime distributions was discussed by Cohen (1963), who introduced progressive Type-II censoring schemes. Mann (1969, 1971), Balakrishnan, Kannan, Lin, and Ng (2003), Balakrishnan, Kannan, Lin, and Wu (2004), Ng (2005), and Ng, Kundu, and Balakrishnan (2006) discussed inference for different lifetime distributions under progressive Type-II censoring schemes. Balakrishnan and Aggarwala (2000) is an excellent reference on progressive censoring. Balakrishnan (2007) studied various distributions and inferential methods for the progressively censored data. Lin and Balakrishnan (2011) discussed the consistency and the asymptotic normality of Maximum Likelihood Estimators (MLEs) based on the progressive Type-II censored samples. Potdar and Shirke (2013, 2014) studied inference for the scale parameter of the half logistic and Rayleigh distribution of k-unit parallel systems

based on progressively Type-II censored data. Ghitany, Alqallaf, and Balakrishnan (2014) discussed estimation of the parameters of Gompertz distributions based on progressively Type-II censored samples. Sultan, Alsadat, and Kundu (2014) studied estimation for the inverse Weibull parameters under progressive Type-II censoring.

As far as the half-logistic distribution is concerned, Balakrishnan and Puthenpura (1986) discussed the best linear unbiased estimation of location and scale parameters. Balakrishnan and Wong (1991) computed the approximate Maximum Likelihood Estimator (AMLE) for the location and scale parameters of the half-logistic distribution. Balakrishnan and Chan (1992) studied estimation for the scale parameter of the half-logistic distribution. Kim and Han (2010) used importance sampling methods to obtain a Bayes estimator for the scale parameter of the half-logistic distribution under progressively Type-II censored samples. Jang, Park, and Kim (2011) studied estimation of the scale parameter of the half-logistic distribution with a multiply Type-II censored sample. Rastogi and Tripathi (2014) studied estimation of parameter and reliability for the exponentiated half-logistic distribution.

The likelihood equation of a half-logistic distribution with scale parameter does not have a closed form solution to obtain MLE. In most of the reported work, an AMLE of the scale parameter is obtained. Following this approach, Balakrishnan and Asgharzadeh (2005) and Wang (2009) reported inference for the scale parameter of a half-logistic distribution based on progressive Type-II censored samples.

Balakrishnan and Asgharzadeh (2005) showed that, if the relative sample fraction is small, then the coverage probability of the confidence interval (CI) based on asymptotic normality of the MLE is unsatisfactory. Wang (2009) paid more attention to length of CI and gave a shorter length CI. Dempster, Laird, and Rubin (1977) introduced the expectation-maximization (EM) algorithm to obtain the MLE for the incomplete data. McLachlan and Krishnan (1997) gave more details about the EM algorithm. Here, the MLE is computed using the EM algorithm, and the focus is on both the coverage probability and length of CI.

Assume that n units having half-logistic lifetime distribution are put on test and failure times of $\sum_{i=1}^{m} R_i = n - m$ units are censored. Lifetimes of these censored units are unknown. Consider the censored data as missing data and use the EM algorithm to compute the MLE. As indicated in Potdar and Shirke (2014), the EM algorithm gives improved inferential results.

Model and Estimation of the Scale Parameter

Suppose progressively Type-II censored data are obtained from the scaled half-logistic distribution with probability density function

$$\mathrm{f}\left(x;\lambda\right)=\frac{1}{\lambda}\frac{2\mathrm{e}^{-x/\lambda}}{\left(1+\mathrm{e}^{-x/\lambda}\right)^{2}},\quad x\geq 0,\lambda>0 \tag{1}$$

and cumulative distribution function

$$\mathrm{F}\left(x;\lambda\right)=\left[\frac{1-\mathrm{e}^{-x/\lambda}}{1+\mathrm{e}^{-x/\lambda}}\right],\quad x\geq 0,\lambda>0$$

Suppose n units are under test and lifetimes of m units are observed under progressive Type-II censoring. Suppose (R_1, R_2,…, R_m), a progressive censoring scheme, is used. The observed lifetimes $x_{(1)}$, $x_{(2)}$,…, $x_{(m)}$ are the progressively Type-II censored sample. The likelihood function for the observed data is given by (Balakrishnan & Aggarwala, 2000)

$$\mathrm{L}\left(\lambda\right)=C\prod_{i=1}^{m}\mathrm{f}\left(x_{(i)};\lambda\right)\left[1-\mathrm{F}\left(x_{(i)};\lambda\right)\right]^{R_i} \tag{2}$$

where

$$C=n\prod_{j=1}^{m-1}\left(n-j-\sum_{i=1}^{j}R_i\right)$$

Maximum Likelihood Estimation

Suppose $\mathbf{z}_1$, $\mathbf{z}_2$,…,$\mathbf{z}_m$ are the censored data. Note $\mathbf{z}_i$ is a vector with R_i element corresponding to R_i removed units after the i^{th} failure is observed (i = 1, 2,…., m). The censored data $\mathbf{Z}$ = ($\mathbf{z}_1$, $\mathbf{z}_2$,…, $\mathbf{z}_m$) can be considered to be the missing data and $\mathbf{X}$ = ($x_{(1)}$, $x_{(2)}$,…, $x_{(m)}$) the observed data. $\mathbf{W}$ = ($\mathbf{X}$, $\mathbf{Z}$) is the complete data set to be used for drawing inference for the scale parameter. The complete log-likelihood function can be written as

$$L_c = -n\log(\lambda) + \sum_{i=1}^{m}\log\left(\frac{2e^{-x_i/\lambda}}{\left(1+e^{-x_i/\lambda}\right)^2}\right) + \sum_{i=1}^{m}\sum_{j=1}^{R_i}\log\left(\frac{2e^{-z_{ij}/\lambda}}{\left(1+e^{-z_{ij}/\lambda}\right)^2}\middle| z_{ij} > x_i\right) \quad (3)$$

By differentiating L_c with respect to λ,

$$\frac{dL_c}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{m}\frac{x_i\left(1-e^{-x_i/\lambda}\right)}{1+e^{-x_i/\lambda}} + \frac{1}{\lambda^2}\sum_{i=1}^{m}\sum_{j=1}^{R_i}\left(\frac{z_{ij}\left(1-e^{-z_{ij}/\lambda}\right)}{1+e^{-z_{ij}/\lambda}}\middle| z_{ij} > x_i\right)$$

The EM algorithm suggested by Dempster et al. (1977) was used to compute the MLE. For the E step in the EM algorithm, the expectation of Z_{ij} was taken. Hence, the above equation becomes

$$\frac{dL_c}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{m}\frac{x_i\left(1-e^{-x_i/\lambda}\right)}{1+e^{-x_i/\lambda}} + \frac{1}{\lambda^2}\sum_{i=1}^{m}R_i\,a(x_i,\lambda) \quad (4)$$

where

$$a(x_i,\lambda) = E\left(\frac{z_{ij}\left(1-e^{-z_{ij}/\lambda}\right)}{1+e^{-z_{ij}/\lambda}}\middle| z_{ij} > x_i\right) = \int_{x_i}^{\infty}\frac{z\left(1-e^{-z_{ij}/\lambda}\right)}{1+e^{-z_{ij}/\lambda}}\frac{f(z)}{1-F(x_i)}dz$$

$$= \frac{1+e^{-x_i/\lambda}}{\lambda e^{-x_i/\lambda}}\int_{x_i}^{\infty}\frac{ze^{-z/\lambda}\left(1-e^{-z/\lambda}\right)}{\left(1+e^{-z/\lambda}\right)^3}dz$$

Solving equation (4) is the M step.

The Newton-Raphson method was used to solve equation (4) by taking the least square estimate as an initial value. Ng (2005) discussed estimation of model parameters of modified Weibull distributions based on progressively Type-II censored data, where the empirical distribution function is computed as

$$\hat{F}\left(x_{(i)}\right) = 1 - \prod_{j=1}^{i}\left(1-\hat{p}_j\right)$$

with

$$\hat{p}_j = \frac{1}{n - p_j^*},\ p_j^* = \sum_{k=2}^{j} R_{k-1} - j + 1, \quad j = 1, 2, \ldots, m$$

The estimate of the parameters can be obtained by the least squares fit of simple linear regression

$$y_i = \beta x_{(i)}$$

with $\beta = -1/\lambda$,

$$y_i = \ln\left[\frac{1 - \frac{\hat{\mathrm{F}}\left(x_{(i-1)}\right) + \hat{\mathrm{F}}\left(x_{(i)}\right)}{2}}{1 + \frac{\hat{\mathrm{F}}\left(x_{(i-1)}\right) + \hat{\mathrm{F}}\left(x_{(i)}\right)}{2}}\right], \quad i = 1, 2, \ldots, m$$

$$\hat{\mathrm{F}}\left(x_{(0)}\right) = 0$$

The least square estimate of λ is given by

$$\hat{\lambda}_0 = -\frac{\sum_{i=1}^{m} x_{(i)}^2}{\sum_{i=1}^{m} x_{(i)} y_i} \tag{5}$$

While obtaining the MLE $\hat{\lambda}_n$ of the scale parameter λ, the above approach was adopted, where $\hat{\lambda}_0$ was taken as an initial value of λ in the Newton-Raphson method. It will be shown that the MLE $\hat{\lambda}_n$ exits and is unique. From equation (2),

$$\mathrm{L}(\lambda) = C \prod_{i=1}^{m} \frac{2^{R_i+1}}{\lambda} \frac{\mathrm{e}^{-(R_i+1)x_i/\lambda}}{\left(1 - \mathrm{e}^{-x_i/\lambda}\right)^{R_i+2}}$$

where C is defined as above.

$$\frac{d\log \mathrm{L}}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{m}\left(R_i+1\right)x_i - \frac{1}{\lambda^2}\sum_{i=1}^{m}\frac{\left(R_i+2\right)x_i \mathrm{e}^{-x_i/\lambda}}{1+x_i \mathrm{e}^{-x_i/\lambda}} = 0 \tag{6}$$

Define

$$\mathrm{g}\left(\lambda\right) = \frac{-\lambda^2 d\log \mathrm{L}}{d\lambda} = \frac{d\log \mathrm{L}}{d\lambda} = n\lambda - \sum_{i=1}^{m}\left(R_i+1\right)x_i + \sum_{i=1}^{m}\frac{\left(R_i+2\right)x_i \mathrm{e}^{-x_i/\lambda}}{1+x_i \mathrm{e}^{-x_i/\lambda}} = 0$$

Note

$$\lim_{\lambda\to 0} g\left(\lambda\right) < 0, \lim_{\lambda\to\infty} g\left(\lambda\right) > 0, \text{ and } \mathrm{g}'\left(\lambda\right) > 0$$

Therefore, the MLE, a solution to g(λ) = 0, exists and is unique.

Fisher Information

We compute observed Fisher information using the idea of the missing information principle of Louis (1982). Thus, observed information = complete information – missing information. Write this as

$$\mathrm{I}_x\left(\lambda\right) = \mathrm{I}_w\left(\lambda\right) - \mathrm{I}_{w|x}\left(\lambda\right) \tag{7}$$

In the following, we obtain complete and missing information given by

$$\mathrm{I}_w\left(\lambda\right) = -\mathrm{E}\left[\frac{d^2 \mathrm{L}}{d\lambda^2}\right]$$

where, L is the log-likelihood function of the complete data. By differentiating L with respect to λ twice

$$\frac{d^2 \mathrm{L}}{d\lambda^2} = \frac{n}{\lambda^2} - \frac{2}{\lambda^4}\sum_{i=1}^{n}\frac{x_i^2 \mathrm{e}^{-x_i/\lambda}}{\left(1+\mathrm{e}^{-x_i/\lambda}\right)^2} - \frac{2}{\lambda^3}\sum_{i=1}^{n}\frac{x_i\left(1-\mathrm{e}^{-x_i/\lambda}\right)}{1+x_i \mathrm{e}^{-x_i/\lambda}}$$

The complete information is given by

$$\mathrm{I}_w(\lambda) = -\frac{n}{\lambda^2} + \frac{2}{\lambda^4}\sum_{i=1}^{n}\mathrm{E}\left[\frac{X_i^2 \mathrm{e}^{-X_i/\lambda}}{\left(1+\mathrm{e}^{-X_i/\lambda}\right)^2}\right] + \frac{2}{\lambda^3}\sum_{i=1}^{n}E\left[\frac{X_i\left(1-\mathrm{e}^{-X_i/\lambda}\right)}{1+x_i\mathrm{e}^{-X_i/\lambda}}\right] \qquad (8)$$

Missing information is given by

$$\mathrm{I}_{W|X}(\lambda) = \sum_{i=1}^{m} R_i \, \mathrm{I}_{W|X}^{(i)}(\lambda) = -\sum_{i=1}^{m}\sum_{j=1}^{R_i}\mathrm{E}_{Z|X}\left[\frac{d^2 \log\left(\mathrm{f}\left(Z_{ij} \mid X_i, \lambda\right)\right)}{d\lambda^2}\right]$$

Consider

$$\mathrm{f}_{z|X}\left(Z_{ij} \mid X_i, \lambda\right) = \frac{\mathrm{f}\left(z_{ij};\lambda\right)}{1-\mathrm{F}\left(x_i;\lambda\right)} = \frac{\frac{1}{\lambda}\frac{2\mathrm{e}^{-z_{ij}/\lambda}}{\left(1+\mathrm{e}^{-z_{ij}/\lambda}\right)^2}}{1-\left[\frac{1-\mathrm{e}^{-x_i/\lambda}}{1+\mathrm{e}^{-x_i/\lambda}}\right]}$$

Therefore,

$$\log \mathrm{f} = -\log\lambda + \log\left[\frac{2e^{-z_{ij}/\lambda}}{\left(1+e^{-z_{ij}/\lambda}\right)^2}\right] - \log\left[1-\left(\frac{1-e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}\right)\right]$$

$$\frac{d\log \mathrm{f}}{d\lambda} = -\frac{1}{\lambda} + \frac{z_{ij}\left(1-\mathrm{e}^{-z_{ij}/\lambda}\right)}{\lambda^2\left(1+\mathrm{e}^{-z_{ij}/\lambda}\right)} - \frac{x_i}{\lambda^2\left(1+\mathrm{e}^{-x_i/\lambda}\right)}$$

and

$$\frac{d^2\log \mathrm{f}}{d\lambda^2} = \frac{1}{\lambda^2} - \frac{2z_{ij}^2\mathrm{e}^{-z_{ij}/\lambda}}{\lambda^4\left(1+\mathrm{e}^{-z_{ij}/\lambda}\right)^2} - \frac{2z_{ij}\left(1-\mathrm{e}^{-z_{ij}/\lambda}\right)}{\lambda^3\left(1+\mathrm{e}^{-z_{ij}/\lambda}\right)} + \frac{2x_i^2\mathrm{e}^{-x_i/\lambda}}{\lambda^4\left(1+\mathrm{e}^{-x_i/\lambda}\right)^2} + \frac{2x_i}{\lambda^3\left(1+\mathrm{e}^{-x_i/\lambda}\right)}$$

Hence

$$\begin{aligned} \mathrm{I}_{W|X}(\lambda) &= \sum_{i=1}^{m} R_i \, \mathrm{I}_{W|X}^{(i)}(\lambda) \\ &= -\frac{n-m}{\lambda^2} + \frac{2}{\lambda^4}\sum_{i=1}^{m}\sum_{j=1}^{R_i}\mathrm{E}\left[\frac{z_{ij}^2 \mathrm{e}^{-z_{ij}/\lambda}}{\left(1+\mathrm{e}^{-z_{ij}/\lambda}\right)^2}\right] + \frac{2}{\lambda^3}\sum_{i=1}^{m}\sum_{j=1}^{R_i}\mathrm{E}\left[\frac{z_{ij}\left(1-\mathrm{e}^{-z_{ij}/\lambda}\right)}{1+\mathrm{e}^{-z_{ij}/\lambda}}\right] \quad (9) \\ &\quad -\frac{1}{\lambda^4}\sum_{i=1}^{m}\frac{R_i x_i^2 \mathrm{e}^{-x_i/\lambda}}{\left(1+\mathrm{e}^{-x_i/\lambda}\right)^2} - \frac{2}{\lambda^3}\sum_{i=1}^{m}\frac{R_i x_i}{\left(1+\mathrm{e}^{-x_i/\lambda}\right)} \end{aligned}$$

Confidence Intervals Based on MLE and log-Transformed MLE

Confidence Interval Based on MLE

Let $\hat{\lambda}_n$ be the MLE of λ and

$$\hat{\sigma}^2\left(\hat{\lambda}_n\right) = \frac{1}{\mathrm{I}\left(\hat{\lambda}_n\right)}$$

be the estimated asymptotic variance of $\hat{\lambda}_n$. Therefore, a 100(1 – α)% asymptotic CI for λ based on asymptotic normality of $\hat{\lambda}_n$ is given by

$$\left(\hat{\lambda}_n - \tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\hat{\lambda}_n\right)}, \hat{\lambda}_n + \tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\hat{\lambda}_n\right)}\right) \quad (10)$$

where $\tau_{\alpha/2}$ is the upper 100(α/2)$^{\text{th}}$ percentile of the standard normal distribution.

Confidence Interval Based on log-Transformed MLE

Meeker and Escobar (1998) reported the asymptotic CI for λ based on $\log\left(\hat{\lambda}_n\right)$. An approximate 100(1 – α)% CI for log(λ) is

$$\left(\log\left(\hat{\lambda}_n\right) - \tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\log\left(\hat{\lambda}_n\right)\right)}, \log\left(\hat{\lambda}_n\right) + \tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\log\left(\hat{\lambda}_n\right)\right)}\right)$$

where $\hat{\sigma}^2\left(\log\left(\hat{\lambda}_n\right)\right)$ is the estimated asymptotic variance of $\log\left(\hat{\lambda}_n\right)$, which is approximated by

$$\hat{\sigma}^2\left(\log\left(\hat{\lambda}_n\right)\right) \approx \frac{\hat{\sigma}^2\left(\hat{\lambda}_n\right)}{\hat{\lambda}_n^2}$$

Hence, an approximate $100(1-\alpha)\%$ CI for λ is

$$\left(\hat{\lambda}_n e^{\left(-\frac{\tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\hat{\lambda}_n\right)}}{\hat{\lambda}_n}\right)}, \hat{\lambda}_n e^{\left(\frac{\tau_{\alpha/2}\sqrt{\hat{\sigma}^2\left(\hat{\lambda}_n\right)}}{\hat{\lambda}_n}\right)}\right) \tag{11}$$

Confidence Interval Based on Pivotal and Generalized Pivotal Quantity

Consider two exact CIs based on the pivotal quantities. To define these CIs, show that the distribution of $V = \hat{\lambda}/\lambda$ is free from λ, where $\hat{\lambda}$ is the MLE of λ, based on the complete data. In the following lemma, it is proved that V is a pivot, following Gulati and Mi (2006):

Lemma 1: The distribution of V is free from λ.

Proof: Consider the probability density function of the half-logistic distribution with scale parameter λ:

$$f\left(x,\lambda\right) = \frac{2e^{-x/\lambda}}{\lambda\left(1+e^{-x/\lambda}\right)^2}, \quad x \geq 0, \lambda > 0$$

Then the log-likelihood function becomes

$$L = -n\log\left(\lambda\right) + n\log\left(2\right) - \frac{1}{\lambda}\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n}\log\left(1+e^{-x_i/\lambda}\right)$$

$dL/d\lambda = 0$ gives the following equation:

$$\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} \frac{x_i e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}} = n\lambda$$

The solution of the above equation is the MLE of λ (say $\hat{\lambda}$). Hence

$$\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} \frac{x_i e^{-x_i/\hat{\lambda}}}{1+e^{-x_i/\hat{\lambda}}} = n\hat{\lambda}$$

$$\sum_{i=1}^{n} \frac{x_i}{\lambda} - 2\sum_{i=1}^{n} \frac{x_i}{\lambda} \frac{e^{\frac{-x_i\lambda}{\lambda\hat{\lambda}}}}{1+e^{\frac{-x_i\lambda}{\lambda\hat{\lambda}}}} = \frac{n\hat{\lambda}}{\lambda}$$

Let $\xi = \lambda/\hat{\lambda}$ and $Y_i = X_i/\lambda$. Then

$$\sum_{i=1}^{n} y_i - 2\sum_{i=1}^{n} y_i \frac{e^{-y_i\xi}}{1+e^{-y_i\xi}} = n\xi^{-1}$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{2}{n}\sum_{i=1}^{n} y_i \frac{e^{-y_i\xi}}{1+e^{-y_i\xi}} - \xi^{-1} = 0$$

Note that $Y_1, Y_2, \ldots, Y_n$ is a random sample from the half-logistic distribution with parameter $\lambda = 1$. Therefore, the distribution of $\xi = \lambda/\hat{\lambda}$ is independent of λ. Hence the proof.

Lemma 2: The distribution of V under progressive Type-II censored data from the half-logistic distribution with scale parameter λ is free from λ.

Proof: This is similar to Lemma 1 and hence is omitted.

This property of the MLE will be used to derive the confidence interval based on pivot and generalized pivot quantity methods.

Remark: V is also a pivot for k-unit parallel and k-unit series systems.

Confidence Interval Based on Pivotal Quantity

From Lemma 2, the distribution of V is free from λ. Define a and b such that

$$\mathrm{P}\left(a < V < b\right) = 1 - \alpha$$

Therefore we obtain the following as a CI for λ:

$$\left(\frac{\hat{\lambda}}{b}, \frac{\hat{\lambda}}{a}\right) \tag{12}$$

The constants a and b are obtained using Monte Carlo simulation by using the following algorithm:

***Algorithm to Obtain Percentiles of* V**

1. Input α, N, m, and progressive Type-II censoring scheme (R_1, R_2,…, R_m).
2. Generate a progressive Type-II censored random sample of size m using censoring scheme (R_1, R_2,…, R_m) from the half-logistic distribution with parameter $\lambda = 1$.
3. Obtain a MLE of λ (say $\hat{\lambda}$) using the EM algorithm.
4. Repeat steps 2 and 3 N times so as to get $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_N$.
5. Arrange the $\hat{\lambda}_i$ in an increasing order. Denote them by $\hat{\lambda}_{(1)}, \hat{\lambda}_{(2)}, \ldots, \hat{\lambda}_{(N)}$.
6. Compute $a = \hat{\lambda}_{([(\alpha/2)N])}$ and $b = \hat{\lambda}_{([(1-\alpha/2)N])}$.

Confidence Interval Based on Generalized Pivotal Quantity

The concept of a generalized confidence interval (GCI) is introduced by Weerahandi (1993). Let x denote the observed value of X. To construct a GCI for λ, first define a generalized pivotal quantity (GPQ), $\mathrm{T}(X; x, \lambda)$, which is a function of the random variable X, its observed value x, and the parameter λ. A quantity $\mathrm{T}(X; x, \lambda)$ is required to satisfy the following two conditions:

i) For a fixed x, the probability distribution of $\mathrm{T}(X; x, \lambda)$ is free of unknown parameters.
ii) The observed value of $\mathrm{T}(X; x, \lambda)$, namely $\mathrm{T}(x; x, \lambda)$, is simply λ.

Let T_α be the $100\alpha^{\text{th}}$ percentile of T. Then T_α becomes the $100(1-\alpha)\%$ lower bound for λ. Therefore a $100(1-\alpha)\%$ two-sided GCI for parameter λ is given by

$$\left(\mathrm{T}_{\alpha/2}, \mathrm{T}_{1-\alpha/2}\right) \tag{13}$$

Define the GPQ as

$$\mathrm{T}\left(X; x, \lambda\right) = \frac{\hat{\lambda}_0}{\hat{\lambda}/\lambda}$$

where $\hat{\lambda}_0$ is the MLE obtained using observed data. Note:

i) The distribution of T(X; x, λ) is free from λ, which follows from Lemma 2, and
ii) T(x; x, λ) = λ, since for the observed data, $\hat{\lambda} = \hat{\lambda}_0$.

A GCI based on T(X; x, λ) is obtained by using following algorithm:

Algorithm to Obtain CI for λ using GPQ

1. Input α, N, m, and progressive Type-II censoring scheme (R_1, R_2,…, R_m).
2. Generate a progressive Type-II censored random sample of size m from the half-logistic distribution with an unknown parameter λ.
3. Based on the data in step 2, obtain a MLE of λ (say $\hat{\lambda}_0$) using the EM algorithm.
4. Generate a progressive Type-II censored random sample of size m from the half-logistic distribution with parameter $\lambda = 1$.
5. Obtain a MLE of λ (say $\hat{\lambda}_i$) using the EM algorithm for step 4 data.
6. Compute $T_i = \hat{\lambda}_0 / \hat{\lambda}_i$.
7. Repeat steps 4 to 6 N times, so as to get T_1, T_2,…,T_N.
8. Arrange the T_i in an increasing order. Denote them by $T_{(1)}$, $T_{(2)}$,…, $T_{(N)}$.
9. Compute a 100(1 – α)% CI for λ as $\left(T_{(\lceil(\alpha/2)N\rceil)}, T_{(\lceil(1-\alpha/2)N\rceil)}\right)$.

Simulation Study

The CIs given in (10) to (13) will now be compared with the CIs given by Balakrishnan and Asgharzadeh (2005) and Wang (2009). A simulation study was

carried out to study the performance of each of the CIs. Asymptotic CIs based on MLE, log-transformed MLE, and GPQ are compared through length and confidence level. Balakrishnan and Sandhu (1995) presented an algorithm for sample generation from progressively Type-II censored schemes. This algorithm was used to generate samples from a half-logistic distribution. Consider the 34 different progressively Type-II censored schemes compiled in Table 1.

Algorithm

1. Generate i.i.d. observations (W_1, W_2,…,W_m) from U(0, 1).
2. For censoring scheme (R_1, R_2,…, R_m),

$$E_i = \frac{1}{\left(i + R_m + R_{m-1} + \ldots + R_{m-i+1}\right)}$$

   for i = 1, 2,…, m.
3. Set $V_i = W_i^{E_i}$ for i = 1, 2,…,m.
4. Set $U_i = 1 - (V_m \cdot V_{m-1} \cdot \ldots \cdot V_{m-i+1})$ for i = 1, 2,…,m. Then (U_1, U_2,…, U_m) is the uniform (0, 1) progressively Type-II censored sample.
5. For given values of the parameter λ, set

$$x_{(i)} = -\lambda \log\left[\frac{1-U_i}{1+U_i}\right]$$

   for i = 1, 2,…, m.

Then ($x_{(1)}$, $x_{(2)}$,…, $x_{(m)}$) is the required progressively Type-II censored sample from the half-logistic distribution. In Table 1, censoring scheme (a, b, c, d) stands for $R_1 = a$, $R_2 = b$, $R_3 = c$, and $R_4 = d$. A similar meaning holds for schemes described through completely specified vector, while scheme (10, 9*0) means $R_1 = 10$ and remaining nine R_i are zero, i.e. $R_2 = R_3 = R_4 = \ldots = R_{10} = 0$. A simulation was carried out with $\lambda = 1$. For each particular progressive censoring scheme, 5,000 sets of observations are generated. The CIs based on asymptotic normal distributions of the MLE and log-transformed MLE are derived.

Table 1. Censoring schemes

Scheme No.	*n*	*m*	*m*/*n*	Scheme
[1]	10	4	0.2500	(0, 0, 0, 6)
[2]	10	4	0.2500	(6, 0, 0, 0)
[3]	10	5	0.5000	(0, 0, 0, 0, 5)
[4]	10	5	0.5000	(5, 0, 0, 0, 0)
[5]	15	4	0.2667	(0, 0, 0, 11)
[6]	15	4	0.2667	(11, 0, 0, 0)
[7]	15	5	0.3333	(0, 0, 0, 0, 10)
[8]	15	5	0.3333	(10, 0, 0, 0, 0)
[9]	15	5	0.3333	(0, 10, 0, 0, 0)
[10]	15	5	0.3333	(0, 0, 10, 0, 0)
[11]	15	5	0.3333	(2, 2, 2, 2, 2)
[12]	15	5	0.3333	(4, 4, 2, 0, 0)
[13]	20	5	0.2500	(0, 0, 0, 0, 15)
[14]	20	5	0.2500	(15, 0, 0, 0, 0)
[15]	20	5	0.2500	(5, 5, 5, 0, 0)
[16]	20	5	0.2500	(3, 3, 3, 3, 3)
[17]	20	5	0.2500	(0, 15, 0, 0, 0)
[18]	20	5	0.2500	(5, 10, 0, 0, 0)
[19]	20	10	0.5000	(9*0, 10)
[20]	20	10	0.5000	(10, 9*0)
[21]	25	5	0.2000	(0, 0, 0, 0, 20)
[22]	25	5	0.2000	(20, 0, 0, 0, 0)
[23]	25	10	0.4000	(9*0, 15)
[24]	25	10	0.4000	(15, 9*0)
[25]	25	15	0.6000	(14*0, 10)
[26]	25	15	0.6000	(10, 14*0)
[27]	50	20	0.4000	(19*0, 30)
[28]	50	20	0.4000	(30, 19*0)
[29]	50	25	0.5000	(24*0, 25)
[30]	50	25	0.5000	(25, 24*0)
[31]	100	20	0.2000	(19*0, 80)
[32]	100	20	0.2000	(80, 19*0)
[33]	100	50	0.5000	(49*0, 50)
[34]	100	50	0.5000	(50, 49*0)

Table 2. Simulated coverage probabilities for confidence intervals

	C_1		C_3		C_4		C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[1]	0.8100	0.8396	0.8108	0.8470	0.8710	0.9176	0.8944	0.9458	0.8992	0.9474
[2]	0.8300	0.8640	0.8338	0.8676	0.8804	0.9282	0.9072	0.9514	0.8986	0.9464
[3]	0.8288	0.8638	0.8330	0.8684	0.8768	0.9256	0.8968	0.9462	0.9025	0.9503
[4]	0.8290	0.8688	0.8382	0.8768	0.8814	0.9286	0.9014	0.9528	0.9036	0.9494
[5]	0.8204	0.8508	0.8160	0.8500	0.8786	0.9204	0.8978	0.9476	0.9016	0.9518
[6]	0.8350	0.8650	0.8364	0.8706	0.8830	0.9306	0.8978	0.9528	0.8948	0.9468
[7]	0.8194	0.8582	0.8278	0.8640	0.8736	0.9230	0.8998	0.9522	0.9058	0.9548
[8]	0.8360	0.8686	0.8418	0.8778	0.8834	0.9284	0.9006	0.9528	0.8998	0.9482
[9]	0.8370	0.8684	0.8398	0.8724	0.8794	0.9240	0.9050	0.9526	0.8986	0.9498
[10]	0.8354	0.8656	0.8364	0.8666	0.8780	0.9306	0.8946	0.9456	0.8978	0.9506
[11]	0.8262	0.8596	0.8308	0.8684	0.8822	0.9274	0.9022	0.9494	0.9050	0.9518
[12]	0.8354	0.8650	0.8408	0.8798	0.8896	0.9336	0.9014	0.9514	0.8934	0.9486
[13]	0.8318	0.8626	0.8418	0.8750	0.8842	0.9348	0.9002	0.9504	0.8966	0.9520
[14]	0.8474	0.8806	0.8474	0.8834	0.8866	0.9342	0.8960	0.9474	0.8974	0.9462
[15]	0.8368	0.8740	0.8388	0.8716	0.8752	0.9250	0.8974	0.9528	0.9008	0.9482
[16]	0.8308	0.8632	0.8312	0.8664	0.8816	0.9260	0.9048	0.9532	0.8950	0.9496
[17]	0.8432	0.8724	0.8492	0.8818	0.8870	0.9296	0.9004	0.9504	0.9000	0.9464
[18]	0.8318	0.8690	0.8390	0.8756	0.8788	0.9260	0.8944	0.9488	0.8998	0.9500
[19]	0.8592	0.8954	0.8790	0.9122	0.8902	0.9416	0.8960	0.9510	0.8950	0.9458
[20]	0.8680	0.9068	0.8706	0.9098	0.8864	0.9358	0.9002	0.9528	0.8958	0.9418
[21]	0.8196	0.8544	0.8280	0.8606	0.8764	0.9284	0.8990	0.9496	0.8976	0.9492
[22]	0.8372	0.8720	0.8400	0.8712	0.8764	0.9304	0.8972	0.9542	0.8970	0.9504
[23]	0.8640	0.9072	0.8636	0.8994	0.8858	0.9364	0.8976	0.9490	0.8980	0.9454
[24]	0.8774	0.9128	0.8780	0.9132	0.8964	0.9434	0.8904	0.9466	0.9010	0.9512
[25]	0.8714	0.9160	0.8770	0.9158	0.8948	0.9432	0.8926	0.9448	0.9006	0.9466
[26]	0.8822	0.9210	0.8848	0.9242	0.8996	0.9504	0.9008	0.9492	0.8938	0.9468
[27]	0.8844	0.9246	0.8790	0.9212	0.8914	0.9388	0.9002	0.9502	0.8970	0.9472
[28]	0.8852	0.9302	0.8880	0.9292	0.8952	0.9470	0.9084	0.9532	0.8948	0.9496

Table 2, continued.

	C_1		C_3		C_4		C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[29]	0.8904	0.9276	0.8950	0.9360	0.9022	0.9494	0.9024	0.9466	0.8948	0.9504
[30]	0.8896	0.9348	0.8918	0.9374	0.8982	0.9484	0.9044	0.9530	0.8978	0.9478
[31]	0.8920	0.9324	0.8856	0.9248	0.8962	0.9460	0.9008	0.9526	0.8968	0.9486
[32]	0.8864	0.9306	0.8876	0.9336	0.8972	0.9478	0.9062	0.9534	0.8958	0.9478
[33]	0.8930	0.9374	0.8938	0.9408	0.8998	0.9454	0.8958	0.9446	0.9046	0.9530
[34]	0.8924	0.9416	0.9010	0.9452	0.9026	0.9522	0.8948	0.9448	0.9070	0.9544

Table 3. The expected lengths of confidence intervals

	C_1		C_2		C_3		C_4		C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[1]	2.0913	2.7742	2.0330	2.7028	1.3723	1.6352	1.4919	1.8397	2.0003	2.6406	2.0432	2.7096
[2]	2.0150	2.6663	1.9223	2.5345	1.3790	1.6432	1.4943	1.8403	1.9281	2.5360	1.9254	2.5328
[3]	1.6829	2.2413	1.6495	2.1395	1.2142	1.4468	1.2952	1.5849	1.6353	2.1214	1.6562	2.1440
[4]	1.6656	2.1061	1.5932	2.0518	1.2246	1.4592	1.3051	1.5965	1.5883	2.0467	1.5690	2.0143
[5]	2.1526	2.8298	2.1217	2.8244	1.4289	1.7026	1.5625	1.9313	2.1204	2.8675	2.0944	2.7809
[6]	2.0219	2.8139	1.9415	2.5615	1.3863	1.6519	1.5039	1.8530	1.9146	2.5256	1.9121	2.5117
[7]	1.8253	2.3360	1.7234	2.2392	1.2655	1.5079	1.3562	1.6627	1.7120	2.2377	1.7132	2.2203
[8]	1.7290	2.2818	1.6054	2.0685	1.2395	1.4770	1.3220	1.6177	1.6076	2.0631	1.5954	2.0493
[9]	1.6816	2.1968	1.6431	2.1214	1.2488	1.4880	1.3343	1.6339	1.6136	2.0929	1.6358	2.1071
[10]	1.8064	2.2591	1.6754	2.1675	1.2566	1.4973	1.3445	1.6474	1.6653	2.1710	1.6636	2.1482
[11]	1.7245	2.2904	1.6782	2.1775	1.2430	1.4812	1.3285	1.6270	1.6886	2.2053	1.6426	2.1253
[12]	1.6759	2.1434	1.6449	2.1252	1.2481	1.4872	1.3333	1.6326	1.6374	2.1200	1.6348	2.1033
[13]	1.8299	2.4993	1.7724	2.3044	1.3030	1.5526	1.4010	1.7199	1.7660	2.2984	1.7672	2.2909
[14]	1.6007	2.0857	1.6130	2.0789	1.2401	1.4776	1.3232	1.6194	1.5938	2.0671	1.5858	2.0396
[15]	1.7540	2.2729	1.6768	2.1690	1.2731	1.5170	1.3625	1.6695	1.6698	2.1834	1.6496	2.1262
[16]	1.7848	2.3377	1.7207	2.2350	1.2532	1.4933	1.3429	1.6464	1.6982	2.2097	1.7251	2.2365
[17]	1.7424	2.1501	1.6597	2.1438	1.2722	1.5159	1.3607	1.6669	1.6277	2.1042	1.6401	2.1126

Table 3, continued.

	C_1		C_2		C_3		C_4		C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[18]	1.7336	2.1373	1.6528	2.1345	1.2618	1.5035	1.3490	1.6523	1.6297	2.1138	1.6378	2.1099
[19]	1.0242	1.2681	1.0099	1.2531	0.8758	1.0436	0.9047	1.0926	1.0153	1.2497	1.0011	2.2410
[20]	1.0137	1.2284	0.9834	1.2145	0.8717	1.0387	0.8998	1.0864	0.9957	1.2302	0.9712	1.1978
[21]	1.8246	2.3465	1.8066	2.3495	1.3169	1.5692	1.4194	1.7442	1.8067	2.3370	1.8018	2.3372
[22]	1.6455	2.0421	1.6180	2.0857	1.2377	1.4748	1.3211	1.6170	1.6001	2.0816	1.5875	2.0391
[23]	1.0462	1.2845	1.0328	1.2825	0.8884	1.0586	0.9189	1.1104	1.0393	1.2960	1.0311	1.2787
[24]	1.0103	1.2819	0.9854	1.2171	0.8753	1.0430	0.9036	1.0911	0.9800	1.2079	0.9812	1.2099
[25]	0.7842	0.9543	0.7775	0.9509	0.7016	0.8360	0.7165	0.8613	0.7766	0.9502	0.7754	0.9475
[26]	0.7846	0.9490	0.7714	0.9407	0.7079	0.8435	0.7229	0.8691	0.7677	0.9354	0.7671	0.9342
[27]	0.6895	0.8386	0.6832	0.8310	0.6328	0.7540	0.6436	0.7723	0.6820	0.8351	0.6820	0.8275
[28]	0.6546	0.8045	0.6550	0.7944	0.6162	0.7343	0.6261	0.7510	0.6526	0.7914	0.6561	0.7941
[29]	0.6009	0.7334	0.5902	0.7144	0.5567	0.6634	0.5640	0.6758	0.5945	0.7184	0.5879	0.7109
[30]	0.5796	0.7047	0.5780	0.6982	0.5513	0.6569	0.5583	0.6688	0.5752	0.6973	0.5761	0.6951
[31]	0.7042	0.8616	0.7249	0.8823	0.6713	0.7999	0.6842	0.8217	0.7312	0.8881	0.7259	0.8817
[32]	0.6482	0.7763	0.6563	0.7960	0.6176	0.7359	0.6275	0.7526	0.6639	0.8022	0.6546	0.7929
[33]	0.4067	0.4736	0.4067	0.4884	0.3951	0.4708	0.3977	0.4752	0.4043	0.4892	0.4047	0.4859
[34]	0.3985	0.4815	0.3992	0.4789	0.3897	0.4644	0.3922	0.4686	0.4014	0.4818	0.3968	0.4754

Table 4. Coverage to Length Ratio (CLR) of confidence intervals

	C_1		C_3		C_4		C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[1]	0.3873	0.3026	0.5908	0.5180	0.5838	0.4988	0.4471	0.3582	0.4401	0.3497
[2]	0.4119	0.3240	0.6046	0.5280	0.5892	0.5044	0.4705	0.3752	0.4667	0.3737
[3]	0.4925	0.3854	0.6860	0.6002	0.6770	0.5840	0.5484	0.4460	0.5449	0.4433
[4]	0.4977	0.4125	0.6845	0.6009	0.6754	0.5816	0.5675	0.4655	0.5759	0.4713
[5]	0.3811	0.3007	0.5711	0.4992	0.5623	0.4766	0.4234	0.3305	0.4305	0.3423
[6]	0.4130	0.3074	0.6033	0.5270	0.5871	0.5022	0.4689	0.3773	0.4680	0.3770

Table 4, continued.

	C1		C3		C4		C5		C6	
Scheme	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**	**90%**	**95%**
[7]	0.4489	0.3674	0.6541	0.5730	0.6442	0.5551	0.5256	0.4255	0.5287	0.4300
[8]	0.4835	0.3807	0.6791	0.5943	0.6682	0.5739	0.5602	0.4618	0.5640	0.4627
[9]	0.4977	0.3953	0.6725	0.5863	0.6591	0.5655	0.5609	0.4552	0.5493	0.4508
[10]	0.4625	0.3832	0.6656	0.5788	0.6530	0.5649	0.5372	0.4356	0.5397	0.4425
[11]	0.4791	0.3753	0.6684	0.5863	0.6641	0.5700	0.5343	0.4305	0.5510	0.4478
[12]	0.4985	0.4036	0.6737	0.5916	0.6672	0.5718	0.5505	0.4488	0.5465	0.4510
[13]	0.4546	0.3451	0.6460	0.5636	0.6311	0.5435	0.5097	0.4135	0.5073	0.4156
[14]	0.5294	0.4222	0.6833	0.5979	0.6700	0.5769	0.5622	0.4583	0.5659	0.4639
[15]	0.4771	0.3845	0.6589	0.5746	0.6423	0.5541	0.5374	0.4364	0.5461	0.4460
[16]	0.4655	0.3693	0.6633	0.5802	0.6565	0.5624	0.5328	0.4314	0.5188	0.4246
[17]	0.4839	0.4057	0.6675	0.5817	0.6519	0.5577	0.5532	0.4517	0.5487	0.4480
[18]	0.4798	0.4066	0.6649	0.5824	0.6514	0.5604	0.5488	0.4489	0.5494	0.4503
[19]	0.8389	0.7061	1.0037	0.8741	0.9840	0.8618	0.8825	0.7610	0.8941	0.7621
[20]	0.8563	0.7382	0.9987	0.8759	0.9851	0.8614	0.9041	0.7745	0.9224	0.7863
[21]	0.4492	0.3641	0.6287	0.5484	0.6174	0.5323	0.4976	0.4063	0.4982	0.4061
[22]	0.5088	0.4270	0.6787	0.5907	0.6634	0.5754	0.5607	0.4584	0.5650	0.4661
[23]	0.8258	0.7063	0.9721	0.8496	0.9640	0.8433	0.8637	0.7323	0.8709	0.7393
[24]	0.8685	0.7121	1.0031	0.8756	0.9920	0.8646	0.9085	0.7836	0.9183	0.7862
[25]	1.1112	0.9599	1.2500	1.0955	1.2488	1.0951	1.1493	0.9943	1.1614	0.9990
[26]	1.1244	0.9705	1.2499	1.0957	1.2444	1.0935	1.1733	1.0148	1.1651	1.0135
[27]	1.2827	1.1026	1.3891	1.2218	1.3850	1.2156	1.3199	1.1378	1.3153	1.1447
[28]	1.3523	1.1562	1.4411	1.2654	1.4298	1.2610	1.3920	1.2045	1.3639	1.1959
[29]	1.4818	1.2648	1.6077	1.4109	1.5996	1.4049	1.5180	1.3177	1.5220	1.3368
[30]	1.5349	1.3265	1.6176	1.4270	1.6088	1.4181	1.5722	1.3668	1.5584	1.3635
[31]	1.2667	1.0822	1.3192	1.1561	1.3099	1.1513	1.2319	1.0727	1.2354	1.0759
[32]	1.3675	1.1988	1.4372	1.2687	1.4298	1.2594	1.3651	1.1885	1.3684	1.1954
[33]	2.1957	1.9793	2.2622	1.9983	2.2625	1.9895	2.2158	1.9311	2.2351	1.9614
[34]	2.2394	1.9556	2.3120	2.0353	2.3014	2.0320	2.2291	1.9611	2.2857	2.0076

We denote by C_1 the CI proposed by Balakrishnan and Asgharzadeh (2005), by C_2 the CI proposed Wang (2009), by C_3 the CI based on the MLE obtained by the EM algorithm, by C_4 the CI based on the log-transformed MLE, by C_5 the CI based on pivotal quantity, and by C_6 the GCI. Coverage probabilities of the CIs for various censoring schemes are displayed in Table 2. Coverage probabilities of C_1 are also displayed in the same table. Coverage probabilities for C_2 are not provided by Wang (2009). Lengths of CIs for the various censoring schemes are given in Table 3. For comparison, lengths of C_1 and C_2 are given in the same table.

For effective comparison of CIs, we compute coverage to length ratio (CLR). CLR for C_1, C_3, C_4, C_5, and C_6 are given in Table 4. It is clear that the CIs having a higher value of CLR are preferred.

Conclusion

Coverage probabilities of C_3, C_4, C_5, and C_6 are better than coverage probabilities of C_1. Comparing coverage probabilities of all four CIs, C_5 and C_6 show the best performance. For small and large sample sizes (n) and the smallest effective sample size (m), C_5 and C_6 show good coverage probability. For large sample sizes, C_3, C_4, C_5, and C_6 show good performance. As n and m increase, coverage probability of C_3 and C_4 increases rapidly as compared to C_5 and C_6. C_6 has higher coverage probability for conventional censoring schemes than progressive censoring schemes, but C_3 and C_4 show higher coverage probability for progressive censoring schemes than conventional censoring schemes.

C_3 has smaller length than the lengths of C_1 and C_2. The MLE by the EM algorithm provides the shortest length CI among all five CIs. For large sample sizes, the length of C_6 approaches the length of C_3. Lengths of all CIs decrease as n and m increase. Lengths of CIs based on progressive censoring schemes are smaller than lengths of CIs based on conventional censoring schemes. There is a minor difference among lengths of C_3, C_4, C_5, and C_6 for large sample sizes. According to the CLR, C_3 is the best among the four CIs for small sample sizes. C_4, C_5, and C_6 also show higher CLR than the CLR of C_1. CLRs of CIs based on progressive censoring schemes are better than CLRs of CIs based on conventional censoring.

Acknowledgements

The first author wishes to thank the University Grants Commission, New Delhi, India for providing fellowship under the Faculty Improvement Programme to carry out this research.

References

Balakrishnan, N. (2007). Progressive censoring methodology: An appraisal (with discussion). *Test, 16*(2), 211-296.doi: 10.1007/s11749-007-0061-y

Balakrishnan, N., & Aggarwala, R. (2000). *Progressive censoring: Theory, methods, and applications*. Boston, MA: Birkhauser.doi: 10.1007/978-1-4612-1334-5

Balakrishnan, N., & Asgharzadeh, A. (2005). Inference for the scaled half-logistic distribution based on progressively Type-II censored samples. *Communications in Statistics – Theory and Methods, 34*(1), 73-87.doi: 10.1081/sta-200045814

Balakrishnan, N., & Chan, P.S. (1992). Estimation for the scaled half logistic distribution under Type-II censoring.*Computational Staistics& Data Analysis, 13*(2), 123-141.doi: 10.1016/0167-9473(92)90001-v

Balakrishnan, N., Kannan, N., Lin, C.T., & Ng, H. K. T. (2003). Point and interval estimation for Gaussian distribution, based on progressively Type-II censored samples.*IEEE Transactions on Reliability, 52*(1), 90-95. doi: 10.1109/tr.2002.805786

Balakrishnan, N., Kannan, N., Lin, C.T., & Wu, S. J. S. (2004). Inference for the extreme value distribution under progressive Type-II Censoring.*Journal of Statistical Computationand Simulation, 74*(1), 25-45.doi: 10.1080/0094965031000105881

Balakrishnan, N., &Puthenpura, S. (1986). Best linear unbiased estimation of location and scale parameters of the half logistic distribution.*Journal of Statistical Computation and Simulation, 25*(3-4), 193-204.doi: 10.1080/00949658608810932

Balakrishnan, N., & Sandhu, R. A. (1995). A simple simulation algorithm for generating progressive Type-II censored samples.*The American Statistician, 49*(2), 229-230.doi: 10.2307/2684646

Balakrishnan, N., & Wong, K. H. T. (1991). Approximate MLEs for the location and scaled parameters of the half logistic distribution with Type-II right censoring.*IEEE Transactions on Reliability, 40*(2), 140-145. doi: 10.1109/24.87114

Cohen, A. C. (1963). Progressively censored samples in life testing.*Technometrics, 5*(3), 327-329.doi: 10.2307/1266337

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm.*Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1-38. Available from: http://www.jstor.org/stable/2984875

Ghitany, M. E., Alqallaf, F., & Balakrishnan, N. (2014). On the likelihood estimation of the parameters of Gompertz distribution based on complete and progressively Type-II censored samples. *Journal of Statistical Computation and Simulation, 84*(8), 1803-1812.doi: 10.1080/00949655.2013.766738

Gulati, S., & Mi, J. (2006). Testing for scale families using total variation distance.*Journal of Statistical Computation and Simulation, 76*(9), 773-792.doi: 10.1080/10629360500282080

Jang, D. H., Park, J., & Kim, C. (2011). Estimation of the scale parameter of the half-logistic distribution with multiply type II censored sample.*Journal of the Korean Statistical Society, 40*(3), 291-301.doi: 10.1016/j.jkss.2010.12.001

Kim, C., & Han, K. (2010). Estimation of the scale parameter of the half-logistic distribution under progressively typeII censored sample.*Statistical Papers, 51*(2), 375-387.doi: 10.1007/s00362-009-0197-9

Lawless, J. F. (1982).*Statistical models and methods for lifetime data*.New York, NY: John Wiley and Sons.

Lin, C.-T., & Balakrishnan, N. (2011). Asymptotic properties of maximum likelihood estimators based on progressively Type-II censoring.*Metrika, 74*(3), 349-360.doi: 10.1007/s00184-010-0306-8

Louis, T. A. (1982). Finding the observed information matrix using the EM algorithm.*Journal of the Royal Statistical Society. Series B (Methodological), 44*(2), 226-233. Available from: http://www.jstor.org/stable/2345828

Mann, N. R. (1969). Exact three-order-statistic confidence bounds on reliable life for a Weibull model with progressive censoring. Journal of the American Statistical Association, 64(325), 306-315.doi: 10.2307/2283740

Mann, N. R. (1971). Best linear invariant estimation for Weibull parameters under progressive censoring.*Technometrics, 13*(3), 521-533.doi: 10.2307/1267165

McLachlan, G. J., & Krishnan, T. (1997).*The EM algorithm and extensions*.New York, NY: John Wiley and Sons.

Meeker, W. Q., & Escobar, L. A. (1998).*Statistical methods for reliability data*.New York, NY: John Wiley and Sons.

Ng, H. K. T. (2005). Parameter estimation for a modified Weibull distribution for progressively Type-II censored samples.*IEEE Transactions on Reliability, 54*(3), 374-380.doi: 10.1109/tr.2005.853036

Ng, H. K. T., Kundu, D., & Balakrishnan, N. (2006). Point and interval estimation for the two-parameter Birnbaum-Saunders distribution based on progressively Type-II censored data.*Computational Statistics & Data Analysis, 50*(11), 3222-3242.doi: 10.1016/j.csda.2005.06.002

Potdar, K. G., & Shirke, D. T. (2013). Reliability estimation for the distribution of a *k*-unit parallel system with Rayleigh distribution as the component life distribution.*International Journal of Engineering Research and Technology, 2*(8), 2362-2371.

Potdar, K. G., & Shirke, D. T. (2014). Inference for the scale parameter of lifetime distribution of *k*-unit parallel system based on progressively censored data. *Journal of Statistical Computation and Simulation, 84*(1), 171-185.doi: 10.1080/00949655.2012.700314

Rastogi, M. K., & Tripathi, Y. M. (2014). Parameter and reliability estimation for an exponentiated half-logistic distribution under progressive type II censoring.*Journal of Statistical Computation and Simulation, 84*(8), 1711-1727. doi: 10.1080/00949655.2012.762366

Sultan, K. S., Alsadat, N. H., & Kundu, D. (2014). Bayesian and maximum likelihood estimations of the inverse Weibull parameters under progressive type-II censoring. Journal of Statistical Computation and Simulation, 84(10), 2248-2265.doi: 10.1080/00949655.2013.788652

Wang, B. (2009). Interval estimation for the scaled half logistic distribution under progressive Type-II censoring.*Communicationsin Statistics – Theory and Methods, 38*(3), 364-371.doi: 10.1080/03610920802213681

Weerahandi, S. (1993). Generalized confidence intervals. Journal of the American Statistical Association, 88(423), 899-905.doi: 10.2307/2290779

Appendix A. Illustrative Examples

Numeric Example

Balakrishnan and Asgharzadeh (2005) gave simulated sample of size $n = 50$ from the half-logistic distribution with scale parameter $\lambda = 25$. This complete sample is

1.7110, 2.0024, 2.3963, 3.9034, 4.6412, 6.4002, 6.7956, 8.5646, 8.6428, 8.8354, 9.3518, 9.7358, 10.5080, 10.5095, 11.8015, 12.8005, 16.3451, 16.9938, 17.2101, 18.5384, 20.3508, 21.1838, 22.1529, 22.4062, 22.4381, 23.0369, 25.8435, 27.0574, 27.1237, 29.0360, 30.6449, 32.5713, 33.6688, 40.3890, 45.4092, 46.4756, 49.8833, 51.1798, 53.0397, 53.8135, 64.9315, 66.1807, 69.9004, 75.2674, 75.4427, 75.7291, 76.1571, 89.5827, 99.8525, 134.6488.

Balakrishnan and Asgharzadeh (2005) and Wang (2009) derived CIs for this complete sample and the censored sample. We also derive CIs by using the MLE obtained by the EM algorithm, and the CIs based on pivot and generalized pivot. In Table 5, we consider two cases suggested by Wang (2009). Also we use the censoring schemes and samples given by Wang (2009) and derive 90% and 95% CIs and their lengths. For comparison, we display CIs and their lengths as stated by Wang (2009).

Table 5. Confidence interval and its length for illustrative example: $n = 50$, $\lambda = 25$

	C_2		C_3	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(24.49, 42.97)	(23.37, 45.72)	(22.76, 40.26)	(21.08, 41.94)
(25*1)	18.48	22.35	17.50	20.86
Case 2	(20.93, 34.82)	(20.05, 36.81)	(19.95, 33.28)	(18.67, 34.56)
(28*0, 10,10)	13.89	16.76	13.33	15.89

	C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(24.52, 42.94)	(23.38, 45.67)	(24.05, 42.82)	(23.18, 45.66)
(25*1)	18.42	22.29	18.77	22.48
Case 2	(21.21, 35.21)	(20.31, 37.23)	(21.42, 34.93)	(20.31, 37.24)
(28*0, 10,10)	14.00	16.92	13.51	16.93

Note: For Case 1, Sr. No. is 1 and $m = 25$. For Case 2, Sr. No. is 2 and $m = 30$.

Table 6. Confidence interval and its length for illustrative example: $n = 50$, $\lambda = 25$

	C_1		C_3	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(19.81, 29.53)	(18.90, 30.45)	(19.88, 29.48)	(18.96, 30.40)
(50*0)	9.72	11.55	9.6	11.44
Case 2	(20.78, 32.12)	(19.72, 33.18)	(18.88, 29.21)	(17.89, 30.20)
(39*0, 10)	11.34	13.46	10.33	12.31
Case 3	(18.66, 31.16)	(17.48, 32.34)	(15.92, 26.62)	(14.89, 27.65)
(29*0, 20)	12.5	14.86	10.7	12.76

	C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(20.59, 30.37)	(19.85, 31.60)	(20.55, 30.26)	(19.92, 31.28)
(50*0)	9.78	11.75	9.71	11.36
Case 2	(19.68, 30.38)	(18.94, 31.81)	(19.53, 30.07)	(18.95, 31.47)
(39*0, 10)	10.7	12.87	10.54	12.52
Case 3	(16.95, 28.23)	(16.23, 29.80)	(16.90, 28.20)	(16.06, 29.92)
(29*0, 20)	11.28	13.57	11.3	13.86

Note: For Case 1, Sr. No. is 1 and $m = 50$. For Case 2, Sr. No. is 2 and $m = 40$. For Case 3, Sr. No. is 3 and $m = 30$.

Balakrishnan and Asgharzadeh (2005) considered three cases, ($n = 50$, $m = 50$), ($n = 50$, $m = 40$), and ($n = 50$, $m = 30$). They used progressive and conventional Type-II censored samples but have not provided samples. To compare the proposed CIs with the CI proposed by Balakrishnan and Asgharzadeh (2005), we considered conventional censored and complete samples considered by Balakrishnan and Asgharzadeh (2005). We obtained 90% and 95% CIs for these schemes. In Table 6, 90% and 95% CIs and their lengths are displayed. Also, the CIs and their length proposed by Balakrishnan and Asgharzadeh (2005) are displayed.

Observe that in the illustrated example, C_3 has shorter length than the lengths of C_1, C_2 and C_5. C_6 has shorter length than that of C_1.

Real Data Example

Lawless (1982) presented real data which represented failure times for a specific type of electrical insulation that was subjected to a continuously increasing voltage stress.

12.3, 21.8, 24.4, 28.6, 43.2, 46.9, 70.7, 75.3, 95.5, 98.1, 138.6, 151.9.

Table 7. Confidence interval and its length for real data: $n = 12$, $\lambda = 50.50$ (BLUE)

	C_3		C_4	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(28.59, 66.24)	(24.98, 69.85)	(31.88, 70.53)	(29.54, 76.10)
(12*0)	37.65	44.87	38.65	46.56
Case 2	(25.55, 73.70)	(20.94, 78.31)	(30.55, 80.61)	(27.84, 88.46)
(7*0, 4)	48.15	57.37	50.06	60.62
Case 3	(23.35, 68.29)	(19.05, 72.59)	(28.06, 74.82)	(25.54, 82.19)
(4, 7*0)	44.94	53.54	46.74	56.65

	C_5		C_6	
Scheme	**90%**	**95%**	**90%**	**95%**
Case 1	(33.37, 75.18)	(31.19, 82.30)	(33.65, 73.96)	(31.88, 83.36)
(12*0)	41.81	51.11	40.31	51.48
Case 2	(33.13, 90.13)	(30.73, 101.89)	(32.60, 86.50)	(30.13, 94.26)
(7*0, 4)	57	71.16	53.9	64.13
Case 3	(30.14, 82.01)	(27.78, 92.25)	(30.55, 83.15)	(27.58, 92.42)
(4, 7*0)	51.87	64.47	52.6	64.84

Note: For Case 1, Sr. No. is 1 and $m = 12$. For Case 2, Sr. No. is 2 and $m = 8$. For Case 3, Sr. No. is 3 and $m = 8$.

The half-logistic distribution fits the data extremely well (Balakrishnan & Chan, 1992). This dataset was used with two censoring schemes, (7*0, 4) and (4, 7*0), and complete data, and the CI is constructed based on the MLE, log-MLE, pivot, and generalized pivot. These 90% and 95% CIs and their lengths are presented in Table 7. Observe that, for real data, C_3 has shorter length than C_4, C_5 and C_6.

The EM algorithm approach works well for small sample size n and the smallest effective sample size m. Overall, the proposed CIs perform better than the CIs proposed by Balakrishnan and Asgharzadeh (2005) and Wang (2009). The proposed CIs are superior to the other two CIs with regard to the length and the coverage probability.

Journal

Communications in Statistics - Theory and Methods ›

Volume 46, 2017 - Issue 20

99 Views | 0 CrossRef citations | 0 Altmetric

Original Articles

A unified approach to adaptive Shewhart control charts

Shashibhushan B. Mahadik

Pages 10272-10293 | Received 17 Jul 2015, Accepted 01 Sep 2016, Accepted author version posted online: 10 Oct 2016, Published online: 24 Jul 2017

Download citation https://doi.org/10.1080/03610926.2016.1235192 Check for updates

ABSTRACT

A *completely adaptive* (CA) $\bar{X}$ chart, that is, an $\bar{X}$ chart in which sampling interval, sample size, control limits, and warning limits are all adaptive and switch between two values, is explored. The exact expressions for the statistical and operational performance measures for this chart are derived. Obviously, these expressions are directly applicable to all the $\bar{X}$ charts in which any one or more of the design parameters are adaptive and switch between two values, as those are particular cases of a CA $\bar{X}$ chart. Thus, a CA $\bar{X}$ chart provides a unified approach to explore all those charts. The simultaneous evaluation of all such charts through extensive numerical comparisons of their performances accentuated how each of the design parameters affects the chart performances when it is made adaptive. Also, the

comparisons facilitated to determine the optimal adaptive $\bar{X}$ charts for different situations in the sense of having the best overall performance. Investigation of a

Journal

Communications in Statistics - Theory and Methods ›

Latest Articles

Original Articles

Consistent and robust variable selection in regression based on Wald test

T. S. Kamble, D. N. Kashid & D. M. Sakate

Received 08 Feb 2017, Accepted 10 Feb 2018, Published online: 06 Mar 2018

Download citation https://doi.org/10.1080/03610926.2018.1440598

ABSTRACT

Selection of relevant predictor variables for building a model is an important problem in the multiple linear regression. Variable selection method based on ordinary least squares estimator fails to select the set of relevant variables for building a model in the presence of outliers and leverage points. In this article, we propose a new robust variable selection criterion for selection of relevant variables in the model and establish its consistency property. Performance of the proposed method is evaluated through simulation study and real data.

KEYWORDS: M-estimator, penalty, outlier, Huber function, leverage points

Additional information

Improved inference for the shape–scale family of distributions under type–II censoring

Article *in* Journal of Statistical Computation and Simulation · March 2018

DOI: 10.1080/00949655.2018.1453812

2 authors:

Hemangi V. Kulkarni
Shivaji University, Kolhapur
27 PUBLICATIONS **68** CITATIONS
SEE PROFILE

Kiran Patil
Shivaji University, Kolhapur
9 PUBLICATIONS **60** CITATIONS
SEE PROFILE

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: http://www.tandfonline.com/loi/gscs20

Improved inference for the shape-scale family of distributions under type-II censoring

H. V. Kulkarni & K. P. Patil

To cite this article: H. V. Kulkarni & K. P. Patil (2018): Improved inference for the shape-scale family of distributions under type-II censoring, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2018.1453812

To link to this article: https://doi.org/10.1080/00949655.2018.1453812

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2018
https://doi.org/10.1080/00949655.2018.1453812

Taylor & Francis
Taylor & Francis Group

Improved inference for the shape-scale family of distributions under type-II censoring

H. V. Kulkarni[a] and K. P. Patil[a,b]

[a]Department of Statistics, Shivaji University, Kolhapur, Maharashtra, India; [b]Department of Statistics, Anandibai Raorane Arts, Commerce and Science College Vaibhavwadi, Sindhudurga, Maharashtra, India

ABSTRACT

The presence of a nuisance parameter may often perturb the quality of the likelihood-based inference for a parameter of interest under small to moderate sample sizes. The article proposes a maximal scale invariant transformation for likelihood-based inference for the shape in a shape-scale family to circumvent the effect of the nuisance scale parameter. The transformation can be used under complete or type-II censored samples. Simulation-based performance evaluation of the proposed estimator for the popular Weibull, Gamma and Generalized exponential distribution exhibits markedly improved performance in all types of likelihood-based inference for the shape under complete and type-II censored samples. The simulation study leads to a linear relation between the bias of the classical maximum likelihood estimator (MLE) and the transformation-based MLE for the popular Weibull and Gamma distributions. The linearity is exploited to suggest an almost unbiased estimator of the shape parameter for these distributions. Allied estimation of scale is also discussed.

ARTICLE HISTORY
Received 22 December 2016
Accepted 14 March 2018

KEYWORDS
Maximal scale invariant transformation; type-II censoring; almost unbiased estimator; Weibull distribution; Gamma distribution; Generalized exponential distribution

1. Introduction

The popular shape-scale probability models like Weibull, Gamma, Pareto, Log-normal, Generalized exponential (GE), etc. are basically skewed in nature and have been employed to model a wealth of real phenomenon in almost all disciplines. See for example, the monographs by Rinne [1], Abernethy [2] and McCool [3], among other references. The shape parameter in a shape-scale family controls the shape of a distribution without shifting or stretching it. Often the inference related to the shape could be of prime importance, see for example, Jiang and Murthy [4]. Krishnamoorthy et al. [5], Powar and Kulkarni [6], SenGupta et al. [7], Bagdonavičius et al. [8] and Patil and Kulkarni [9] among others discussed various applications of shape-scale family of distributions and related inferential procedures. Kulkarni and Patil [10] discussed the two sample comparisons including zero-inflated continuous data with the applications of shape-scale as well as location-scale distributions in the field of molecular biology. Abundant literature exists

CONTACT K. P. Patil kiran.patil9020@gmail.com Department of Statistics, Shivaji University, Kolhapur, Maharashtra, India Department of Statistics, Anandibai Raorane Arts, Commerce and Science College Vaibhavwadi, Sindhudurga, Maharashtra, India

Supplemental data for this article can be accessed here. https://doi.org/10.1080/00949655.2018.1453812

on the estimation of parameters of shape-scale distributions, see for example,Zaigraev and Podraza-Karakulska [11] and Tanaka et al. [12] among others for estimation of the Gamma shape, while the MLE still remains the most popularly used one in real applications. MLEs, under mild-regularity conditions enjoy nice asymptotic properties, however, the quality of their small sample performance can be often perturbed by the existence of an unknown nuisance scale parameter. We refer to Severini [13] and Berger et al. [14] among others, who critically addressed the problem of nuisance parameters. In the present work, we employ the invariance principle for eliminating the scale parameter, to get a maximal scale invariant transformation of the data coming from a shape-scale distribution.

The resulting scale invariant likelihood can be used for all kind of scale invariant inference including point and interval estimation and tests for the shape parameter. The scale invariant likelihood-based inference turned out to be much efficient than classical procedures for the commonly encountered Weibull, Gamma and GE distribution. For the Weibull and the Gamma distribution, Monte-Carlo studies based on a large number of simulations, revelled an almost exact linear relation between the bias of the proposed transformation-based MLE of shape and that of its classical MLE. Exploiting this linearity, we propose an almost unbiased estimator of the shape parameter for these two distributions. In the sequel, Section 2 presents the proposed scale invariant transformation. The resulting likelihoods are functions of only the shape parameter. The results are illustrated for popular distributions, namely the Weibull, the Gamma and the GE distribution for complete and type-II censored samples. While the proposed estimator being an MLE with respect to a proper likelihood function, enjoys all asymptotic properties under regular conditions, the Section 3 reports simulation-based small sample performance assessment of the resulting likelihood-based inference procedures and presents further refined estimation procedures. The related problem of estimation of scale is also addressed. Section 4 reports concluding remarks.

2. The proposed transformation and scale invariant inference

Throughout the article, we assume that a random sample under consideration $\mathbf{X}_n = \{X_1, X_2 \ldots, X_n\}$ comes from a shape-scale family with density $\frac{1}{a} f((x/a), b)$, $a > 0$ where $f(., b)$ is indexed by a single-shape parameter b. This section employs a maximal scale invariant transformation for eliminating the scale parameter, leading to a nuisance free likelihood for the shape parameter. In the sequel, L^* denotes the likelihood function for the transformed data.

2.1. Complete data

Let $\mathbf{x}_n = \{x_1, x_2 \ldots, x_n\}$ be n i. i. d. observations with the joint probability density function

$$\begin{aligned} \frac{1}{a^n} f((\mathbf{x}/a), b) &= L(a, b \mid \mathbf{x}) \\ &= \frac{1}{a^n} \prod_{i=1}^{n} f((x_i/a), b), \end{aligned}$$

$a, b, \mathbf{x} \geq 0$, where the density $f(., b)$ is indexed only by the shape b. Suppose the interest is in inferring the shape parameter, the scale a being a nuisance parameter. The classical MLE of the shape based on L is likely to be a complex function of the nuisance scale parameter. Consider the following transformation to eliminate the scale parameter:

$$y_i = \begin{cases} x_i/x_n & \text{for } i = 1, 2, \ldots, n-1, \\ x_n & \text{for } i = n. \end{cases}$$

The Jacobian of transformation is, $|J| = y_n^{n-1}$, and the transformed joint density for $\mathbf{y}_n \equiv \{y_1, y_2 \ldots, y_n\}$ is,

$$g(\mathbf{y}_{n-1}, (y_n/a), a, b) = \frac{1}{a^n} \prod_{i=1}^{n-1} f((y_i y_n/a), b) f((y_n/a), b)|J|.$$

Integrating over y_n, the resulting scale invariant density function for $\mathbf{y}_{n-1}$ is,

$$\begin{aligned} L^*(b \mid \mathbf{y}_{n-1}) &= g^*(\mathbf{y}_{n-1} \mid b) \\ &= \int_{y_n} g(\mathbf{y}_{n-1}, (y_n/a), a, b)\, \mathrm{d}(y_n). \end{aligned} \tag{1}$$

Note that although the scale a is appearing in the right-hand expression, the process of integration eliminates it rendering the final result free from a. Inference for the shape b based on L^* is considered in the next section. The following comments are notable:

i. Although apparently it seems as if L^* is based on only $n-1$ observations $\mathbf{y}_{n-1}$, computation of L^* is based on all the n original observations, hence L^* utilizes the entire information in the original sample of size n.
ii. Often the integrand in the LHS of (1) may not be available in the close form and needs to be numerically computed. In such cases sometime built-in functions in any software for computing the integrals are observed to give absurd results. As a way out, a simple computational trick when the support of y_n is $(0, \infty)$ is to use the importance sampling to evaluate the integral by writing

$$L^* = \int_{y_n} \{g(\mathbf{y}_{n-1}, (y_n/a), a, b) e^{y_n}\} e^{-y_n}\, \mathrm{d}(y_n),$$

which is the expected value of $h(\mathbf{y}_{n-1}, Y_n) = g(\mathbf{y}_{n-1}, (Y_n/a), a, b) e^{Y_n}$ under Y_n distributed as standard exponential distribution. Using the weak law of large numbers (WLLN) the resulting integral is then well approximated by simulating a fairly large number M (say $M = 10,000$) of standard exponential random numbers w_i, $i = 1, 2, \ldots, M$ leading to the close approximation

$$L^* \approx \frac{1}{M} \sum_{i=1}^{M} g(\mathbf{y}_{n-1}, (w_i/a), a, b) e^{w_i},$$

for fixed observed data $\mathbf{y}_{n-1}$.

iii. Since L^* is also a proper likelihood function, all kind of likelihood-based inference procedures using L^* enjoy the asymptotic properties of regular likelihood-based inference, for example, asymptotic normality of MLE and consistency among others, with an additional advantage of being nuisance scale free.

2.2. Type-II censored data

Let $\mathbf{x}_{(r)} = \{x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(r)}\}$, $(r < n)$ be a type-II censored sample from a shape-scale family $(1/a)f((x/a), b)$. Then the likelihood function is

$$L_C(a, b \mid \mathbf{x}_{(r)}) = \frac{1}{a^r} \prod_{i=0}^{r} f((x_{(i)}/a), b))\{\bar{F}((x_{(r)}/a), b)\}^{n-r},$$

where $\bar{F}((x_{(r)}/a), b)$ is the underlying survival function evaluated at $(x_{(r)}/a)$. The scale invariant transformation to be employed here is

$$y_{(i)} = \begin{cases} x_{(i)}/x_{(r)} & \text{for } i = 1, 2, \ldots, r-1, \\ x_{(r)} & \text{for } i = r, r+1, r+2, \ldots, n. \end{cases}$$

The Jacobian of transformation is $|J| = y_{(r)}^{r-1}$, leading to the transformed likelihood function

$$\begin{aligned} L_C(a, b \mid \mathbf{y}_{(r)}) &= g_C(\mathbf{y}_{(r-1)}, (y_{(r)}/a), a, b), \\ &= \frac{1}{a^r} \prod_{i=1}^{r-1} f((y_{(i)} y_{(r)}/a), b) f((y_{(r)}/a), b)\{\bar{F}((y_{(r)}/a), b)\}^{n-r} y_{(r)}^{r-1}. \end{aligned}$$

Integration over $y_{(r)}$ leads to the scale invariant likelihood function under type-II censoring

$$\begin{aligned} L_C^*(b \mid \mathbf{y}_{r-1}) &= g_C^*(\mathbf{y}_{r-1}, b), \\ &= \int_{(y_r)} g_C(\mathbf{y}_{r-1}, (y_{(r)}/a), a, b)\, \mathrm{d}y_{(r)}. \end{aligned}$$

Here, the suffix C on L^* indicates that the likelihood is under type-II censoring. If the number of uncensored observations r is equal to n then the sample is considered to be a complete sample. Comments similar to (i)–(iii) at the end of previous sub section hold for this case also.

The next subsection derives the scale invariant likelihood for popular lifetime distributions, namely the Weibull, the Gamma and the GE distributions.

3. Inference based on the transformed likelihood

The L^* (L_C^*) can be used for all kind of nuisance free likelihood-based inference about the shape parameter b for complete (type-II censored) case. In the sequel, proposed maximal scale invariant likelihood estimator (MSILE) of the shape parameter b is the maximizer $\hat{b}^*$ of the transformed likelihood L^* (L_C^*). The transformed likelihood can also be used for Likelihood ratio tests (LRT) related to the shape, and the resulting tests can also be inverted

Table 1. Linearity between bias of MLE and that of MSILE with proposed AUE.

Distribution	Relation	Proposed AUE ($\tilde{b}$)
Weibull	$E(\hat{b}^* - b) \approx 0.528E(\hat{b} - b)$	$(\hat{b}^* - 0.528\hat{b})/(1 - 0.528)$
Gamma	$E(\hat{b}^* - b) \approx 0.668E(\hat{b} - b)$	$(\hat{b}^* - 0.668\hat{b})/(1 - 0.668)$

to form interval estimates of b in the usual manner, leading to scale invariant inference for shape in these cases.

Most often a closed form expression does not exist for $\hat{b}^*$ and commonly used numerical methods can be employed for its computation. Note that the computational load in maximizing a function of a single parameter b would be much less than that of maximizing a function of two arguments as in the regular likelihood.

3.1. An improved almost unbiased estimator (AUE)

A simulation study for the Weibull and Gamma distributions revealed an almost exact relationship between the bias of $\hat{b}$ ($E(\hat{b} - b)$) and that of $\hat{b}^*$ ($E(\hat{b}^* - b)$). Exploiting this linearity an improved almost unbiased estimator $\tilde{b}$ of b is suggested based on 100 000 simulated random samples from various parametric combinations of both the distributions. The details are reported in Table 1, where the relations are not exact but were found to be very close to exact through simulations. Moreover, it is to be noted that these relations are between population biases and may not closely hold for a particular observed data set.

A linear relationship between the biases of the two estimators for GE distribution was also visible but was not sharp to the extent of producing an AUE for the shape parameter.

3.2. Examples

In the sequel, we use following notation:

$$T_s(\mathbf{y}_n, b) = \sum_{i=1}^{n} y_i^b, \quad T_s(\mathbf{y}_{(r)}, b) = \sum_{i=1}^{r} y_{(i)}^b,$$

$$T_p(\mathbf{y}_n, b) = \prod_{i=1}^{n} y_i^b, \quad T_p(\mathbf{y}_{(r)}, b) = \prod_{i=1}^{r} y_{(i)}^b.$$

Also L and L^* denote the classical and transformed likelihood functions, respectively. $\mathbf{x}$ ($\mathbf{y}$) denote the original(transformed) observations. Routine computations as per Subsection 2.1 and 2.2 yield the following transformed likelihood functions for the Weibull (W), the Gamma (G) and the GE distributions.

3.2.1. Weibull distribution

(i) Complete sample:
Regular likelihood function:

$$L_W(b, a \mid \mathbf{x}) = \left(\frac{b}{a^b}\right)^n T_p(\mathbf{x}_n, b - 1)e^{-T_s(\mathbf{x}_n, b)/a^b},$$

$$x_i, a, b > 0, \quad i = 1, 2, \ldots, n.$$

Transformed likelihood function:

$$L_W^*(b \mid \mathbf{y}_{n-1}) = \frac{\Gamma(n) b^{n-1} T_p(\mathbf{y}_{n-1}, b-1)}{(T_s(\mathbf{y}_{n-1}, b)+1)^n},$$
$$y_i,\ b > 0, \quad i = 1, 2, \ldots, n-1.$$

(ii) Type-II censored sample:
Regular likelihood function:

$$L_{W\_C}(b, a \mid \mathbf{x}) = \left(\frac{b}{a^b}\right)^r T_p(\mathbf{x}_{(r)}, b-1) e^{-T_s(\mathbf{x}_{(r)}, b)/a^b} [e^{-T_s(x_{(r)}, b)/a^b}]^{n-r},$$
$$a,\ b > 0,\ 0 \le x_{(1)} \le x_{(2)} \le \cdots \le x_{(r)}.$$

Transformed likelihood function:

$$L_{W\_C}^*(b \mid \mathbf{y}_{(r-1)}) = \frac{\Gamma(r) b^{r-1} T_p(\mathbf{y}_{(r-1)}, b-1)}{(T_s(\mathbf{y}_{(r-1)}, b) + n - r + 1)^r},$$
$$0 \le y_{(1)} \le \cdots \le y_{(r-1)} \le 1, \quad b > 0.$$

3.2.2. Gamma distribution

(i) Complete sample:
Regular likelihood function:

$$L_G(b, a \mid \mathbf{x}) = \left(\frac{1}{a^b \Gamma(b)}\right)^n T_p(\mathbf{x}_n, b-1) e^{-T_s(\mathbf{x}_n, 1)/a},$$
$$x_i, a, b > 0, \quad i = 1, 2, \ldots, n.$$

Transformed likelihood function:

$$L_G^*(b \mid \mathbf{y}_{n-1}) = \frac{\Gamma(nb) T_p(\mathbf{y}_{n-1}, b-1)}{\Gamma(b)^n (T_s(\mathbf{y}_{n-1}, b)+1)^{nb}},$$
$$y_i,\ b > 0, \quad i = 1, 2, \ldots, n-1.$$

(ii) Type-II censored sample:
Regular likelihood function:

$$L_{G\_C}(b, a \mid \mathbf{x}) = \left(\frac{1}{a^b \Gamma(b)}\right)^r T_p(\mathbf{x}_{(r)}, b) e^{-T_s(\mathbf{x}_{(r)}, 1)/a} [\bar{G}(x_{(r)}, b, a)]^{n-r},$$
$$a, b > 0, \quad 0 \le x_{(1)} \le \cdots \le x_{(r)},$$

where $\bar{G}(x_{(r)}, b, a)$ is the survival function of gamma (b,a) distribution evaluated at $x_{(r)}$, b is shape parameter and a the scale parameter.

Transformed likelihood function:

$$L^*_{G\_C}(b \mid \mathbf{y}_{(r-1)}) = \frac{T_p(\mathbf{y}_{(r-1)}, b-1)}{\Gamma(b)^r} I_1(\mathbf{y}_{(r-1)}, b),$$
$$b > 0,$$

where, $0 \le y_{(1)} \le \cdots \le y_{(r-1)} \le 1$,

$$I_1(\mathbf{y}_{(r-1)}, b) = \int_0^\infty e^{-T_s(\mathbf{y}_{(r-1)},1)u} u^{rb-1} [\bar{F}(u,1,b)]^{n-r} \, \mathrm{d}u.$$

3.2.3. GE distribution

(i) Complete sample:
Regular likelihood function:

$$L_{GE}(b, a \mid \mathbf{x}) = \left(\frac{b}{a}\right)^n e^{-T_s(\mathbf{x}_n,1)/a} \prod_{i=1}^n (1 - e^{-(x_i/a)})^{b-1},$$
$$x_i, a, b > 0, \quad i = 1, 2, \ldots n.$$

Transformed likelihood function:

$$L^*_{GE}(b \mid \mathbf{y}_{n-1}) = b^n I_2(\mathbf{y}_{n-1}, b),$$
$$y_i > 0, \quad i = 1, 2, \ldots, n-1.$$
$$b > 0,$$

where,

$$I_2(\mathbf{y}_{n-1}, b) = \int_0^\infty \prod_{i=1}^{n-1} (1 - e^{-uy_i})^{b-1} (1 - e^{-u})^{b-1} e^{-T_s(\mathbf{y}_{n-1},1)u} u^{n-1} \, \mathrm{d}u,$$

(ii) Type-II censored sample:
Regular likelihood function:

$$L_{GE\_C}(b, a \mid \mathbf{x}) = \left(\frac{b}{a}\right)^r \prod_{i=1}^r (1 - e^{-(x_{(i)}/a)})^{b-1} e^{-T_s(\mathbf{x}_{(i)},1)/a} [1 - (1 - e^{-(x_{(r)}/a)})^{b-1}]^{n-r},$$
$$a, b > 0,$$
$$0 \le x_{(1)} \le x_{(2)} \le \cdots \le x_{(r)}.$$

Transformed likelihood function:

$$L^*_{GE\_C}(b \mid \mathbf{y}_{(r-1)}) = b^r I_3(\mathbf{y}_{(r-1)}, b),$$
$$b > 0,$$

where $0 \le y_{(1)} \le \cdots \le y_{(r-1)} \le 1$ and

$$I_3(\mathbf{y}_{(r-1)}, b) = \int_0^\infty \prod_{i=1}^{r-1} (1 - e^{-y_{(i)}u})^{b-1} (1 - e^{-u})^{b-1} e^{-T_s(\mathbf{y}_{(r-1)},1)u}$$
$$\times [1 - (1 - e^{-u})^b]^{n-r} \, u^{r-1} \, \mathrm{d}u.$$

Note that, in practical usage, the integrals $I_1(.,b)$, $I_2(.,b)$ and $I_3(.,b)$ need to be numerically computed. The trick mentioned in comment (ii) of Section 2.1 can be employed for simpler computation.

4. Empirical assessment

This section reports the results of an empirical assessment of the proposed inferential procedures in comparison to the classical MLE. A total of 100,000 samples are simulated from the above mentioned 3 distributions. The parametric combinations considered are: sample sizes $n = 10,20,30,50$, shape parameters $b = 0.5, 1, 2, 5$, scale parameters $a = 1,5,10$ and censoring fractions $r = 0.5, 0.7, 1$ where $[nr]$ observations are actually observed. The quantities $r_b^*, \tilde{r}_b, r_m^*$ and $\tilde{r}_m$ defined below for quantifying the extent of reduction in bias and MSE in comparison to MLE are computed for each sample.

$$r_b^* = \left|\frac{\text{Bias } \hat{b}}{\text{Bias } \hat{b}^*}\right|, \quad \tilde{r}_b = \left|\frac{\text{Bias } \hat{b}}{\text{Bias } \tilde{b}}\right|, \quad r_m^* = \frac{\text{MSE } \hat{b}}{\text{MSE } \hat{b}^*}, \quad \tilde{r}_m = \frac{\text{MSE } \hat{b}}{\text{MSE } \tilde{b}}.$$

4.1. Assessment of the MSILE of the shape parameter

The average bias (MSE) of the three estimators namely MLE, MSILE and AUE for the Weibull and the Gamma distribution and MLE and MSILE of the GE distribution are displayed in Table S1 in the supplementary material. Figure 1 displays the box plots of the ratios r_b^* (panels (a)–(c)) and r_m^* (panels (d)–(f)) for the three distributions. All the ratios are well above 1 indicating that the biases and MSEs of MSILE are uniformly smaller than those of MLE. For GE distribution under large sample sizes with small shape parameter the precision of MSILE over MLE in terms of both bias and MSE was not notable and these cases are not included in the box plots. For small samples with high-censoring fraction a similar thing was observed with respect to MSE for the GE distribution. For Gamma distribution the precision of MSILE with respect to MSE for small sample size and high-censoring fraction was markedly large in comparison to other cases and is displayed separately in Figure 2 for better visibility. Figure 3 displays the ratios $\tilde{r}_b$ and $\tilde{r}_m$ for the

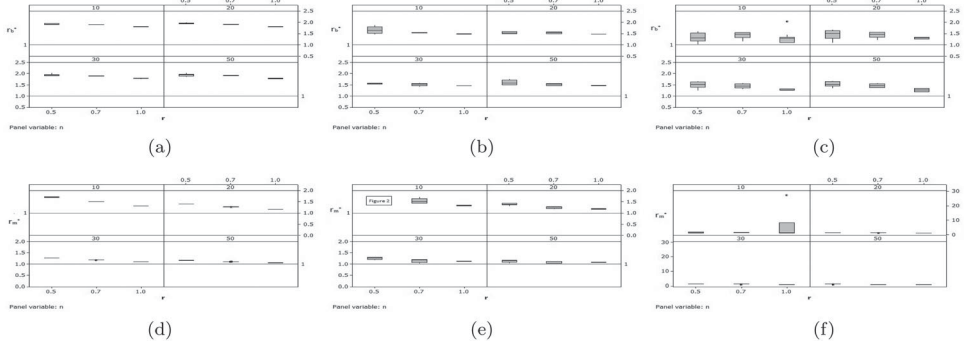


Figure 1. The extent of reduction in bias ((a)–(c)) and MSE ((d)–(f)) for MSILE in comparison to MLE: Weibull, Gamma distribution and GE distribution. (a) Weibull distribution, (b) Gamma distribution, (c) GE distribution, (d) Weibull distribution, (e) Gamma distribution and (f) GE distribution.

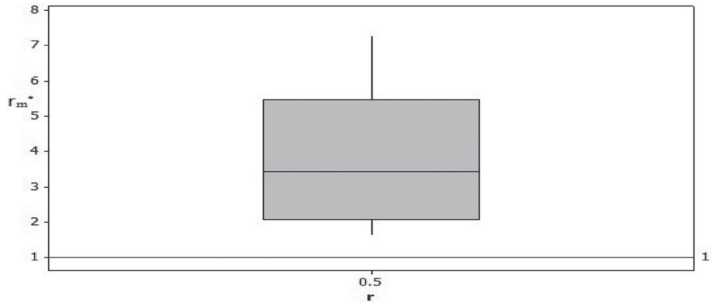


Figure 2. The extent of reduction in MSE for MSILE in comparison to MLE for Gamma distribution with sample size $n = 10$ and $r = 0.5$.

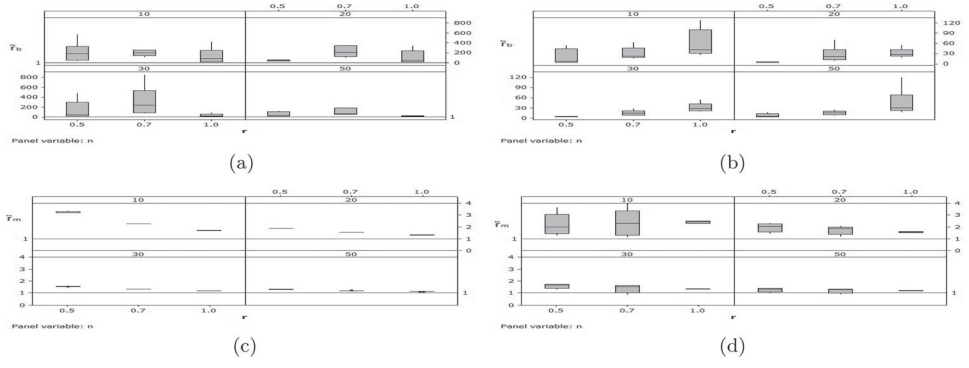


Figure 3. The extent of reduction in bias ((a) and (b)) and MSE ((c) and (d)) for AUE in comparison to MLE: Weibull and Gamma distribution. (a) Weibull distribution, (b) Gamma distribution, (c) Weibull distribution and (d) Gamma distribution.

Weibull and Gamma distributions. A similar observation of Figure 3 reveals that the AUE further uniformly and markedly refines the performance of MSILE by reducing the bias to almost zero for all sample sizes and all censoring fractions for the Weibull and Gamma distributions. The improvement was much more noticeable for the Weibull distribution. The efficiency increases with the extent of censoring for small sample sizes. Owing to the consistency of MLEs, the extent of the reduction in MSE reduces with increased sample size.

Tanaka et al. [12] suggested two improved estimators of shape parameter of Gamma distribution which exhibit superiority over MLE for the case of complete sample. In the sequel we refer these estimators as *Tanaka*_1 and *Tanaka*_2. Figure 4 shows the box plots of bias (a) and MSE (b) of improved estimators suggested by Tanaka et al. [12], MLE, MSILE and AUE for shape parameter of Gamma distribution. The sub-panels of each sub-figure therein show the box plots of bias and MSE for these estimators with different sample sizes. Figure 5 displays similar plots varying the shape parameters. The graphs reveal that the proposed AUE has uniformly smaller bias over all the estimators. The bias of *Tanaka*_1 are also close to zero at small shape parameter but comparatively larger than AUE for shape parameter greater than 2. MSEs of *Tanaka*_2 and AUE are comparable and reasonably small. However note that the estimators suggested by Tanaka et al. [12] are valid only

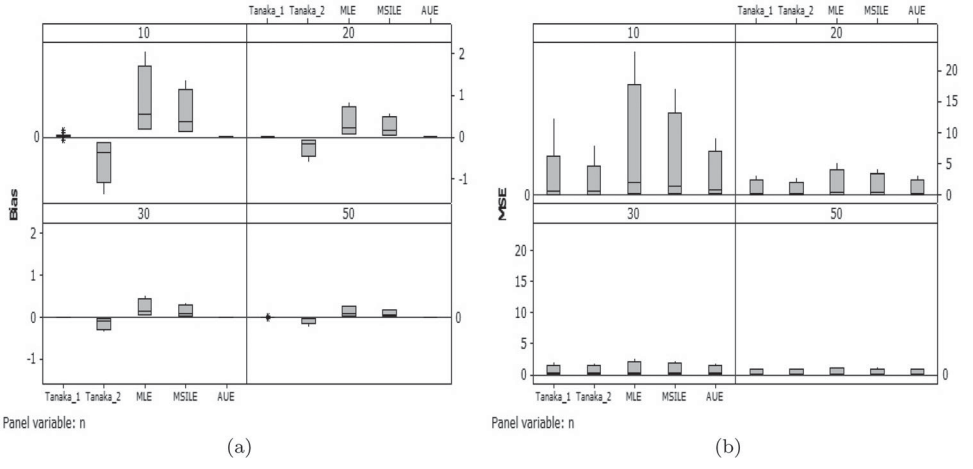


Figure 4. Sample size wise Bias (a) and MSE (b) of estimators of shape parameters of Gamma distribution.

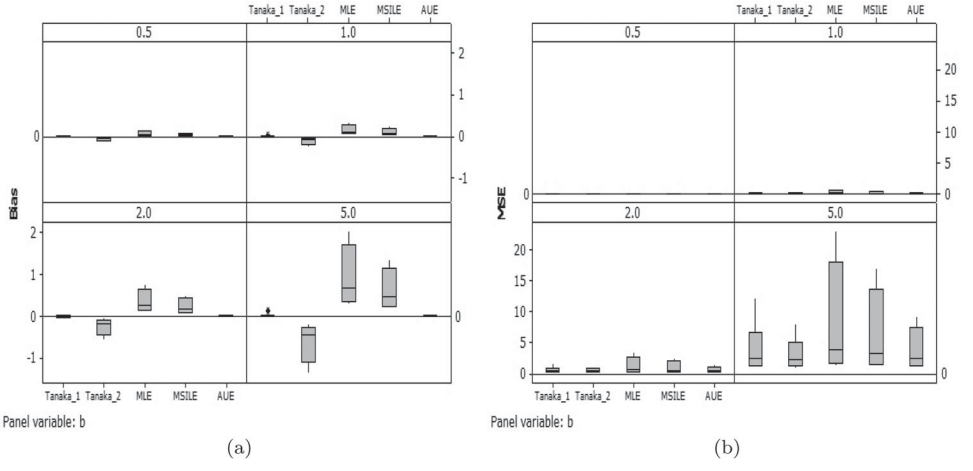


Figure 5. Shape parameter wise Bias (a) and MSE (b) of estimators of shape parameters of Gamma distribution.

under complete sample case while AUE is available for both complete sample and type-II censoring.

4.2. One sample test for the shape

LRT under the transformed (original) likelihood based on $\hat{b}^*$ ($\hat{b}$) is compared empirically based on 100,000 simulations, for the same set of parametric combinations as in subsection 4.1 at 5% level. The absolute difference (D) between observed type-I errors and the nominal level $\alpha = 0.05$ are displayed in Table S2 of the supplementary material. The box plots of the absolute difference (D) are displayed in Figure 6. The differences based on MSILE are clearly very close to zero compared to MLE for all the three distributions and all parametric

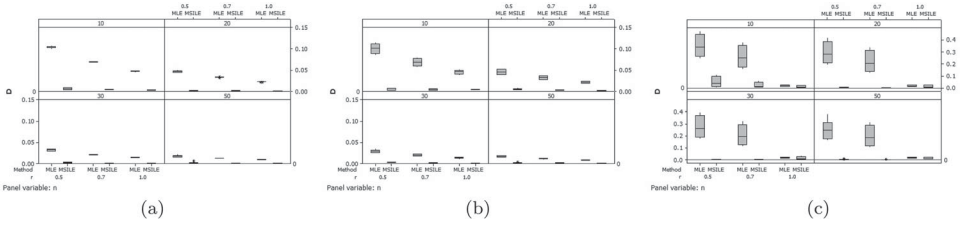


Figure 6. Box plots of absolute differences (D) between simulated type-I error (size) and actual level $\alpha = 0.05$. (a) Weibull distribution, (b) Gamma distribution and (c) GE distribution.

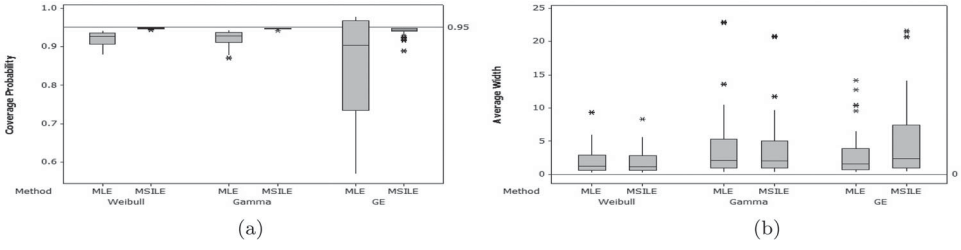


Figure 7. Coverage probability and average widths of CI for the shape parameter. (a) Coverage probability and (b) Average width.

combinations with increasing degree of efficiency with the extent of censoring indicating the superiority of MSILE for testing the shape parameter.

4.3. Interval estimation

The proposed LRT can be inverted to find a confidence interval (CI) for the shape parameter. The coverage probabilities and average widths of CI based on the MSILE and MLE are displayed in the Figure 7. It is clear that the MSILE has uniformly well concentrated coverages around the true confidence coefficient 0.95. The extent of benefit of MSILE with respect to coverage probability was more prominently seen for GE distribution although the widths in this case are little larger than those of MLE.

4.4. Assessment of the scale parameter

Note that for a fixed shape b, the MLE of the scale parameter a is a function of b. Let it be denoted by $\hat{a}(b)$. Similarly, let $\hat{a}_{\mathrm{KS}}(b)$ denote the estimator obtained by minimizing the Kolmogrov–Smirnov distance between $F(., \hat{a}, b)$ and the empirical distribution function $F_n(.)$ for fixed b. Let $\hat{a}_{\mathrm{CV}}(b)$ and $\hat{a}_{\mathrm{AD}}(b)$ denote similar estimators based on the Crammer–Von Mises and the Anderson–Darling distances, respectively. We compare the following point estimators for the scale parameter a empirically:

(1) M1–M3: $\hat{a}(\hat{b})$, $\hat{a}(\hat{b}^*)$ and $\hat{a}(\tilde{b})$, respectively.
(2) M4–M6: $\hat{a}_{KS}(\hat{b})$, $\hat{a}_{KS}(\hat{b}^*)$, $\hat{a}_{KS}(\tilde{b})$, respectively.
(3) M7–M9: $\hat{a}_{CV}(\hat{b})$, $\hat{a}_{CV}(\hat{b}^*)$, $\hat{a}_{CV}(\tilde{b})$, respectively.
(4) M10–M12: $\hat{a}_{AD}(\hat{b})$, $\hat{a}_{AD}(\hat{b}^*)$, $\hat{a}_{AD}(\tilde{b})$, respectively.

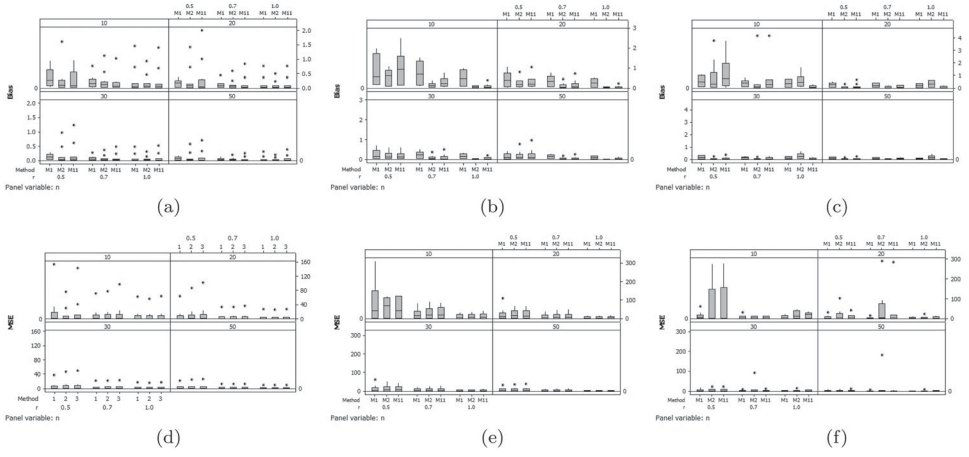


Figure 8. Bias and MSE of the three estimators (*M*1, *M*2 and *M*11) of the scale parameter in the three distributions. (a) Weibull distribution, (b) Gamma distribution, (c) GE distribution, (d) Weibull distribution, (e) Gamma distribution and (f) GE distribution.

Out of these 12 estimators, box plots of the three estimators having smallest simulated absolute bias and MSE are given in Figure 8 for the Weibull, the Gamma and the GE distribution. The Anderson–Darling minimum distance estimator (M11) based on MSILE $\hat{b}^*$, classical MLE based on $\hat{b}$ (M1) and the one based on $\hat{b}^*$ (M2) were found to exhibit the smallest bias. The MSE for all the 12 estimators was comparable. The performance of M2 is satisfactory with the Weibull and Gamma distributions and is recommended for point estimation of the scale. For GE distribution a similar observation leads to the recommendation of M1 under small samples and M11 under moderate to large samples.

The proposed procedure is illustrated with real life examples in the sequel.

4.5. Real life application

Krishnamoorthy et al. [5], Powar and Kulkarni and SenGupta et al. [7] among others discussed the importance of shape-scale family of distributions in ground water monitoring, assessment of air pollution and prediction of environmental events as well. In the context of ground water monitoring, Krishnamoorthy et al. [5] and Powar and Kulkarni [6] analysed vinyl chloride concentration in micro grams per litre of water (μg/L) from 34 clean upgradient wells with observed values: 5.1, 2.4, 0.4, 0.5, 2.5, 0.1, 6.8, 1.2, 0.5, 0.6, 5.3, 2.3, 1.8, 1.2, 1.3, 1.1, 0.9, 3.2, 1.0, 0.9, 0.4, 0.6, 8.0, 0.4, 2.7, 0.2, 2.0, 0.2, 0.5, 0.8, 2.0, 2.9, 0.1, 4.0. The nominal level of vinyl chloride suggested by U.S. Environmental Protection Agency is 2.0 to 2.4 μg/L. Note that increased percentage of vinyl chloride is a major cause for cancer or liver damage. The *p*-values based on KS statistics for fitting Gamma and Weibull distribution are 0.9694 and 0.9366, respectively with respective Akaike Information Criteria (AIC) values 114.8263 and 114.8992. As per the minimum AIC criteria and maximum *p*-value of KS-test, the given data is best with fitted Gamma distribution. The estimated maximum likelihood parameters are: $\hat{b} = 1.0627$; and $\hat{a} = 1.7685$; and the proposed estimates are $\hat{b}^* = 1.0381$;, $\tilde{b} = 0.9887$; and $\hat{a}^* = 1.8104$;, leading to an estimate of the percentage of wells having vinyl chloride concentration greater than the prescribed upper bound of 2.4

is 26.16%, that is almost $\frac{1}{4}$th of the wells have critically large percentage of vinyl chloride, indicating that monitoring of this wells is essential to avoid future risks.

5. Concluding remarks

The maximal invariant transformation-based likelihood inference for the shape parameter has exhibited uniform marked improvement over their regular likelihood-based counterparts under small samples and and high-censoring fractions and is recommended as a substitute for MLE point and interval estimation as well as testing problem. The proposed AUE for the Weibull and the Gamma distribution further improves the scenario. MLE of the scale as a function of MSILE of shape also turns out to be more efficient than its regular MLE under Weibull and Gamma distributions and is recommended.

Acknowledgments

We would like to thank two referees and associate editor for insightful comments to greatly improvement of this manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The Authors are very much thankful to Council of Scientific and Industrial Research (CSIR), India, for financial support for this research work, under the grant sanction Order No. 25(0211)/13/EMR-II.

References

[1] Rinne H. The Weibull distribution: a handbook. Boca Raton: Taylor and Francis Group; 2009; LLC, ISBN 978-1-4200-8743-7.
[2] Abernethy RB. The New Weibull Handbook, Library of Congress Cataloging in Publication Data, ISBN 0-9653062-3-2;2010.
[3] McCool JI. Using the Weibull distribution: reliability, modelling and inference. Hoboken: John Wiley and Sons, Inc.; 2012; ISBN 978-1-118-21798-6 (cloth).
[4] Jiang R, Murthy DNP. A study of Weibull shape parameter: properties and significance. Reliab Eng Syst Saf. 2011;96:1619–1626.
[5] Krishnamoorthy K, Lin Y, Xia Y. Confidence limits and prediction limits for a Weibull distribution based on the generalized variable approach. J Stat Plan Infer. 2009;139:2675–2684.
[6] Powar SK, Kulkarni HV. Estimation of confidence interval for hydrological design value for some continuous distributions under complete and censored samples. Stoch Environ Res Risk Assess. 2015;29:1691–1708. doi:10.1007/s00477-015-1022-8
[7] SenGupta A, Kulkarni HV, Hubale UD. Prediction intervals for environmental events based on Weibull distribution. Environ Ecol Stat. 2015;22(1):87–104. doi:10.1007/s10651-014-0286-3
[8] Bagdonavičius V, Nikulin M, Zerbet A. On outliers detection for location-scale and shape-scale families. J Math Sci. 2017;225(5):723–732.
[9] Patil KP, Kulkarni HV. On the interval estimation of stress–strength reliability for exponentiated scale family of distributions. Qual Reliab Eng Int. 2017;33(7):1447–1453. doi:10.1002/qre.2117

[10] Kulkarni HV, Patil KP. Two sample comparisons including zero inflated continuous data: a parametric approach with applications to microarray experiment. Math Biosci. 2018;298:19–28. doi:10.1016/j.mbs.2018.01.009
[11] Zaigraev A, Podraza-Karakulska A. On estimation of the shape parameter of the gamma distribution. Stat Probabil Lett. 1998;78(3):286–295. doi:10.1016/j.spl.2007.07.003
[12] Tanaka H, Pal N, Lim WK. On improved estimation of a gamma shape parameter. Statistics. 2014;49(1):84–97. doi:10.1080/02331888.2014.915842
[13] Severini TA. Likelihood functions for inference in the presence of a nuisance parameter. Biometrika. 1998;85(3):507–522.
[14] Berger JO, Liseo B, Wolpert RL. Integrated likelihood methods for eliminating nuisance parameters. Stat Sci. 1999;14(1):1–28.

Steady-State Behavior of Nonparametric Synthetic Control Chart Using Signed-Rank Statistic

V. Y. Pawar
Department of Statistics, PDVP College, Tasgaon, (MS) India
vypawar.stats@gmail.com

D.T. Shirke
Department of Statistics, Shivaji University, Kolhapur, (MS) India 416004
dts_stats@unishivaji.ac.in

S.K. Khilare
Department of Statistics, R. B. N. B. College, Shrirampur, (MS) India 413709
shashi.khilare@gmail.com

Abstract

The article studied the steady-state behaviour of the synthetic control chart using signed-rank statistic for shifts in the process median. The steady-state ATS (Average Time to Signal) values are computed using Markov chain approach. To compute steady-state ATS, the performance of the synthetic control chart and two-of-L+1 control chart can be made identical over all samples with head start features. When subgroup sample size n=10, the steady-state performance of the synthetic control chart is worth for small to moderate shifts under all considered symmetric distributions. When subgroup sample size n=5, steady-state ATS values are larger under normal and double exponential distributions only for small shifts. However, under the Cauchy distribution zero-state ATS values are larger but not significantly larger as compared to steady-state ATS values. Usefulness of proposed control chart explored using numerical example. Proposed control chart is simple and easy to use for practitioners.

Keyword: Nonparametric, signed-rank, synthetic, runs rule, steady-state and average time to signal.

1. Introduction

A control chart is one of the most useful tools for monitoring quality of the characteristic of an interest in a manufacturing process. Most of the control charts are based on the assumption that the process characteristic follows a normal distribution. Many researchers have pointed out that all the processes are not normally distributed; see for example (Chou et al. 2001) and the references cited therein. The standard control charts do not perform well, if the assumption of normality is not satisfied. The effects of non-normality on the $\overline{X}$ chart have been studied in the literature and includes among others (Schilling and Nelson 1976, Bradley 1973). This demands the construction of nonparametric control charts. A chart is said to be nonparametric if the run length distribution of the chart does not depend on the underlying process distribution, when there is no shift in the process parameter under study. Hence, the in-control Average Time to Signal (ATS) of nonparametric control chart does not depend on the underlying process distribution.

In the review of literature related to the nonparametric control charts, (Bakir and Reynolds 1979) provided a control chart based on within group ranking. (Hackl and

Ledolter 1991) suggested a control chart based on ranks. (Amin et al. 1995) proposed nonparametric quality control charts for location and scale parameters based on the sign statistic. (Bakir 2004) reported a control chart based on signed-rank statistic, which was further improved in terms of Average Run Length (ARL) by (Chakraborti and Eryilmaz 2007). (Bakir 2006) proposed distribution free quality control charts based on signed-rank-like statistics. (Chakraborti and Van de Wiel 2008) proposed Mann-Whiteny statistic based control chart. (Human et al. 2010) studied nonparametric Shewhart-type sign control charts based on runs. (Khilare and Shirke 2010, 2012) developed nonparametric synthetic control charts using sign statistics for shifts in location and variability respectively. (Ho and Costa 2011) proposed monitoring a wandering mean with an np chart and this chart also works with sign statistic. (Yang et al. 2011) provided a new EWMA Control Chart based on a simple statistic to monitor the small mean shifts in the process with non-normal or unknown distributions. (Majid and Neda 2013) developed nonparametric signed-rank control charts with variable sampling interval. (Abbasi et al. 2013) proposed nonparametric progressive mean control chart for monitoring the process target. (Liu et al. 2014) developed dual nonparametric cusum control chart based on ranks. (Riaz and Abbasi 2016) suggested double EWMA control chart for process monitoring. (Abid et al. 2016) reported the use of ranked set sampling in nonparametric EWMA control charts based on sign test statistic. (Abid et al. 2016) proposed nonparametric EWMA control chart based on Wilcoxon signed-rank statistic for monitoring location. (Coelho et al. 2017) reported nonparametric signed-rank control charts with variable sampling intervals.

If process is running in an in-control state for a long period, it will reach in steady-state mode. In order to characterize long-term properties of a control chart, it is an appropriate to investigate the steady-state ARL. (Crosier 1986) suggested a technique for obtaining steady-state ARL of CUSUM chart using the Markov chain approach. (Saccucci and Lucas 1990) have given a FORTRAN computer program for the computation of ARL of EWMA and combined Shewhart-EWMA control schemes. The program calculates zero-state and steady-state ARL using the Markov chain approach. (Champ 1992) computed steady-state ARL of Shewhart control chart with supplementary runs rules. (Davis and Woodall 2002) studied the steady-state properties of synthetic control chart to monitor shifts in process mean. (Lim and Cho 2009) developed a control charts with m-of-m runs rules to study the economical-statistical properties of control chart using steady-state ARL. (Khilare and Shirke 2015) studied the steady-state behavior of nonparametric control charts using sign statistic.

In present article, we proposed the synthetic control chart using runs rules for monitoring the median of a continuous characteristic of the underlying process. The main purpose of the paper is to study the steady-state behavior of the synthetic control chart based on signed-rank statistic when process runs in an in-control state for long time. Rest of paper is organized as follows.

Section 2 gives a control chart based on the signed-rank statistic. In section 3 conforming run length control chart is described briefly. Section 4 gives the operations and design procedure of the synthetic control chart using signed-rank statistic. Section 5 gives runs rule representation, Markov chain model and steady-state ATS of the synthetic control

chart. The steady-state performance of the synthetic control chart is given in section 6. Section 7 gives numerical example. Concluding remarks are given in section 8.

2. A Control Chart Based On the Signed-Rank Statistic:

Let (X_{t1}, X_{t2} , ….. , X_{tn}) be a random sample (subgroup) of size $n>1$ observed from a continuous process with median θ at sampling instances t =1, 2, …..**.** It is assumed that the underlying process distribution is continuous symmetric and that the in-control process median is known or specified to be equal to θ_0. We further assume that θ_0 is known and when $\theta \neq \theta_0$ the process is out-of-control. (Bakir 2004) provided a nonparametric control chart based on the signed-rank statistic. For the t^{th} subgroup sample (x_{t1}, x_{t2}, ... ,x_{tn}), the signed- rank statistic is defined as

$$\psi_t = \sum_{j=1}^{n} sign(x_{tj} - \theta_0) R_{tj}^{+}, \qquad t = 1, 2, \ldots \tag{1}$$

Where, sign (u) = -1, 0, 1 if $u< 0, = 0, > 0$ and

$$R_{tj}^{+} = 1 + \sum_{i=1}^{n} I(|x_{ti} - \theta_0| < |x_{tj} - \theta_0|)$$

with $I(a < b) = 1$ if $a < b$ and 0 otherwise.

We can rewrite (1) as

$$\psi_t = 2w_t^{+} - \frac{n(n+1)}{2}. \tag{2}$$

Where w_t^{+} is the well-known Wilcoxon Signed-rank Statistic (the sum of the ranks of the absolute values of the deviations corresponding to the positive deviations). One can therefore use ψ_t given in (2) as a charting statistic instead using (1). Let *UCL* be the upper control limit corresponding to a positive-sided control chart. The chart gives an out-of-control signal at the first sampling instance t for which $\psi_t \geq UCL$. In the following section we briefly describe conforming run length control chart.

3. The Conforming Run Length Control Chart

The conforming run length (*CRL)* chart was originally developed for attribute quality control by (Bourke 1991). In 100% inspection, the *CRL* is the number of inspected units between two consecutive nonconforming units (including the ending nonconforming unit). The *CRL* chart uses the *CRL* as the charting statistic. The idea behind the *CRL* chart is that the conforming run length will change when the fraction nonconforming '*p*' in the process changes. The *CRL* is shortened as *p* increases and lengthened as *p* decreases. The charting statistic (*CRL*) follows a geometric distribution with parameter *p*. The mean value of *CRL* (i.e. the average number of inspected units in a *CRL* sample) is

$$\mu_{CRL} = \frac{1}{p}, \tag{3}$$

and its cumulative distribution function (c. d. f.) is given by,

$$F_p(CRL) = 1 - (1-p)^{CRL}; \quad CRL = 1,2,... \tag{4}$$

If our only concern is the detection of an increase in *p*, the lower control limit (denoted *L*) is sufficient for the *CRL* chart. If α_{CRL} is the specified/desired type I error of the *CRL* chart and p_0 is the in-control fraction nonconforming, *L* can be derived from the following equation.

$$\alpha_{CRL} = F_{p_0}(L) = 1-(1-p_0)^L,$$

which gives $$L = \frac{\ln(1-\alpha_{CRL})}{\ln(1-p_0)}. \quad (5)$$

Note that L must be rounded to the largest integer smaller than or equal to the calculated value in (5). If the sample *CRL* (i.e. the charting statistic) is smaller than or equal to *L* then it is very likely that the fraction nonconforming *p* has increased and therefore, an out-of-control signal will be given. ARL_{CRL} is the average number of CRL samples required to detect change in *p*. The ARL_{CRL} is given by

$$ARL_{CRL} = \frac{1}{P[(Unit\ is\ nonconfor\min g)(CRL\ betwwen\ two\ nonconfor\min g\ units \le L)]}, \quad (6)$$

$$ARL_{CRL} = \frac{1}{p.P(CRL \le L)} = \frac{1}{p.F(L)}.$$

Where,

$$p = P(Unit\ is\ nonconfor\min g) \text{ and } F(L) = 1-(1-p)^L.$$

Therefore,

$$ARL_{CRL} = \frac{1}{p.\left(1-(1-p)^L\right)} \quad (7)$$

In section 4 we briefly discuss synthetic control chart using signed-rank statistic.

4. A Nonparametric Synthetic Control Chart

A nonparametric synthetic control chart proposed by (Pawar and Shirke 2010) is a combination of the nonparametric signed-rank statistic ψ_t (called the ψ_t chart hereafter) and the *CRL* chart. Basically the operation of the nonparametric synthetic control chart is similar to that of the synthetic control chart for monitoring the process mean as was proposed by (Wu and Spedding 2000), except that the subgroup mean is replaced by the signed-rank statistic ψ_t and the upper control limit is changed accordingly. However, we do not follow the same design procedure due to (Wu and Spedding 2000) in order to ensure that the synthetic control chart is nonparametric.

4.1. Operations

The operations of the nonparametric synthetic control chart are as follows:

1. Decide on the upper control limit of the ψ_t chart and the lower limit *L* of the *CRL* chart. The design of these control parameters will be described shortly.
2. At each inspection point '*t*' take a random sample of *n* observations and calculate ψ_t.
3. If ψ_t<*UCL*, (the sample is called a conforming sample) then control flow goes back to step (2) (That is continue to draw random samples from the process and calculate the statistic ψ_t).Otherwise, the sample is called a nonconforming sample and control flow goes to the next step.
4. Check the number of samples between the current and the last nonconforming sample (including the current sample). This number is taken as the value of the plotting statistic (i.e. *CRL*) of the *CRL* chart in the synthetic chart.
5. If this *CRL* is larger than the lower control limit of the *CRL* chart, then the process is thought to be under control and the charting procedure is continued. Otherwise, the process is declared to be out of control and control flow goes to the next step.
6. Take the necessary action to find and remove the assignable cause(s).

4.2. ARL of the synthetic control chart:

The probability that a synthetic control chart produces an out-of-control signal is given by

$$Q(\delta) = p(\delta).P(\delta),$$

where,

$$p(\delta) = P(sample\, sample\, is\, nonconfor \min g),$$
$$p(\delta) = P(\psi_t \geq UCL / \theta = \theta_0 + \delta), \text{ and}$$
$$P(\delta) = P(CRL\, between\ \ two\ \ nonconfor \min g\ samples \leq L),$$
$$P(\delta) = P(CRL \leq L),$$
$$P(\delta) = \left(1 - (1 - p(\delta))^L\right).$$

Hence, ARL of synthetic control chart is given by

$$ARL_S(\delta) = \frac{1}{Q(\delta)} = \frac{1}{p(\delta)\left(1 - (1 - p(\delta))^L\right)}.$$

4.3. Design

The synthetic chart has two parameters namely, *L* and *UCL*. For given in-control *ARL* and subgroup sample size *n*, the parameters *L* and *UCL* are obtained as follows.

Let $ARL_S(\delta)$ be the out-of-control ARL of the synthetic control chart and can be obtained using formula given below.

$$ARL_S(\delta) = \frac{1}{p(\delta)\left(1 - (1 - p(\delta))^L\right)}.$$

Here $p(\delta)$ is the probability that the sample is nonconforming, when the permanent upward step shift of δ units occurs in the process. When there is no shift, δ is equal to zero. We note that in equation (7), 'p' is the probability that a unit is nonconforming, while $p(\delta)$ defined above is the probability that the sample is nonconforming. Thus $p(\delta)$ plays the role of p in equation (7).

We note that the in-control *ARL* of the synthetic chart is given by *ARLs(0)*, where

$$ARLs(0) = \frac{1}{p(0)\left(1-(1-p(0))^{L}\right)}. \qquad (8)$$

Suppose the desired in-control *ARL* is *ARL(0)* and the subgroup sample size is n. We compute the *ARLs(0)* values using equation (8) for $UCL=1,2,\ldots,(n(n+1)/2)$ and L=1,2,…… and choose that pair of (L, UCL) for which the *ARLs(0)* is close to *ARL(0)*. We may note that for a fixed value of UCL, *ARLs(0)* is a decreasing function of L, while for a fixed value of L, *ARLs(0)* is a non-decreasing function of UCL. Table 1 gives values of *ARLs(0)* for n =5. As an example, suppose we wish to set *ARL(0)* at 32. Then, from Table 1, we see that L=4 and UCL=10 is the required pair as the *ARLs(0)* corresponding to these values is 32.77.

Table 1: In control ARL values for positive sided chart for various values of UCL and L then n=5

	L									
UCL↓	1	2	3	4	5	6	7	8	9	10
1	4	2.67	2.29	2.13	2.06	2.03	2.02	2.01	2	2
2	6.06	3.8	3.11	2.81	2.66	2.57	2.53	2.5	2.48	2.48
3	6.06	3.8	3.11	2.81	2.66	2.57	2.53	2.5	2.48	2.48
4	10.24	6.07	4.74	4.12	3.78	3.58	3.45	3.37	3.31	3.28
5	10.24	6.07	4.74	4.12	3.78	3.58	3.45	3.37	3.31	3.28
6	20.9	11.73	8.74	7.29	6.45	5.92	5.56	5.31	5.13	4.99
7	20.9	11.73	8.74	7.29	6.45	5.92	5.56	5.31	5.13	4.99
8	40.96	22.22	16.03	12.98	11.18	10.01	9.2	8.61	8.17	7.83
9	40.96	22.22	16.03	12.98	11.18	10.01	9.2	8.61	8.17	7.83
10	113.78	59.69	41.71	32.77	27.44	23.91	21.42	19.57	18.15	17.03
11	113.78	59.69	41.71	32.77	27.44	23.91	21.42	19.57	18.15	17.03
12	256	132.13	90.9	70.32	58.01	49.83	44.02	39.67	36.32	33.65
13	256	132.13	90.9	70.32	58.01	49.83	44.02	39.67	36.32	33.65
14	1024	520.13	352.23	268.32	218.01	184.49	160.58	142.67	128.75	117.64
15	1024	520.13	352.23	268.32	218.01	184.49	160.58	142.67	128.75	117.64

5. Runs Rule Representation of the Synthetic Control Chart

The runs rule representation of synthetic control chart to detect shifts in the location parameter for $\overline{X}$ control chart has been studied by (Davis and Woodall 2002). This section presents the runs rule representation of a nonparametric synthetic control chart using signed-rank statistic. For the runs rule representation of the proposed nonparametric synthetic control chart using sign-rank statistic, the procedure of (Davis and Woodall 2002) is followed. Let '0' denotes conforming sample and '1' denotes nonconforming sample. If value of signed-rank statistic falls within control limit, the sample is conforming and if it falls out-side the control limit then sample is nonconforming. Thus a sequence of ψ_t can be represented by a string of '0' and '1'. For example 100100 would indicate that in a sequence of six samples, the first and third samples are nonconforming samples, while the rest are conforming. For simplicity, suppose that L of CRL chart is equal to 4. This means that any sequence of ψ_t with pattern 10001, 1001, 101 or 11 will generate an out-of-control signal for the synthetic chart. In general, such sequence also generates signal under the following runs rule:

If two out-of-$L+1$ consecutive signed-rank statistics fall out-side of the control limit, the control chart signals an out-of-control status.

On initial pattern of 0001, the synthetic control chart will signal using $L=4$, while two-of-$L+1$ control chart would not. The performance of both the control charts can be made identical over all the samples using head start feature in the runs rule representation; that is , it is assumed that the there is signed-rank statistic at time zero and that falls out-side of the control limit. With this head start, both control charts will signal on initial patterns 1, 01, and 001 but not on the initial pattern 0001.Thus, performance of the synthetic and two-of-L+1 charts is now identical for all possible sequences of ψ_t. If CRL value is less than or equal to L, then declare that the process is out-of-control. Thus, the synthetic control chart using ψ_t identical to the above runs rule with the head start a ψ_t at time zero is observed and is nonconforming. In the following, we present the Markov chain model and ATS results of synthetic control chart.

5.1. Steady-State Average Time to Signal of the Synthetic Control Chart:

The steady-state ARL of the proposed synthetic control chart can be obtained using the Markov chain approach. The states of transition probability matrix (t.p.m.) are based on the lower control limit of the CRL chart.

Consider the case where $L=3$. This chart is an identical to a chart which signals if two-of-four signed-rank statistics fall out-side of the control limit, assuming that a signed-rank statistic at time zero is out-side of control limit.

Let

A= Pr(next observed signed-rank statistic will be below upper control limit).

The probability of next observed signed-rank statistic will be lies below the upper control limit is

$$A = \Pr(\psi_t \le UCL),$$

and $B = 1 - A.$

Davis and Woodall (2002) suggested that the following t.p.m. would govern the Markov chain for the synthetic control chart.

- The row contains 'A' in first column and 'B' in second column.
- The last row contains 'A' in first column.
- In all other rows, the entry above the diagonal is 'A'.
- In all other locations, the entry is zero.

Table 2: The transition probability matrix for the synthetic control chart using signed-rank statistic when L= 3

States $\downarrow \rightarrow \rightarrow$	000	001	010	100	Signal
000	A	B	0	0	0
001	0	0	A	0	B
010	0	0	0	A	B
100	A	0	0	0	B
Signal	0	0	0	0	1

With this Markov chain model, the zero-state ARL (0SARL) is

$$0SARL = s'(I - Q)^{-1}1, \tag{9}$$

hence, zero-state average time to signal (0SATS) is given by

$$0SATS = (0SARL - 0.5) * h, \tag{10}$$

where, Q is an $L+1$ by $L+1$ matrix of probabilities obtained by deleting last row and last column from the above matrix, 1 is column vector of appropriate order having all elements unity and I is an $L+1$ by $L+1$ identity matrix, s is an initial probabilities of an order $L+1$, 1 for initial state and 0 for the rest of the cases, s'= [0, 1, 0,…, 0, 0]. A state '001' is an initial state.

If the process is running smoothly for a longer time, it reaches in the steady-state. Therefore, it is necessary to study steady-state behaviour of the process. To study the steady-state performance of the proposed synthetic control chart, the measure average time to signal (ATS) is used. The steady-state average time to signal (SSATS) measures average number of samples required to signal when the effect of head start has disappeared.

Let Q_0 be the stochastic matrix obtained from matrix Q. Let π be a row vector corresponding to the stationary probability distribution of Q_0. The SSARL of the synthetic chart using sign-rank statistic is given by

$$SSARL = \pi'(I - Q_0)^{-1}1. \tag{11}$$

The π can be obtained as

$$\pi = Q_0'\pi,$$

subject to constraint

$$\sum_{i=1}^{n} \pi_i = 1.$$

Finally SSATS is given by,

$$SSATS = h\left(SSARL - \frac{1}{2}\right). \tag{12}$$

Where, sampling interval (h) is adjusted according to the desired false alarms rate.

We provide steady-state performance of the synthetic control chart in the following section.

6. Steady-State Performance of the Synthetic Control Chart

When there is a shift in the process median, the distribution of the charting statistic is difficult to obtain. Therefore, we use simulation to obtain the ATS values for various shifts in the process median. A simulation study based on 10000 runs is performed for sample of sizes *n*=5, n=10 and the corresponding in-control ATS values are 32 and 380 respectively for computing probabilities of next observed signed-rank statistic will falls below upper control limit for different shifts. The simulation study is carried out for three continuous symmetric distributions namely the normal, double exponential and Cauchy. As in (Bakir 2004), the scale parameter is set to be $\lambda = 1/\sqrt{2}$ for the double exponential distribution to achieve a standard deviation of 1.0. For the Cauchy distribution, $\lambda = 0.2605$ is chosen to achieve a tail probability of 0.05 above θ + 1.645, the same as that of a normal distribution with a mean θ and a standard deviation of 1.0. These three distributions are continuous symmetric about their median but have different tail behavior. Moreover, the tail probabilities, say above 3 are 0.0013499, 0.007185 and 0.0275707, while the tail probabilities above 4 are 0.00003167, 0.0017467 and 0.0207007 respectively for the normal, the double exponential and the Cauchy distributions. In most of times practitioners are interested only in upward shifts in the process median; therefore, in this paper we computed zero-state and steady-state ATS values only for up-ward shifts. Similarly we can compute zero-state and steady-state ATS values for down ward shifts and two-sided shifts in process median. Table 3 and Table 4 give the zero-state and steady-state ATS profile of the synthetic control chart to detect upward shifts in the process median.

Table 3: Zero-state and steady-state ATS values of the synthetic control chart with n=5, L=4, ARL(0)= 32.77 and UCL=10.

$(\theta-\theta_0)$	Normal distribution		Double exponential distribution		Cauchy distribution	
	0SATS	SSATS	0SATS	SSATS	0SATS	SSATS
0	32.77	32.77	32.77	32.77	32.77	32.77
0.2	10.06	10.51	7.45	7.79	3.7	3.69
0.4	4.51	4.60	3.17	3.08	1.92	1.59
0.6	2.49	2.28	1.96	1.64	1.53	1.09
0.8	1.74	1.36	1.52	1.08	1.36	0.87
1	1.38	0.90	1.29	0.79	1.26	0.74
1.2	1.19	0.66	1.16	0.62	1.22	0.69

From Table 3 we observed that:

- For normal and double exponential distributions the steady-state ATS values are large as compared to zero-state ATS values only for small shifts in median.
- For Cauchy distribution zero-state ATS values are large as compared to the steady-state ATS values but not significantly different.

Table 4: Zero-state and steady-state ATS values of the synthetic control chart with n=10, L=8, ARL(0)= 380 and UCL=40

$(\theta-\theta_0)$	Normal distribution		Double exponential distribution		Cauchy distribution	
	0SATS	SSATS	0SATS	SSATS	0SATS	SSATS
0	380.00	380.00	380.00	380.00	380.00	380.00
0.2	37.98	43.86	21.12	25.48	6.44	8.28
0.4	7.56	9.67	4.43	5.69	2.10	2.54
0.6	2.64	3.29	1.87	2.21	1.36	1.51
0.8	1.32	1.46	1.13	1.21	1.12	1.19
1	0.85	0.85	0.82	0.81	0.95	0.98
1.2	1.15	1.18	1.17	1.21	1.38	1.50

Following are the findings from Table 4:

- When subgroup sample size n=10, the steady-state ATS performance is worth as compared to the zero-state ATS for all considered distributions.
- Steady-state ATS performance of the synthetic control chart is better under Cauchy distribution than double exponential and normal distributions.
- Steady-state ATS performance of the synthetic control chart is worth for normal distribution.

7. Numerical Example

The operations of the proposed control chart can be illustrated using data related to the diameter of casting taken from Montgomery-2009 (Exercise example no.-6.69, page no.-286). The data set contains 20 samples each of size five. The median of the data set is to be 11.7531. To have an in-control ARL equal to 32, the parameters of the upper-sided synthetic control chart are UCL=10 and L = 4. A sample is conforming one when $\psi_t < UCL$. Table 5 depicts the values of the signed-rank statistic defined in equation (1) for 20 samples. Figure 1 gives the upper-sided synthetic control chart using signed-rank statistic. The synthetic control chart signals an out-of-control status, if $CRL \leq L$. Figure 1 show that the signed-rank statistic of sample two is plotted above UCL. That is sample two is nonconforming and CRL at this time epoch is 2 which is less than L; hence, synthetic control chart signals an out-of-control status at time epoch 2. The synthetic control chart also signals at time epochs 13, 17 and 19.

Table 5: Sample number and Signed-rank statistic

Sr. No.	Sign-rank statistic
1	-11
2	13
3	5
4	5
5	-15
6	5
7	-15
8	-8
9	15
10	-1
11	-8
12	-9
13	15
14	-15
15	-6
16	-9
17	15
18	-1
19	15
20	9

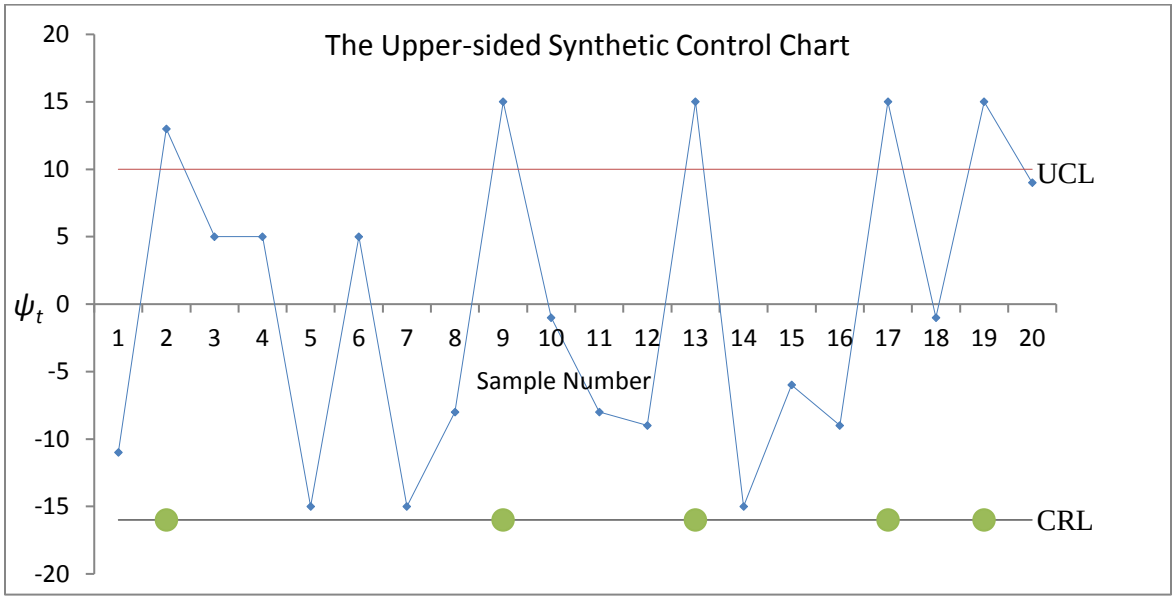


Figure 1: The upper-sided synthetic control chart

8. Conclusions

In this article we studied the steady-state behaviour of the synthetic control chart using signed-rank statistic for shifts in the process median. The steady-state ATS values are computed using Markov chain approach. To compute steady-state ATS, the performance of the synthetic control chart and two-of-L+1 control chart can be made identical over all samples with head start features. When subgroup sample size n=10, the steady-state performance of the synthetic control chart is worth for small to moderate shifts under all considered symmetric distributions. When subgroup sample size n=5, steady-state ATS values are larger under normal and double exponential distributions only for small shifts. However, under the Cauchy distribution zero-state ATS values are larger but not significantly larger as compared to steady-state ATS values. Usefulness of proposed control chart explored using numerical example. Proposed control chart is simple and easy to use for practitioners.

References

1. Abbasi S. A., Miller A. and Riaz M. (2013). Nonparametric progressive mean control chart for monitoring process target. Quality and Reliability Engineering International 29: 1069-1080.

2. Abid M., Nazir H. Z., Riaz M. and Lin Z. (2016). Use of ranked set sampling in nonparametric control charts. Journal of the Chinese Institute of Engineers 39: 627-636.

3. Abid M., Nazir M. and Lin Z. (2016). An efficient non-parametric EWMA Wilcoxon signed-rank chart for monitoring location. Quality and Reliability Engineering International 33: 669-685.

4. Amin R.W., Reynolds M. R. Jr. and Bakir S. T. (1995). Nonparametric quality control charts based on the sign statistic. Communications in Statistic-Theory and Methods 24(6): 1597-1623.

5. Bakir S. T. (2004). A distribution-free Shewhart quality control chart based on signed-ranks. Quality Engineering 16(4): 613-623.

6. Bakir S. T. (2006). Distribution-free quality control charts based on signed-rank-like statistic. Communications in Statistics-Theory and Methods 35: 743-757.

7. Bakir S. T. and Reynolds M .R. Jr. (1979). A nonparametric procedure for process control based on within-group ranking. Technometrics 2: 175-183.

8. Bourke P. D. (1991). Detecting a shift in fraction nonconforming using run-length control charts with 100% inspection. Journal of Quality Technology 23: 225- 238.

9. Bradley J. V. (1973). The central limit effect for a variety of populations and the influence of population moments. Journal of Quality Technology 5: 171-177.

10. Chakraborti S. and Eryilmaz S. (2007). A nonparametric Shewhart-type signed-rank control chart based on runs. Communications in Statistics-Simulation and Computations 36: 335-356.

11. Chakraborti S. and Van de Wiel M. A. (2008). A nonparametric control charts based on mann-whitney statistic. IMS Collection 1: 156-172.

12. Chakraborti S., Van der Laan P. and Bakir S. (2001). Nonparametric control charts: An overview and some results. Journal of Quality Technology 33: 304–315.

13. Champ W. C. (1992). Steady-state run length analysis of a shewhart control chart with supplementary runs rules. Communications in Statistics- Theory and Methods 21: 765-777.

14. Charts. Industrial Quality Control 23(11): 563-568.

15. Coelho M. L. I., Graham M. A. and Chaktraborti S. (2017). Monitoring location: A nonparametric control charts with variable sampling intervals. Quality and Reliability Engineering International 33:2181-2192.

16. Crosier R. B. (1986). A new two-sided cumulative sum quality control scheme. Technometrics 28(3): 187-194.

17. Davis R. B. and Woodall W. H. (2002). Evaluating and improving the synthetic control chart. Journal of Quality Technology 34(2): 200-208.

18. Hackl L. and Ledolter J. (1991). A control chart based on ranks. Journal of Quality Technology 23: 117-124.

19. Ho L. L. and Costa A. F. B. (2011). Monitoring a wandering mean with an np chart. Producao 21(2): 254-258.

20. Human S. W., Chakraborti S. and Smit C. F. (2010). Nonparametric Shewhart-type sign control charts based on runs. Communications in Statistics-Theory and Methods 39: 2046-2062.

21. Khilare S. K. and Shirke D. T. (2010). A nonparametric synthetic control chart using sign statistic. Communications in Statistics-Theory and Methods 39: 3282-3293.

22. Khilare S. K. and Shirke D. T. (2012). Nonparametric synthetic control charts for process variation. Quality and Reliability Engineering International 28(2): 193-202.

23. Khilare S. K. and Shirke D. T. (2015). Steady-state behavior of nonparametric control charts using sign statistic. Production 25:739-749.

24. Lim T. and Cho M. (2009). Design of control charts with m-of-m runs rules. Quality and Reliability Engineering International 25: 1085-1101.

25. Liu L., Zhang J. and Zi X. (2014). Dual nonparametric CUSUM control chart based on ranks. Communications in Statistics-Simulation and Computation 44:756-772.

26. Montgomery D. C. (2009). Introduction to statistical quality control. Arizona State University. Johan Willey and Sons.

27. Majid N. and Neda N. (2013). Nonparametric Shewhart-type signed-rank control chart with variable sampling intervals. Quality and Reliability Engineering International 24(2): 184-189.

28. Pawar V. Y. and Shirke D. T. (2010). A nonparametric Shewhart-type synthetic control chart. Communications in Statistics-Simulation and Computation 39(8): 1493-1505.

29. Riaz M. and Abbasi S. (2016). Nonparametric double EWMA control chart for process monitoring. Revista Colombian de Estadistica 39: 167-184.

30. Saccucci M. S. and Lucas J. M. (1990). Average run length for exponentially weighted moving average control schemes using the markov chain approach. Journal of Quality Technology 22(2): 154-162.

31. *Schilling E. G. and Nelson P. R. (1976). The effect of non-normality on the control limits of x-bar charts. Journal of Quality Technology 8: 183–188.*

32. Wu Z. and Spedding T. A. (2000). A synthetic control chart for detecting small shifts in the process mean. Journal of Quality Technology 32: 32-38.

33. *Yang S. F., Tsai W. C., Huang R. M., Yang C. C. and Cheng S. (2011). Monitoring process mean with a new EWMA control chart. Producao 21(2): 217-222.*

Received: 9 December 2016 | Revised: 22 November 2017 | Accepted: 5 February 2018

DOI: 10.1002/qre.2295

RESEARCH ARTICLE

WILEY

A nonparametric CUSUM chart for process dispersion

D.T. Shirke | M.S. Barale

Department of Statistics, Shivaji University, Kolhapur 416004, Maharashtra, India

Correspondence
M. S. Barale, Department of Statistics, Shivaji University, Kolhapur 416004, Maharashtra, India.
Email: baralemahesh12@gmail.com

Funding information
UGC-Major Research Project, Grant/ Award Number: 43-542/2014

Abstract

In the present article, we propose a nonparametric cumulative sum control chart for process dispersion based on the sign statistic using in-control deciles. The chart can be viewed as modified control chart due to Amin et al,[6] which is based on in-control quartiles. An average run length performance of the proposed chart is studied using Markov chain approach. An effect of non-normality on cumulative sum S^2 chart is studied. The study reveals that the proposed cumulative sum control chart is a better alternative to parametric cumulative sum S^2 chart, when the process distribution is non-normal. We provide an illustration of the proposed cumulative sum control chart.

KEYWORDS

average run length, nonparametric chart, process control, sign test

1 | INTRODUCTION

Control charts are used to monitor process parameter, such as location, dispersion, and proportion of defectives. The widely used control charts are $\bar{X}$ chart for process location and R chart or S^2 chart for the process dispersion; these charts are also known as Shewhart's charts. The main drawback of these charts is that these charts are less efficient against small shifts. One can resolve this problem by using runs rules. There are some operational issues while implementing the runs rules charts. An alternative to detect the small shift is to use memory chart, as like cumulative sum (CUSUM) chart proposed by Page[1] or exponentially weighted moving average (EWMA) charts proposed by Roberts.[2] These charts consider the past as well as current information about the process, which makes charts very sensitive to small shifts in the process parameters. In the literature, various parametric CUSUM procedures are available for monitoring process location and dispersion, but very few nonparametric CUSUM procedures are available to monitor the process dispersion.

The traditionally used CUSUM S and CUSUM S^2 charts are based on the assumption of normality, but when the process distribution is not normal, the false alarm probability of the chart varies. Therefore, one of the robust alternatives to these charts is to use the nonparametric control charts. A control chart is said to be nonparametric, if its in-control average run length (ARL) does not depend on underlying process distribution. The performance of a control chart is usually measured using the ARL, which is defined as an average number of samples required to get an out-of-control signal.

Till date, there are several parametric as well as nonparametric Shewhart's control charts reported in the literature for process location and dispersion. Bakir[3] developed a distribution-free Shewhart quality control chart based on a signed-rank like statistic for process location. Chakraborti and Eryilmaz[4] have proposed a control chart based on a signed-rank statistic for process center. Khilare and Shirke[5] developed a nonparametric synthetic control chart using a sign statistic for process location. Amin et al[6] developed a nonparametric control chart based on sign statistic for the process center and variability. They have also developed CUSUM chart by using sign statistic for process center and reported that it can be extended for variability also. Rendtel[7] and Reynolds et al[8] described a CUSUM chart with variable sampling intervals for process mean. Yang and Cheng[9] have proposed a nonparametric CUSUM mean chart based on the sign

statistic. Das[10] developed a nonparametric control chart for variability based on the squared rank statistic. Khilare and Shirke[11] have proposed a nonparametric synthetic control chart for the process variation. Shirke et al[12] have proposed a nonparametric control chart for process variability based on in-control deciles. Zhou et al[13] provided a nonparametric quality control chart based on Ansari-Bradly test statistic for variability. Chowdhury et al[14] constructed a nonparametric control chart for joint location and scale monitoring, which is based on the Lepage test. Guo and Wang[15] have proposed a variable sampling interval S^2 chart with known or unknown in-control variance. Zombade and Ghute[16] provided 4 nonparametric control charts for the process variation based on Sukhatme's 2 sample test and Mood's test. In the proposed work, we propose a CUSUM chart based on sign statistic defined by Shirke et al.[12] The sign statistic is defined using in-control deciles of the process distribution.

The remaining article is organized as follows. Section 2 describes the effect of non-normality on S^2 chart, a nonparametric CUSUM chart based on in-control deciles, and method for obtaining its ARL. Section 3 provides the performance study of control charts for various process distributions. Sections 4 and 5 provide the illustrative example and conclusions, respectively.

2 | A NONPARAMETRIC CUSUM CHART BASED ON IN-CONTROL DECILES

Suppose we are monitoring the process for detecting variation in the quality characteristic of interest say X. Let variance of X is σ^2 and when the process is in-control $\sigma^2 = \sigma_0^2$. We monitor the process by drawing a random sample of size n at fixed time epoch. Let X_{ij} be the j^{th} observation at time epoch i, where $i = 1, 2, \ldots$ and $j = 1, 2, \ldots, n$. In the literature, parametric CUSUM S^2 chart is used to monitor small changes in process dispersion, where S^2 be the sample variance. We have to detect a shift in process dispersion quickly. Let σ_1^2 be the process variance after change in the process dispersion.

The charting statistic for CUSUM S^2 chart are as follows:

$$\begin{aligned} C_i^+ &= max(0, S_i^2 - k + C_{i-1}^+) \\ C_i^- &= max(0, k - S_i^2 + C_{i-1}^-), \end{aligned} \tag{1}$$

where $k = [2ln(\sigma_0/\sigma_1)\sigma_0\sigma_1/(\sigma_0 - \sigma_1)]$ and $C_0^+ = C_0^- = 0$. The statistic C_i^+ and C_i^- are called as an upper and lower CUSUM's respectively and initial values of C_i^+ and C_i^- are taken to be zero. The chart signals, if any of the C_i^+ or C_i^- exceeds a prespecified control limit h. The parameters h is chosen to meet in-control ARL specified by an experimenter. Therefore, ARL is a function of n, h, and k for CUSUM procedure. One can use the C_i^+ to detect an increase in the process dispersion only when corresponding upper one-sided ARL is denoted as ARL^+.

The construction of CUSUM S^2 chart is based on the assumption of normality or at least approximately normality of the process quality characteristic. Amin et al[6] discussed the effect of non-normality on control charts for location. If the process distribution deviates from normal, the ARLs obtained by assuming normality will differ. Table 1 gives the ARL values for CUSUM S^2 chart with sample size $n = 10$, upper control limit $h = 1.5362$ and $k = 1.24$ for the normal, double exponential, uniform, exponential and gamma distributions. Here double exponential is the example of heavy tailed and uniform is the example of light tailed distribution. An effect of skewed distributions on CUSUM S^2 chart is also studied. The upper control limit only considered with various shifts in a standard deviation that is $\sigma = \delta\sigma_0$, where δ be the extent of increase in process standard deviation. Table 1

TABLE 1 ARL^+ performance of CUSUM S^2 chart for n = 10

δ	Normal	Laplace	Uniform	Exponential	Gamma
01	284.2	36.8	448631.0	23.6	37.1
1.2	11.8	9.2	328.6	8.6	9.3
1.4	3.9	4.3	21.2	4.6	4.3
1.6	2.4	2.7	7.9	3.0	2.8
1.8	1.7	2.1	4.8	2.3	2.0
2	1.4	1.7	3.5	1.9	1.7
3	1.0	1.1	1.7	1.2	1.1
4	1.0	1.0	1.3	1.1	1.0
5	1.0	1.0	1.1	1.0	1.0

depicts that if the process distribution is heavy tailed or skewed and control limit is set under normality assumption, then ARLs are very small as compared with the normal. While for light tailed distribution, the ARLs tend to be larger as compared with normal. This implies that for heavy tailed distribution, a false alarm will occur frequently.

Shirke et al[12] developed a sign chart for variability based on in-control deciles, which is a modification of a sign chart based on in-control quartiles given by Amin et al.[6] The chart procedure proposed by Shirke et al[12] is as follows. Consider D_2 and D_8 respectively be the 2^{nd} and 8^{th} deciles, when the process is in-control. We assume that D_2 and D_8 are known from the past data.

Define

$$W_{ij} = \begin{cases} 1 & X_{ij} < D_2 \text{ or } X_{ij} > D_8 \\ 0 & X_{ij} = D_2 \text{ or } X_{ij} = D_8 \\ -1 & D_2 < X_{ij} < D_8, \end{cases} \quad (2)$$

and $W_i = \sum_{j=1}^{n} W_{ij}$. Define a random variable $V_i=(W_i+n)/2$ and has binomial distribution with parameters n and p, where $p=P\{X_{ij}<D_2 \text{ or } X_{ij}>D_8|\sigma=\delta\sigma_0\}$. Moreover, when the process is in-control $p = p_0 = 0.4$. The two-sided chart gives signal if $V_i > c$ or $V_i < n - c$, where c is chosen such that

$$\alpha = \sum_{j=0}^{n-c-1} \binom{n}{j} p_0^j (1-p_0)^{n-j} + \sum_{j=c+1}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j}. \quad (3)$$

In the upper one-sided case, c is chosen such that

$$\alpha = 1 - \sum_{j=0}^{c} \binom{n}{j} p_0^j (1-p_0)^{n-j}, \quad (4)$$

where α be the false alarm probability, when the process is in-control.

Shirke et al[12] have shown that the chart based on deciles outperforms chart based on quartiles proposed by Amin et al.[6] We extend this approach and provide a nonparametric CUSUM chart to monitor the process dispersion σ^2. We define

$$U_{ij} = \begin{cases} 1 & X_{ij} \le D_2 \text{ or } X_{ij} \ge D_8 \\ 0 & X_{ij} > D_2 \text{ or } X_{ij} < D_8, \end{cases} \quad (5)$$

and $U_i = \sum_{j=1}^{n} U_{ij}$. U_i has binomial distribution with parameters n and p, where $p = P(X_{ij} \le D_2 \text{ or } X_{ij} \ge D_8|\sigma = \delta\sigma_0)$.

The small shifts in process dispersion can be monitored with the help of the proportion of the observations which falls in the tails. When there is a change in the process variation, we denote p by p_1. Consider $\psi = |p_0 - p_1|$, $\psi > 0$ and we wish to detect a shift of size p_1 quickly. Define a CUSUM monitoring statistic for the i^{th} subgroup sample,

$$\begin{aligned} C_i^+ &= max(0, U_i - (np_0 + k_1) + C_{i-1}^+) \\ C_i^- &= max(0, (np_0 - k_1) - U_i + C_{i-1}^-), \end{aligned} \quad (6)$$

where k_1 is the reference value with $k_1 = \frac{n\psi}{2}$. The initial starting values are mostly chosen as zero, that is, $C_0^+ = 0$ and $C_0^- = 0$. Let H be the parameter of a nonparametric CUSUM chart. If $C_i^+ > H$ or $C_i^- > H$, then the process is thought to be out-of-control. Moreover, $C_i^+ > H$ is used to detect an increase in process dispersion, while $C_i^- > H$ for to detect the decrease in process dispersion. It is noted that

TABLE 2 The (k_1,H) values under $ARL_0^+ \approx 370$

$ARL^+ \approx 370$						
ψ	**0.1**		**0.2**		**0.3**	
n	k_1	***H***	k_1	***H***	k_1	***H***
5	0.25	8.69	0.50	5.00	0.75	3.61
6	0.30	8.11	0.60	5.00	0.90	3.49
7	0.35	8.10	0.70	5.10	1.05	3.50
8	0.40	8.20	0.80	5.00	1.20	3.60
9	0.45	8.80	0.90	5.00	1.35	3.30
10	0.50	9.30	1.00	5.00	1.50	3.51
11	0.55	9.00	1.10	5.00	1.65	3.81
12	0.60	9.98	1.20	5.00	1.80	3.58
13	0.65	9.00	1.30	5.30	1.95	3.69
14	0.70	9.30	1.40	5.00	2.10	3.58
15	0.75	10.20	1.50	5.00	2.25	3.73

TABLE 3 ARL^{+} comparison for various process distributions

	δ	Normal		Laplace		Uniform		Exponential		Gamma	
		N-CUSUM	CUSUM S^2	N-CUSUM	CUSUM S^2	N-CUSUM	CUSUM S^2	N-CUSUM	CUSUM S^2	N-CUSUM	CUSUM S^2
h			1.5362		3.643		0.936		6.957		3.713
ARL*			284.2		36.6		448631.0		23.8		37.0
n=10	1	284.0	284.2	284.0	284.4	284.0	285.5	284.0	283.5	284.0	283.4
H=8.2	1.2	20.2	11.7	28.2	25.1	15.5	16.6	75.3	29.7	56.9	26.1
k=0.5	1.4	9.2	3.9	12.1	8.4	7.6	5.7	29.4	11.5	18.6	8.5
	1.6	6.3	2.3	8.0	4.7	5.5	3.4	16.4	6.9	10.1	4.8
	1.8	5.1	1.7	6.2	3.3	4.5	2.5	11.3	4.9	7.0	3.4
	2	4.4	1.4	5.3	2.6	3.9	2.0	8.8	3.8	5.5	2.6
	3	3.1	1.0	3.5	1.4	2.9	1.2	4.7	1.9	3.2	1.4
	4	2.7	1.0	3.0	1.1	2.6	1.1	3.6	1.4	2.7	1.1
	5	2.5	1.0	2.7	1.0	2.4	1.0	3.2	1.2	2.4	1.0
h			1.1		3.4804		0.39		5.1082		2.887
ARL*			283.6		32.1		605126.4		20.1		32.3
n=15	1	284.4	283.6	284.4	283.0	284.4	285.0	284.4	285.5	284.4	284.5
H=8.8	1.2	15.3	8.2	21.9	16.2	11.6	14.6	64.4	22.9	47.2	18.8
k=0.75	1.4	6.8	2.8	9.0	6.1	5.6	4.2	22.9	8.6	14.0	6.2
	1.6	4.7	1.8	5.9	3.7	4.1	2.3	12.3	5.2	7.5	3.6
	1.8	3.8	1.4	4.6	2.7	3.4	1.7	8.4	3.7	5.2	2.6
	2	3.3	1.2	3.9	2.1	3.0	1.4	6.5	2.9	4.1	2.0
	3	2.4	1.0	2.7	1.2	2.2	1.0	3.5	1.5	2.4	1.2
	4	2.1	1.0	2.3	1.0	2.1	1.0	2.8	1.2	2.1	1.0
	5	2.0	1.0	2.1	1.0	2.0	1.0	2.4	1.1	2.0	1.0
h			0.9		2.3		0.2993		3.495		2.367
ARL*			283.8		29.7		735294.2		17.9		29.3
n=20	1	283.1	283.8	283.1	284.3	283.1	283.8	283.1	283.6	283.1	284.7
H=9.2	1.2	12.4	6.5	17.9	14.3	9.4	12.0	56.0	20.3	40.1	14.8
k=1	1.4	5.5	2.3	7.3	4.8	4.6	3.4	18.8	7.1	11.4	5.0
	1.6	3.8	1.5	4.8	2.9	3.3	1.9	10.0	4.1	6.1	2.9

(Continues)

TABLE 3 (Continued)

δ	Normal		Laplace		Uniform		Exponential		Gamma	
	N-CUSUM	CUSUM S^2	N-CUSUM	CUSUM S^2	N-CUSUM	CUSUM S^2	N-CUSUM	CUSUM S^2	N-CUSUM	CUSUM S^2
1.8	3.1	1.2	3.8	2.1	2.8	1.4	6.8	2.9	4.2	2.1
2	2.7	1.1	3.2	1.7	2.5	1.2	5.3	2.3	3.4	1.7
3	2.0	1.0	2.2	1.1	2.0	1.0	2.9	1.3	2.1	1.1
4	1.9	1.0	2.0	1.0	1.8	1.0	2.3	1.0	1.8	1.0
5	1.8	1.0	1.9	1.0	1.7	1.0	2.1	1.0	1.7	1.0

ARL* is the in-control ARL when the control limit set under the normality assumption and N-CUSUM is the proposed nonparametric CUSUM chart.

the reference value k_1 and control limit H can be chosen such that they would satisfy the specified ARL.

It is easy to compute ARLs for Shewhart-type control chart and not so for CUSUM and EWMA control charts. There are different methods in the literature to compute ARLs of a CUSUM chart. Brook and Evans[17] have given Markov chain approach to obtain ARL of a CUSUM chart. Yang and Cheng[9] used Markov chain approach to computing ARLs of a CUSUM chart based on sign statistic. We first obtain ARL for the upper one-sided CUSUM chart. We divide the region $(0, H)$ into $M-1$ subintervals of equal width of $2w$, where $w = H/(2(M-1))$. Take 1^{st} subinterval as $(-\infty, 0]$, the k^{th} interval is $(m_k - w, m_k+w)$, where m_k be the midpoints of k^{th} subinterval with $m_1 = 0$, $m_k=(2k-3)H/(2(M-1))$ for $k = 2,3,...,M$; and $(M+1)^{th}$ interval as (H, ∞). These all $M+1$ subintervals can be viewed as states of Markov chain. Moreover, the state $M+1$ is the action state, which is absorbing state and remaining M states are transient states of Markov chain $\{C_i^+; i = 0, 1, ...\}$.

Consider the transition probability matrix corresponding to transient states $1, 2, ...M$ be $R^p = ((p_{kj}^p))$, $(k, j = 1, 2, ..., M)$, whose kj^{th} element represents the transition probability that statistic C_i^+ reaches state j at time i, given that C_{i-1}^+ was in state k at time $(i-1)$. The transition probabilities can be calculated as

$$\begin{aligned} p_{k1}^p &= P(C_i^+ \le 0 | C_{i-1}^+ = m_k) = P(U_i - (np_0 + k_1) \\ &\quad + C_{i-1}^+ \le 0 | C_{i-1}^+ = m_k) \\ &= P(U_i \le np_0 + k - m_k) \\ &= \sum_{s=0}^{[np_0+k_1-m_k]} \binom{n}{s} p^s (1-p)^{n-s}, \end{aligned}$$

$k = 1, 2, ..., M; i = 1, 2, 3, ...$

$$\begin{aligned} p_{kj}^p &= P(m_j - w \le C_i^+ < m_j + w | C_{i-1}^+ = m_k) \\ &= P(m_j - w \le U_i - (np_0 + k_1) + C_{i-1}^+ \\ &\quad < m_j + w | C_{i-1}^+ = m_k) \\ &= P(m_j - m_k - w + np_0 + k_1 \le U_i < m_j - m_k \\ &\quad + w + np_0 + k_1) \\ &= \sum_{s=0}^{[(m_j-m_k+w+np_0+k_1)^-]} \binom{n}{s} p^s (1-p)^{n-s} \\ &\quad - \sum_{s=0}^{[(m_j-m_k-w+np_0+k_1)^-]} \binom{n}{s} p^s (1-p)^{n-s}, \end{aligned}$$

$k = 1, 2, ..., M, j = 2, 3, ..., M$ and $i = 1, 2, 3, ...,$ where $(\boldsymbol{\beta})^-$ be the largest integer not greater than $\boldsymbol{\beta}$. Let b be the $M \times 1$ vector of probabilities that the process started in state $1, 2, ...M$. In this case $b = (b_1, b_2, ..., b_M)'$. Since we considered that $C_0^+ = C_0^- = 0$, we get $b_1=1$ and $b_k = 0$

for $k \neq 1$. Consider $P^p = ((p^p_{kj}))$ be a $(M+1)\times(M+1)$ transition probability matrix such that

$$P^p = \left[\begin{array}{c|c} R^p_{M\times M} & P_{M\times 1} \\ \hline 0'_{1\times M} & \mathbf{1} \end{array}\right].$$

Then ARL corresponding to upper one-sided CUSUM chart can be obtained as $ARL^+ = b'(I - R^p)^{-1}\mathbf{1}'$, where $\mathbf{1}' = (1,1,...,1)$ be the $1\times M$ vector with elements 1. The in-control ARLs can be calculated by substituting $p=p_0$, therefore $ARL^+ = ARL_0^+$ be the in-control ARL and if $p=p_1$ then $ARL^+ = ARL_1^+$ be the out-of-control ARL. Similar way, one can compute ARL for lower one-sided CUSUM chart, which is denoted by ARL^-. Then the ARL for nonparametric CUSUM chart can be calculated as follows:

$$ARL = \frac{1}{1/ARL^+ + 1/ARL^-}. \quad (7)$$

Table 2 gives values of k_1 and H under $ARL_0^+ \approx 370$ for sample size 5 to 15 and ψ=0.1,0.2,0.3.

3 | PERFORMANCE STUDY OF THE NONPARAMETRIC CUSUM CHART BASED ON DECILES

The performance of control charts can be studied to measure its ability to detect a change in the process parameter quickly. ARL is one of the performance measures, which is used for comparison of control charts. The chart is more efficient, when in-control ARL is large and corresponding out-of-control ARL is small. We have studied the performance of the proposed chart for various process distributions (normal, Laplace, uniform, exponential, and gamma). In the literature, no any standard nonparametric CUSUM chart is available to monitor process dispersion. Therefore, We have compared the performance of proposed nonparametric CUSUM chart with parametric CUSUM S^2 chart.

In most of the situations, early detection of an increase in the process dispersion is of interest and in that case, a one-sided control chart is desirable. The performance of the proposed chart is reported for sample sizes n = 10, 15, 20 with shift δ in process standard deviation. Based on 20,000 runs, the ARL for CUSUM S^2 chart is computed. Table 3 provide ARLs along with various shifts in process standard deviation for Normal $(0,1)$, Laplace $(0,1)$, Uniform $(a=0, b=\sqrt{12}+a)$, Exponential $(\theta=1)$ and Gamma $(a=2, b=\sqrt{a})$ distribution for sample size $n = 10, 15, 20$. It is clear from Table 3 that an out-of-control ARLs for CUSUM S^2 chart are smaller than nonparametric CUSUM chart, which indicate that CUSUM S^2 chart is more efficient than nonparametric CUSUM chart for all distributions under study. But, such comparison is meaningless because in-control ARL is obtained by using control limit which is set under normality assumption (ARL^*). For example, ARL^* is quite low (36.6) for Laplace distribution. Suppose we are interested to enhance ARL^* from 36.6 to 284 using multiplicative factor (284/36.6=7.75), we get out-of-control ARL to detect shift in variation of 1.2σ as 194.5. This is significantly larger than corresponding out-of-control ARL 28.2 for nonparametric CUSUM chart. Here, we can see that the ARL^* of CUSUM S^2 chart changes from 284.2 (for normal distribution) to 36.6, 448631.0, 23.8 and 37, when the process distribution is Laplace, uniform, exponential and gamma

TABLE 4 The ARL_1^+ values under $ARL_0^+ \approx 370$, ψ=0.1, and p_0=0.4

	p_1					
n	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
9	365.1	17.9	7.1	4.4	3.2	2.6
10	377.9	16.8	6.8	4.3	3.2	2.5
11	367.8	15.6	6.1	3.8	2.8	2.2
12	379.7	14.8	6.1	3.9	2.9	2.2
13	374.7	13.9	5.4	3.4	2.5	2.1
14	367.2	13.0	5.2	3.4	2.5	2.1
15	375.3	12.4	4.9	3.1	2.3	2.0
16	376.0	11.9	4.7	2.9	2.2	2.0
17	364.4	11.3	4.4	2.8	2.1	2.0
18	367.1	10.8	4.4	2.9	2.2	2.0
19	375.3	10.4	4.2	2.7	2.1	1.9

TABLE 5 Piston rings data and values of charting statistic

Sample	X_1	X_2	X_3	X_4	X_5	V	C^+
1	74.030	74.002	74.019	73.992	74.008	3	0.75
2	73.995	73.992	74.001	74.011	74.004	2	0.50
3	73.988	74.024	74.021	74.005	74.002	3	1.50
4	74.002	73.996	73.993	74.015	74.009	1	0.25
5	73.992	74.007	74.015	73.989	74.014	4	3.25
6	74.009	73.994	73.997	73.985	73.993	1	2.00
7	73.995	74.006	73.994	74.000	74.005	0	1.00
8	73.985	74.003	73.993	74.015	73.988	3	4.00
9	74.008	73.995	74.009	74.005	74.004	0	1.75
10	73.998	74.000	73.990	74.007	73.995	1	2.75
11	73.994	73.998	73.994	73.995	73.990	1	2.75
12	74.004	74.000	74.007	74.000	73.996	0	1.75
13	73.983	74.002	73.998	73.997	74.012	2	3.75
14	74.006	73.967	73.994	74.000	73.984	2	3.75
15	74.012	74.014	73.998	73.999	74.007	2	3.75
16	74.000	73.984	74.005	73.998	73.996	1	2.75
17	73.994	74.012	73.986	74.005	74.007	2	3.75
18	74.006	74.010	74.018	74.003	74.000	2	3.75
19	73.984	74.002	74.003	74.005	73.997	1	2.75
20	74.000	74.010	74.013	74.020	74.003	3	4.75
21	73.982	74.001	74.015	74.005	73.996	2	4.50
22	74.004	73.999	73.990	74.006	74.009	1	3.50
23	74.010	73.989	73.990	74.009	74.014	4	6.50
24	74.015	74.008	73.993	74.000	74.010	2	6.25
25	73.982	73.984	73.995	74.017	74.013	4	8.25
26	74.012	74.015	74.030	73.986	74.000	4	10.00
27	73.995	74.010	73.990	74.015	74.001	3	10.75
28	73.987	73.999	73.985	74.000	73.990	3	11.50
29	74.008	74.010	74.003	73.991	74.006	2	11.25
30	74.003	74.000	74.001	73.986	73.997	1	10.25
31	73.994	74.003	74.015	74.020	74.004	2	11.25
32	74.008	74.002	74.018	73.995	74.005	1	10.25
33	74.001	74.004	73.990	73.996	73.998	1	10.25
34	74.015	74.000	74.016	74.025	74.000	3	12.25
35	74.030	74.005	74.000	74.016	74.012	3	13.00
36	74.001	73.990	73.995	74.010	74.024	3	13.75
37	74.015	74.020	74.024	74.005	74.019	4	15.50
38	74.035	74.010	74.012	74.015	74.026	5	18.25
39	74.017	74.013	74.036	74.025	74.026	5	21.00
40	74.010	74.005	74.029	74.000	74.020	3	21.75

respectively for sample size n=10. Moreover, ARL* changes from 283.8 to 735294.2 for uniform distribution with n=20. It means, there is very low false alarm.

The nonparametric CUSUM chart has smaller out-of-control ARLs when process distributions are uniform and normal. But it has larger ARLs when the process distribution is heavy tailed or skewed like Laplace, exponential and gamma. It is observed that from Table 3 that out-of-control ARLs decreases as sample size increases. In Table 4, one-sided out-of-control ARL values for various values of p_1 and sample size n=9 to 19 are reported. It can be observed that out-of-control ARLs decreases as sample size increases and as tail proportion p_1 increase, that is shift in process dispersion is increases.

4 | EXAMPLE

Here, we illustrate the construction of a nonparametric sign chart based on deciles and proposed CUSUM chart based on deciles with the example inside diameter measurements (mm) for automobile engine piston rings data Montgomery.[18] There are 25 primary samples and 15 additional samples each of size 5, which is described in Table 5. Figures 1 and 2 show that a nonparametric control chart based on in-control deciles with control limit

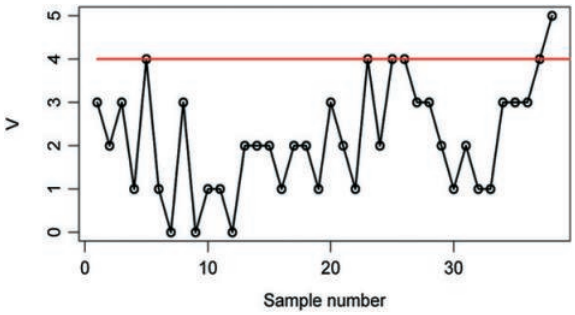


FIGURE 1 A nonparametric control chart based on deciles for piston rings data [Colour figure can be viewed at wileyonlinelibrary.com]

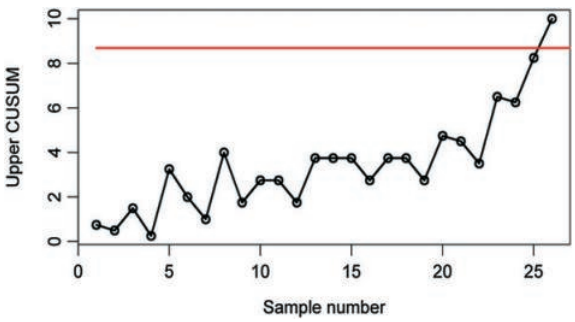


FIGURE 2 A nonparametric CUSUM chart for piston rings data [Colour figure can be viewed at wileyonlinelibrary.com]

c=4 and a nonparametric CUSUM chart with k_1=0.25 and H=8.69. We can see that a nonparametric control chart based on in-control deciles gives the signal on 38^{th} sample while a CUSUM chart gives the signal on 26^{th} sample.

5 | CONCLUSION

In the present article, we present a nonparametric CUSUM chart based on in-control deciles for detecting small shifts in process dispersion. Since, whatever be the process distribution the proposed nonparametric CUSUM chart give same in-control ARL. Therefore, the proposed nonparametric chart is a better alternative to CUSUM S^2 chart when process distribution is not known in advance. Moreover, it does not require any distributional assumption. The performance in terms of ARL of the proposed control chart for various distributions is studied. Due to the simplified procedure of proposed CUSUM chart, we recommend for use of proposed CUSUM chart.

ACKNOWLEDGEMENTS

Both the authors would like to acknowledge the financial support received from University Grants Commission under Major Research Project (F. No. 43-542/2014 (SR)) to conduct the research work.

REFERENCES

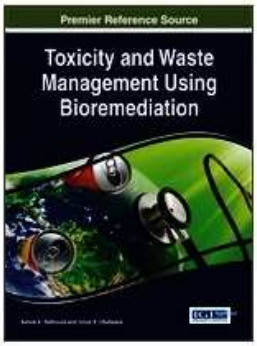
1. Page ES. Cumulative sum charts. *Technometrics*. 1961;3(1):1-9.
2. Roberts SW. Control chart tests based on geometric moving averages. *Technometrics*. 2000;42(1):97-101.
3. Bakir ST. Distribution-free quality control charts based on signed-rank-like statistics. *Commun Stat Theory Methods*. 2006;35(4):743-757.
4. Chakraborti S, Eryilmaz S. A nonparametric Shewhart-type signed-rank control chart based on runs. *Commun Stat Simul Comput*. 2007;36(2):335-356.
5. Khilare SK, Shirke DT. A nonparametric synthetic control chart using sign statistic. *Commun Stat Theory Methods*. 2010; 39(18):3282-3293.
6. Amin RW, Reynolds MR, Bakir ST. Nonparametric quality control charts based on the sign statistic. *Commun Stat-Theory Methods*. 1995;24(6):1597-1623.
7. Rendtel U. CUSUM-schemes with variable sampling intervals and sample sizes. *Stat Pap*. 1990;31(1):103-118.
8. Reynolds MR, Amin RW, Arnold JC. CUSUM charts with variable sampling intervals. *Technometrics*. 1990;32(4):371-384.
9. Yang S, Cheng SW. A new non-parametric CUSUM mean chart. *Qual Reliab Eng Int*. 2011;27(7):867-875.
10. Das N. A non-parametric control chart for controlling variability based on squared rank test. *J Ind Syst Eng*. 2008;2(2):114-125.

11. Khilare SK, Shirke DT. Nonparametric synthetic control charts for process variation. *Qual Reliab Eng Int*. 2012;28(2):193-202.
12. Shirke DT, Pawar VY, Chakraborti S. A nonparametric control chart for monitoring variability based on the deciles, Under review; 2016.
13. Zhou M, Zhou Q, Geng W. A new nonparametric control chart for monitoring variability. *Qual Reliab Eng Int*. 2016; 32(7):2471-2479.
14. Chowdhury S, Mukherjee A, Chakraborti S. Distribution free phase II CUSUM control chart for joint monitoring of location and scale. *Qual Reliab Eng Int*. 2008;31(1):135-151.
15. Guo B, Wang BX, Guo B. The variable sampling interval s^2 chart with known or unknown in-control variance. *Int J Prod Res*. 2016;54(11):3365-3379.
16. Zombade DM, Ghute VB. Nonparametric CUSUM charts for process variability. *J Academia Ind Res*. 2014;3(1):53.
17. Brook D, Evans DA. An approach to the probability distribution of CUSUM run length. *Biometrika*. 1972;9(3):539-549.
18. Montgomery DC. *Introduction to Statistical Quality Control*. New York: Wiley; 2009.

D. T. Shirke is a Professor of Statistics at the Shivaji University, Kolhapur, India. He received his PhD in Statistics from Shivaji University, Kolhapur, India. His research areas include statistical inference and statistical process control. He is an elected member of International Statistical Institute, the Netherlands.

M. S. Barale is PhD student at Department of Statistics, Shivaji University, Kolhapur, India. He received his master's degree in Statistics in 2015 from the Shivaji University, Kolhapur, India.

How to cite this article: Shirke DT, Barale MS. A Nonparametric CUSUM Chart for Process Dispersion. *Qual Reliab Engng Int*. 2018;34: 858–866. https://doi.org/10.1002/qre.2295



Application of Genomics and Proteomics in Bioremediation

Amol Uttam Hivrale (Shivaji University, India), Pankaj K. Pawar (Shivaji University, India), Niraj R. Rane (Shivaji University, India) and Sanjay P. Govindwar (Shivaji University, India)

Source Title: Toxicity and Waste Management Using Bioremediation
Copyright: © 2016 | Pages: 16
DOI: 10.4018/978-1-4666-9734-8.ch005

Abstract

Bioremediation mediated by microorganisms is proving to be cost effective, ecofriendly and sustainable technology. Genome enable experimental and modeling techniques are of a great help in evaluating physiology and enhancing performance of life forms to be used for bioremediation purpose. Similarly, the application of proteomics in bioremediation research provides a global view of the protein composition of microbial cell and offers promising approach to understand the molecular mechanism of removal of toxic material from the environment. Combination of proteomics and genomics in bioremediation is an insight into global metabolic and regulatory network that can enhance the understanding of gene functions. Present chapter give a bird's eye view of genomics and proteomics and their potential utilization in bioremediation and for the clearer understanding of the cellular responses to environmental stimuli. An understanding of the growth conditions governing the expression of proteome in a specific environment is essential for developing rational strategies for successful bioremediation.

Chapter Preview

Top

1. Introduction

Bioremediation is a process in which naturally occurring organisms are used for rapid degradation / removal of hazardous pollutants from environment in order to obtain healthy soil, sediments, substances and ground water (Kumar et al., 2011). In natural way biodegradation is the recycling of waste or breaking down organic matter in to nutrients for the other organisms (Alexander, 1994). Bioremediation is carried out with the help of life forms, including bacteria, fungi, insects, worms, plants, etc. by taking nutrients such as C, N and P from the contaminant ultimately transforming xenobiotics in to environment friendly products (Vidali, 2001). Bioremediation approach becomes important when it comes to remediation of water reserves. Industrial effluents especially textile industry waste are responsible for contamination of water bodies which result in limiting the water availability for drinking and agriculture purpose (King et al., 1997).

Microbes and Bioremediation

Dynamic behavior, flexibility in nutritional requirements and ability to adopt under extreme stress conditions makes the microbe the most eligible life forms for survival. This virtue of the microbe is proving to be beneficial to human kind especially when it comes to removal of contaminants / toxic entities from

Heavy Metal Contamination of Soils pp 433-470 | Cite as

Genetic Engineering of Plants for Heavy Metal Removal from Soil

Authors | Authors and affiliations

Umesh B. Jagtap, Vishwas A. Bapat

Part of the Soil Biology book series (SOILBIOL, volume 44)

Abstract

A large amount of hazardous materials including heavy metals were released into the environment from natural and extensive anthropogenic activities, which cause soil, air, and water pollution and deterioration. At higher concentration, these metals exert toxic effects on plant and animal health including human. Among various traditional soil remediation technologies, use of phytoremediation to clean up metal(loid)-contaminated sites has gained increasing attention as an inexpensive, eco-friendly, and publicly acceptable remediation technology but has experienced varied successes in practice. Recent scientific discoveries that resulted from the application of molecular biology, bioinformatics, omics, and next-generation DNA sequencing technologies have assisted the remarkable impact of these immensely parallel platforms on genetics. In this context, genetic engineering has contributed rapid and significant changes in the crop improvement by offering a wide array of novel genes and traits which can be effectively inserted into candidate plants to raise its phytoremediation potential for metal

14

Microbial Degradation Mechanism of Textile Dye and Its Metabolic Pathway for Environmental Safety

Rahul V. Khandare and Sanjay P. Govindwar

CONTENTS

14.1 Introduction....400
14.2 Water Consumption in Textile Dying Processes is a Key Problem....401
14.3 Biological Methods for Dye Removal from Effluents....402
14.3.1 Bacteria as Remediators of Textile Dyes....403
14.3.1.1 Pure and Mixed Cultures of Bacteria for Dye Decolorization....403
14.3.1.2 Decolorization of Dyes under Anaerobic Conditions....404
14.3.1.3 Decolorization of Dyes under Anoxic Conditions....406
14.3.1.4 Decolorization of Dyes under Aerobic Conditions....408
14.4 Combinatorial Systems of Bacteria with Fungi and/or Plants Put Up a Better Fight....409
14.5 Development of Bioreactors to Explore the Bacterial Dye Degradation Potential....411
14.6 Understanding the Mechanism of Dye Metabolism by Bacteria....412
14.6.1 Bacterial Enzymes to Breakdown the Complex Dye Structure....413
14.6.1.1 Lignin Peroxidase (EC 1.11.1.14)....413
14.6.1.2 Aryl Alcohol Oxidase (EC 1.1.3.7)....413
14.6.1.3 Laccase (EC 1.10.3.2)....414
14.6.1.4 Azo Reductase (EC 1.7.1.6)....415
14.6.1.5 NADH-DCIP Reductase (EC 1.6.99.3)....416
14.6.1.6 Tyrosinase (E.C. 1.14.18.1)....416
14.6.1.7 Flavin Reductase (EC 1.5.1.30)....417
14.6.2 Analyses of Dyes and Effluents and Their Degradation Metabolites....418
14.6.3 Metabolic Pathway of Dyes Involving Bacterial Enzymes....421
14.7 Factors Affecting Bacterial Remediation of Textile Dyes....422
14.7.1 Oxygen....423
14.7.2 Pollutant Availability....423
14.7.3 Dye Structure and Concentration....423
14.7.4 Temperature....424
14.7.5 pH....424
14.7.6 Electron Donor....425
14.7.7 Redox Mediator and Its Potential....425
14.8 Toxicity Studies of Dyes and Their Degradation Products....425
14.8.1 Microbial Toxicity Studies....426
14.8.2 Phytotoxicity Studies....426
14.8.3 Animal Toxicity Study....427
14.8.4 Cytotoxicity Studies....427

Banana: Genomics and Transgenic Approaches for Genetic Improvement pp 93-105 | Cite as

Molecular Analysis of Fruit Ripening in Banana

Authors Authors and affiliations

Antara Ghosh, T. R. Ganapathi, V. A. Bapat

Chapter
First Online: 09 September 2016

Abstract

Banana is a climacteric fruit and has a very short postharvest life. Many varieties of banana fruits are ripened artificially by treating them with hydrocarbons. The current methods of postharvest management practices used for fruits are not enough to control the ripening in banana. Recent advances in recombinant DNA technology and genetic engineering have resulted in the modification of fruit ripening in banana. Towards this, many genes involved in ripening have been cloned and characterized. Ripening in banana is characterized by a biphasic ethylene production with a sharp early peak and a post climacteric small peak. During banana fruit ripening, ethylene production induces a developmental cascade which results in the conversion of starch into sugars, an associated burst of respiratory activity, and an increase in the protein synthesis. Other changes include fruit softening, flavor and aroma development, change in pigmentation, and increased susceptibility to pathogens; also, banana fruit softening is attributed to activities of various cell wall hydrolases. The participation of various cell wall hydrolases in banana softening during ripening has also been reported recently. The enhancing and suppressive effects of ABA and IAA on activities of different cell wall hydrolases have been noticed during ethylene-induced ripening in banana. Simultaneously, decrease in polyphenols, higher alcohol acetyl transferase activity, chlorophyll degradation etc., have been earlier reported during ripening in banana. Recently efforts are made to delay the ripening by

Banana: Genomics and Transgenic Approaches for Genetic Improvement pp 261-275 | Cite as

Molecular Farming: Prospects and Limitation

Authors | Authors and affiliations

Himanshu Tak, Sanjana Negi, T. R. Ganapathi ✉, V. A. Bapat

Chapter
First Online: 09 September 2016

Abstract

Plant molecular farming is the production of recombinant pharmaceutical and nonpharmaceutical proteins of commercial importance utilizing plants as bioreactors. Research and development on plant-derived recombinant proteins have gained momentum in recent years. Advantages of employing plants as bioreactors for recombinant protein generation are many including low cost of production, easier scale-up, cost-effective storage, and absence of animal pathogens in protein preparations. This article reviews the various technologies developed for employing plants as bioreactors, different plant systems being used as expression host, and limitations and research advances to overcome these limitations. An overview of different plant-derived products whether currently in market or are in different stages of development, including phases of clinical trials, is described. Special emphasis has been given on banana being used as an expression host, advantages and limitations of using banana in plant molecular farming, and different approaches which can be utilized to overcome those limitations have been described.

Keywords

Molecular farming | Recombinant pharmaceuticals | Magnification | Glycosylation

Chapter

Phytoremediation as a Green and Clean Tool for Textile Dye Pollution Abatement in Phytoremediation of Environmental Pollutants.

January 2017

In book: Phytoremediation of Environmental Pollutants · Chapter: Phytoremediation as a Green and Clean Tool for Textile Dye Pollution Abatement · Publisher: CRC Press, Taylor & Francis Group, Boca Raton, FL 33487-2742 · Editors: Ram Chandra, N. K. Dubey, Vineet Kumar

Projects: Construction of wetland, phytobeds for efficient cleaning of textile effluent. · Development of phytoreactors for cost effective and ecofriendly treatment of textile effluents

Niraj Rane · Rahul V. Khandare · Anuprita D Watharkar · Sanjay P Govindwar

Chapter 15
Hairy Roots: Production of Metabolites to Environmental Restoration

N.S. Desai, P. Jha, and V.A. Bapat

Abstract Hairy roots (HRs) have been proven as a potential source of secondary metabolites and also, for the biotransformation of desirable metabolites. Recently, HRs have emerged as an efficient *in vitro* model systems for screening the capabilities of different plant species to tolerate, accumulate, and/or to remove environmental pollutants. HRs offer benefits of greater genotypic and phenotypic stability than the dedifferentiated cultures, thus providing a more reliable and a reproducible experimental system, and even for flexibility of insertion of gene of interest to the HR gene construct for efficient applications. Additionally, absence of soil matrix and microbes is the key advantage in HRs for precise removal of toxic products as well as for elucidating metabolic pathways for conversion of hazardous chemicals to non hazardous products. The feasibility of scale up of HRs in bioreactors offers an attractive avenue for industrial processes both for metabolite synthesis as well as for phytoremediation. The present review highlights current knowledge, recent progress, areas which need to be explored and future perspectives related to the application and improvement of the efficiency of HRs for phytoremediation research.

Keywords Hairy roots (HRs) • Inorganic pollutant • Organic pollutant • Phytoremediation

N.S. Desai (✉)
Department of Biotechnology and Bioinformatics, D.Y. Patil University, Navi Mumbai 400614, India
e-mail: neetindesai@gmail.com

P. Jha
Department of Biotechnology, Centre for Energy Biosciences, Institute of Chemical Technology, Matunga, Mumbai 400019, India

V.A. Bapat
Department of Biotechnology, Shivaji University, Kolhapur 416004, India

K.-Y. Paek et al. (eds.), *Production of Biomass and Bioactive Compounds Using Bioreactor Technology*, DOI 10.1007/978-94-017-9223-3_15

In: Ecological Restoration
Editor: Victor R. Squires
ISBN: 978-1-63484-611-0

Chapter 5

ROLE OF MICROBES AND PLANTS IN PHYTOREMEDIATION: POTENTIAL OF GENETIC ENGINEERING

Umesh B. Jagtap[1], Vishwas A. Bapat[1], Gaëlle Saladin[2], Ewa Chudzińska[3], Magdalena Krzesłowska[4], Ewa M. Pawlaczyk[3], Tayyaba Komal[5], Alvina Gul[5], Irena Sherameti[6] and Zeshan Ali[7,*]

[1]Department of Biotechnology, Shivaji University Vidyanagar, Kolhapur, India
[2]Laboratoire de Chimie des Substances Naturelles, Université de Limoges, Faculté des Sciences et Techniques, Limoges Cedex, France
[3]Department of Genetics, Faculty of Biology, Adam Mickiewicz University, Umultowska, Poznań, Poland
[4]Laboratory of General Botany, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland
[5]Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology (NUST), Islamabad, Pakistan
[6]Friedrich-Schiller University Jena, Institute of General Botany and Plant Physiology, Jena, Germany
[7]National Institute of Bioremediation, National Agricultural Research Centre, Islamabad, Pakistan

ABSTRACT

Toxic metal pollution of soils is a major environmental problem. This review chapter focuses on the progress achieved in the last years to remediate soils contaminated with heavy metals. Genetically modified plants for metal removal, use of microbes in

* Corresponding Author Email: eco4nd@yahoo.com.

Xenobiotics in the Soil Environment pp 197-215 | Cite as

Transgenic Approaches for Building Plant Armor and Weaponry to Combat Xenobiotic Pollutants: Current Trends and Future Prospects

Authors Authors and affiliations

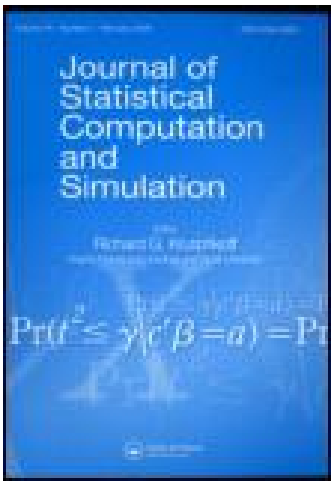
Umesh B. Jagtap, Vishwas A. Bapat

Chapter
First Online: 16 February 2017

Part of the Soil Biology book series (SOILBIOL, volume 49)

Abstract

Environment and living organisms are often threatened by xenobiotic contaminants released into the environment from extensive anthropogenic activities. Some plants and plant-associated microbes have developed a massive arsenal of specialized tactics to combat xenobiotic pollutants, while other plants were unable to do so. However, transgenic approaches offer various ways to keep them secure as well as building plant armor and weaponry to combat xenobiotic pollutants. This book chapter highlights various strategies for genetic engineering of plant and associated microbes considering the fine-tuning of transgene in transgenic plants/microbes for a better remediation response and constraints in lab to land transfer. Furthermore, a role of modern scientific and technological advances in addition to synthetic biology for building ultimate plants/microbes with enhanced remediation potential is also discussed.



Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

A modified test for testing exponentiality using transformed data

B. R. Dhumal[a] & D. T. Shirke[b]

[a] Krantisinh Nana Patil College, Walwe, Sangli, Maharashtra, India

[b] Department of Statistics, Shivaji University, Kolhapur, Maharashtra, India

Published online: 01 Aug 2012.

To cite this article: B. R. Dhumal & D. T. Shirke (2014) A modified test for testing exponentiality using transformed data, Journal of Statistical Computation and Simulation, 84:2, 397-403, DOI: 10.1080/00949655.2012.710850

To link to this article: http://dx.doi.org/10.1080/00949655.2012.710850

Journal of Statistical Computation and Simulation, 2014
Vol. 84, No. 2, 397–403, http://dx.doi.org/10.1080/00949655.2012.710850

Taylor & Francis
Taylor & Francis Group

A modified test for testing exponentiality using transformed data

B.R. Dhumal[a]* and D.T. Shirke[b]

[a]*Krantisinh Nana Patil College, Walwe, Sangli, Maharashtra, India;* [b]*Department of Statistics, Shivaji University, Kolhapur, Maharashtra, India*

(*Received 2 February 2012; final version received 7 July 2012*)

In this article, we present a test for testing uniformity. Based on the test, we provide a test for testing exponentiality. Empirical critical values for both the tests are computed. Both the tests are compared with the tests proposed by Noughabi and Arghami [H. Alizadeh Noughabi, and N.R. Arghami, *Testing exponentiality using transformed data*, J. Statist. Comput. Simul. 81 (4) (2011), pp. 511–516] using simulation experiments for a wide class of alternatives. The tests possess attractive power properties.

Keywords: test for uniformity; test for exponentiality; nonparametric kernel density estimation

1. Introduction

In much statistical inference and model building, there is a need to test the validity of assumptions made about the underlying populations from which observations have been drawn. Many times an experimenter begins the analysis of data by proposing a probability distribution for the observed data. If the assumption regarding underlying distribution is not tested, we may lead to incorrect conclusions and questions may be raised on the reliability of results obtained using the assumption. An exponential distribution is probably one of the most commonly used distributions in statistical work after normal distribution. It has important connections with life testing, reliability theory, theory of stochastic processes and is closely related to several other well-known distributions with statistical applications, for example, the gamma and the Weibull distributions.

Recently, Noughabi and Arghami [1] have provided a test for testing uniformity and using this test they have proposed a test for testing exponentiality. We have modified the above test and our simulation study reveals that the modified test gives better performance than the test proposed by them. The rest of the article is organized as follows.

In Section 2, we propose a test by modifying the test for uniformity due to Noughabi and Arghami [1] and study power of the same for various alternatives. Section 3 uses the modified test for testing uniformity for constructing a test for testing exponentiality. Monte Carlo study to estimate power of the test is discussed in Section 4. Section 5 gives concluding remarks.

*Corresponding author. Email: brd_stats@yahoo.in

2. Testing uniformity

Suppose a random sample $X_1, X_2, \ldots, X_n$ from a population with absolutely continuous density function $f(x)$ concentrated on the interval [0, 1] and having distribution function $F(x)$ is available. Consider the problem of testing the following null hypothesis,

H_0: A random sample of n X-values come from uniform distribution, denoted by $U(0, 1)$.

The test statistic proposed by Noughabi and Arghami [1] for testing H_0 is

$$T = \frac{1}{n}\sum_{i=1}^{n} |x_i \hat{f}(x_i) - F_0(x_i)|,$$

where $F_0(x)$ is the uniform distribution function. Also,

$$\hat{f}(x_i) - \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right),$$

where the kernel function $K(\cdot)$ is chosen to be the standard normal density function. The bandwidth h is obtained from the normal optimal smoothing formula, $h = 1.06sn^{-1/5}$, where s is the sample standard deviation [2].

2.1. *Test based on S statistic*

To test H_0, we suggest the following modified test statistic:

$$S = \frac{1}{n}\sum_{i=1}^{n} |\hat{f}(x_i) - f_0(x_i)|,$$

where $f_0(\cdot)$ is the probability density function of uniform. Also,

$$\hat{f}(x_i) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{h}\right),$$

kernel function $K(\cdot)$ is chosen to be the standard normal density function and bandwidth h is obtained from the normal optimal smoothing formula, $h = 1.06sn^{-1/5}$, where s is the sample standard deviation.

Large values of S indicate that the sample is from a non-uniform distribution. Therefore, we reject the null hypothesis at the significant level α, if $S \geq C(\alpha)$. The critical point $C(\alpha)$ is determined by the αth quantile of the distribution of the S statistic by means of Monte Carlo simulations. In Table 1, we present the results of Monte Carlo study conducted at an α-nominal level with 10,000 replications to assess the empirical critical values for S statistic.

Table 1. Critical values of S statistic.

	α		
n	0.01	0.05	0.10
5	2.031	1.037	0.702
10	0.731	0.503	0.398
15	0.579	0.406	0.338
20	0.477	0.358	0.299
25	0.425	0.324	0.280
30	0.397	0.305	0.266
50	0.324	0.265	0.231

2.2. *Performance study of test based on S statistic*

Noughabi and Arghami [1] have used several statistical tests that first appeared in Stephens [3]. These statistical tests are

(1) Kolmogorov–Smirnov (D) test,
(2) Kuiper's (V) test,
(3) the Cramér–von Mises (W^2) test,
(4) the Watson (U^2) test,
(5) the Anderson–Darling (A^2) test.

To study the performance of our test we consider the above tests, along with the test due to Noughabi and Arghami [1].

The null hypothesis is that we have a uniform random number on the interval (0, 1). The seven alternative distributions, which have been considered by several authors for studying power of

Table 2. Power comparisons of the tests for uniform (0, 1) with size 0.10.

n	Alternative	D	W^2	V	U^2	A^2	T	S
10	$F_{k=1.5}$	0.250	0.270	0.182	0.189	0.258	0.097	0.220
	$F_{k=2.0}$	0.525	0.574	0.334	0.335	0.551	0.116	0.408
	$G_{k=1.5}$	0.086	0.070	0.220	0.232	0.050	0.313	0.330
	$G_{k=2.0}$	0.123	0.092	0.446	0.479	0.059	0.581	0.602
	$G_{k=3.0}$	0.234	0.235	0.823	0.870	0.175	0.913	0.924
	$H_{k=1.5}$	0.198	0.174	0.218	0.228	0.208	0.045	0.054
	$H_{k=2.0}$	0.308	0.258	0.454	0.480	0.365	0.038	0.049
20	$F_{k=1.5}$	0.405	0.447	0.260	0.265	0.437	0.146	0.298
	$F_{k=2.0}$	0.811	0.860	0.596	0.582	0.857	0.218	0.609
	$G_{k=1.5}$	0.129	0.110	0.346	0.371	0.100	0.454	0.471
	$G_{k=2.0}$	0.265	0.273	0.730	0.783	0.289	0.817	0.827
	$G_{k=3.0}$	0.671	0.798	0.987	0.996	0.837	0.996	0.997
	$H_{k=1.5}$	0.251	0.214	0.340	0.372	0.269	0.047	0.052
	$H_{k=2.0}$	0.466	0.425	0.730	0.781	0.556	0.061	0.051

Table 3. Power comparisons of the tests for uniform (0, 1) with size 0.05.

n	Alternative	D	W^2	V	U^2	A^2	T	S
10	$F_{k=1.5}$	0.159	0.169	0.101	0.103	0.163	0.042	0.133
	$F_{k=2.0}$	0.400	0.435	0.232	0.224	0.417	0.050	0.280
	$G_{k=1.5}$	0.040	0.027	.130	0.137	0.015	0.189	0.212
	$G_{k=2.0}$	0.048	0.023	0.313	0.339	0.010	0.410	0.456
	$G_{k=3.0}$	0.095	0.053	0.713	0.760	0.021	0.810	0.852
	$H_{k=1.5}$	0.112	0.099	0.128	0.141	0.127	0.024	0.030
	$H_{k=2.0}$	0.206	0.158	0.311	0.335	0.235	0.025	0.025
20	$F_{k=1.5}$	0.281	0.316	0.167	0.164	0.318	0.068	0.184
	$F_{k=2.0}$	0.699	0.770	0.468	0.440	0.761	0.094	0.468
	$G_{k=1.5}$	0.056	0.039	0.224	0.246	0.035	0.324	0.321
	$G_{k=2.0}$	0.122	0.101	0.591	0.651	0.103	0.707	0.714
	$G_{k=3.0}$	0.411	0.508	0.969	0.984	0.561	0.987	0.988
	$H_{k=1.5}$	0.149	0.122	0.225	0.243	0.162	0.023	0.025
	$H_{k=2.0}$	0.310	0.248	0.593	0.652	0.378	0.054	0.045

various test statistics, are

$$F: \quad F(x) = 1 - (1-x)^k, \quad 0 \le x \le 1$$

for k equal to 1.5 and 2.

$$G: \quad F(x) = \begin{cases} 2^{(k-1)}x^k & 0 \le x \le 0.5, \\ 1 - 2^{(k-1)}(1-x)^k & 0.5 \le x \le 1 \end{cases}$$

for k equal to 1.5, 2 and 3

$$H: \quad F(x) = \begin{cases} 0.5 - 2^{(k-1)}(0.5-x)^k & 0 \le x \le 0.5, \\ 0.5 + 2^{(k-1)}(x-0.5)^k & 0.5 \le x \le 1 \end{cases}$$

for k equal to 1.5 and 2.

Alternative F, G and H were first used by Stephens [3] in his study of power comparisons of several tests for uniformity. According to Stephens, alternative F gives points closer to zero than expected under the hypothesis of uniformity, whereas G gives points near to 0.5 and H give two clusters (close to 0 and 1). The same were used by Noughabi and Arghami [1].

For the nominal levels 5% and 10%, Tables 2 and 3 show the power estimates of the test based on S statistic and also for the tests mentioned above. The entries are the 10,000 Monte Carlo samples of size $n = 10, 20$ that resulted in the rejection of H_0.

3. Testing exponentiality using transformed data

Suppose that n independent observations are made on X with density $f(x)$ over a non-negative support and with mean $\lambda^{-1} < \infty$, the hypothesis of interest is

$$H_0 : f(x) = f_0(x) = \lambda e^{-\lambda x},$$

where λ is unspecified. The alternative to H_0 is

$$H_1 : f(x) \ne f_0(x).$$

Noughabi and Arghami [1] have proposed a goodness-of-fit test for testing H_0. The test statistic is based on the following theorem, which is due to Alzaid and Al-Osh [4] and is also mentioned in [5].

THEOREM 3.1 *Let X_1 and X_2 be two independent observations from a distribution F. Then $X_1/(X_1 + X_2)$ is distributed as $U(0,1)$ if and only if F is exponential*

Let $X_{(i)}, i = 1, \ldots, n$, be the order statistics of a random sample of size n. Furthermore, transform the sample data to

$$Y_i j = \frac{X_{(i)}}{X_{(i)} + X_{(j)}} \quad i \ne j, i, j = 1, 2, \ldots, n.$$

Using Theorem 3.1, under the null hypothesis H_0, each Y_i has a uniform distribution. Noughabi and Arghami [1] used the proposed test for uniformity (introduced in Section 2) to test the uniformity

Table 4. Critical values of U statistic.

	α		
n	0.01	0.05	0.10
5	1.297	0.626	0.419
10	0.636	0.388	0.287
15	0.463	0.309	0.244
20	0.384	0.265	0.215
25	0.326	0.251	0.201
30	0.315	0.228	0.188
50	0.225	0.139	0.109

of Y_i's and thus the exponentiality of X_i's. The test statistic due to Noughabi and Arghami [1] is

$$T' = \frac{1}{n(n-1)} \sum_{i=1}^{n(n-1)} |y_i \hat{f}(y_i) - F_0(y_i)|.$$

Based on the modified test for uniformity proposed in Section 2, we define the U statistic as

$$U = \frac{1}{n(n-1)} \sum_{i=1}^{n(n-1)} |\hat{f}(y_i) - f_0(y_i)|.$$

Large values of U indicate that the sample is from a non-exponential distribution. In Table 4, we present the results of Monte Carlo study conducted at an α-nominal level with 10,000 replications (from exponential distribution with mean one) to assess the empirical critical values for U statistic.

4. Performance study of the test based on U statistic

For power comparisons, we considered the following alternatives:

(1) The Weibull distribution with density function

$$f(x) = \beta\lambda^{\beta}(x)^{\beta-1} \exp[-(\lambda x)^{\beta}], \quad x \geq 0, \lambda > 1, \beta > 0.$$

(2) The gamma distribution with density function

$$f(x, \beta, \lambda) = \frac{\lambda^{\beta} x^{\beta-1} \exp(-\lambda x)}{\Gamma(\beta)}, \quad x \geq 0, \lambda > 1, \beta > 0.$$

(3) The log-normal distribution with density function

$$f(x, \nu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(\ln(x) - \nu)^2\right\}, \quad x > 0, -\infty < \nu < \infty, \sigma > 0$$

We choose the parameters so that $E(X) = 1$, i.e. $\lambda = \Gamma(1 + 1/\beta)$ for the Weibull, $\lambda = \beta$ for the gamma and $\nu = \sigma 2/2$ for the log-normal family of distributions.

For the nominal levels 5% and 10%, Tables 5–7 show the power estimates of the test based on U statistic and the test proposed by Noughabi and Arghami [1] for testing exponentiality. The entries are the 10,000 Monte Carlo samples of size $n = 10, 20$, which resulted in the rejection of H_0.

Table 5. Power comparisons of the tests for exponential against the Gamma distribution.

n	β	α	T'	U
10	2	0.01	0.107	0.119
		0.05	0.338	0.346
	3	0.01	0.301	0.335
		0.05	0.662	0.672
	4	0.01	0.523	0.555
		0.05	0.861	0.867
20	2	0.01	0.256	0.342
		0.05	0.619	0.642
	3	0.01	0.714	0.794
		0.05	0.946	0.953
	4	0.01	0.932	0.962
		0.05	0.997	0.998

Table 6. Power comparisons of the tests for exponential against the Weibull distribution.

n	β	α	T'	U
10	2	0.01	0.354	0.386
		0.05	0.695	0.702
	3	0.01	0.858	0.878
		0.05	0.981	0.983
	4	0.01	0.987	0.989
		0.05	1.000	1.000
20	2	0.01	0.722	0.836
		0.05	0.959	0.964
	3	0.01	1.000	1.000
		0.05	1.000	1.000
	4	0.01	1.000	1.000
		0.05	1.000	1.000

Table 7. Power comparisons of the tests for exponential against the log-normal distribution.

n	v	α	T'	U
10	−0.3	0.01	0.089	0.102
		0.05	0.307	0.315
	−0.2	0.01	0.252	0.280
		0.05	0.611	0.619
	−0.1	0.01	0.761	0.790
		0.05	0.970	0.972
20	−0.3	0.01	0.304	0.294
		0.05	0.600	0.616
	−0.2	0.01	0.713	0.735
		0.05	0.941	0.944
	−0.1	0.01	0.998	1.000
		0.05	1.000	1.000

5. Concluding remark

Simulation results presented in Tables 2 and 3 show that for almost all alternatives the modified test for testing uniformity performs better than the test due to Noughabi and Arghami [1] for small and moderate sample sizes. Even though the results are less competitive than the remaining group

of tests especially for H and F alternatives, our aim was to develop a modified test for testing exponentiality and not for testing uniformity.

The simulation study presented in Tables 5–7 shows the superiority of the modified test for testing exponentiality over the test proposed by Noughabi and Arghami [1] for testing exponentiality for the Weibull, gamma as well as log-normal alternatives. The test possesses attractive power properties for small and moderate sample sizes.

References

[1] H. Alizadeh Noughabi, and N.R. Arghami, *Testing exponentiality using transformed data*, J. Statist. Comput. Simul. 81 (4) (2011), pp. 511–516.
[2] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, 1986.
[3] M.A. Stephens, *EDF statistics for goodness-of-fit and some comparisons*, J. Am. Statist. Assoc. 69 (347) (1974), pp. 730–737.
[4] A.A. Alzaid and M.A. Al-Osh, *Characterization of probability distributions based on the relation* $X = U(X_1 + X_2)$, Sankhya Ser. B 53 (1992), pp. 188–190.
[5] N. Balakrishnan and A.P. Basu, *The Exponential Distribution: Theory, Methods and Applications*, Gordon and Breach, Amsterdam, 1995.

Medha Nanivadekar

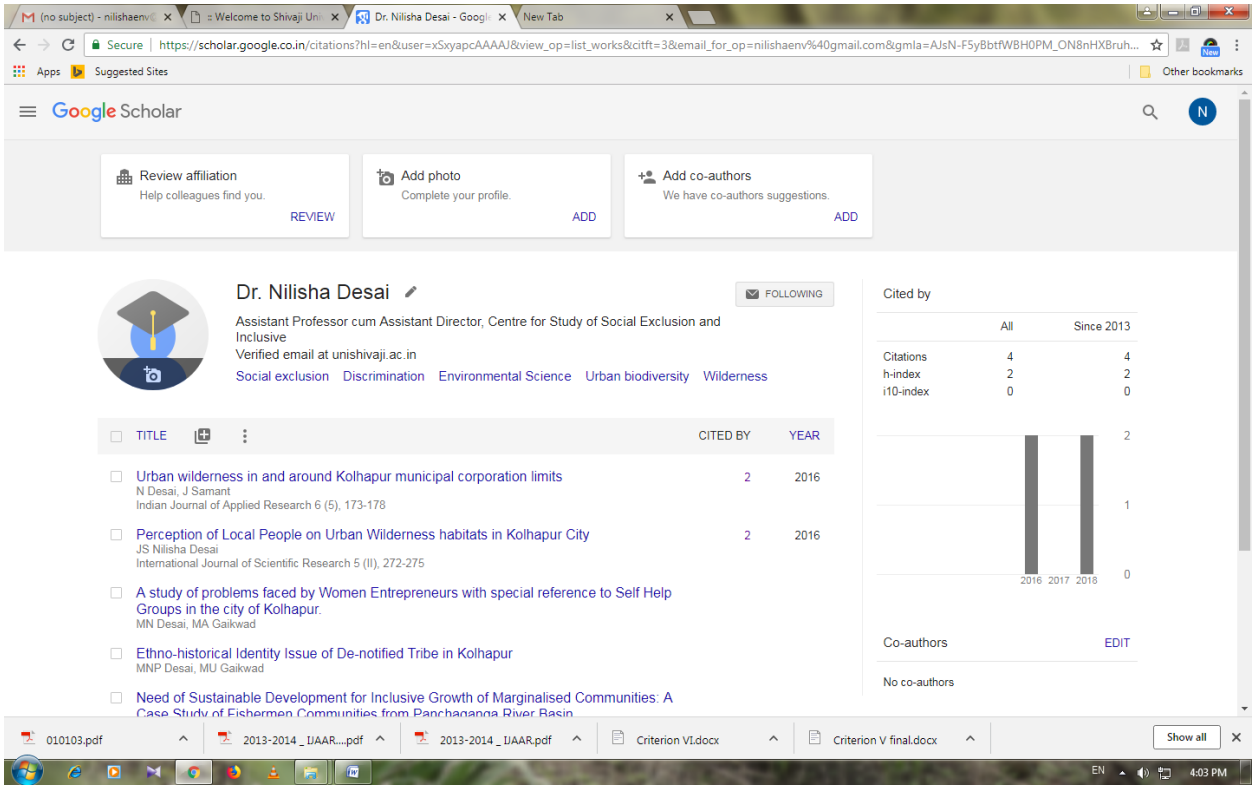
Associate Professor of Women's Studies, Shivaji University, Kolhapur
, INDIA

Women and Politics

	All	Since 2013
Citations	137	53
h-index	5	3
i10-index	2	1

TITLE	CITED BY	YEAR
Are quotas a good idea? The Indian experience with reserved seats for women M Nanivadekar Politics & Gender 2 (1), 119-128	74	2006
Reservation for Women: Challenge of Tackling Counter-Productive Trends M Nanivadekar Economic and political weekly, 1815-1819	16	1998
Empowering Women: Assessing the Policy of Reservations in Local Bodies: a Report M Nanivadekar Rambhau Mhalgi Prabodhini	8	1997
Reservation for Women in Local Bodies: Lessons from Maharastra M Nanivadekar Unpublished manuscript	7	2003
Indian experience of women’s quota in local government: implications for future strategies M Nanivadekar UNDAW, Expert group on the equal participation of women and men in decision …	6	2005
Dual-Member Constituencies: Resolving Deadlock on Women's Reservation M Nanivadekar Economic and Political Weekly, 4506-4510	5	2003
Reservation for Women M Nanivadekar Economic and Political Weekly, 1815	4	1998
Electoral Process in Corporation Elections: A Gender Study M Nanivadekar Bharatiya Stree Shakti, Mumbai, 20-22	4	1997
Women’s quota in urban local government: A cross-national comparison M Nanivadekar A paper presented at an International Seminar organized by the French …	3	2003
Partners in Politics, Competing in Crime Growing Importance of Independents and Rebels in Local Elections: Part Two M Nanivadekar Manushi, 26-34	3	2003

TITLE	CITED BY	YEAR
Feminist Fundamentalism over Women's Reservation Bill: Lessons from the Quota Debate in India M Nanivadekar, FS Scholar IWPR's Eighth International Women's Policy Research Conference on 'When …	2	2005
Feasible Option of Dual-Member Constituencies: Do We Have the Political Will M Nanivadekar Indian Legislator	2	2003
Electoral Process in Corporation Election M Nanivadekar A Gender	2	1997
Dual Member Constituencies: Resolving Deadlock on Women's Reservation' M Nanivadekar Reservation for Women, 355-70	1	2008
Overview: Women's Leadership in the Global Context M Nanivadekar Gender and Women's Leadership: A Reference Handbook 1, 293-303		2010
Partners in politics, competing in crime: fallouts of women's reservations in Maharashtra (part I) M Nanivadekar		2003


Google Scholar
Review affiliation
Help colleagues find you.
REVIEW
Add photo
Complete your profile.
ADD
Add co-authors
We have co-authors suggestions.
ADD
Dr. Nilisha Desai
FOLLOWING
Assistant Professor cum Assistant Director, Centre for Study of Social Exclusion and Inclusive
Verified email at unishivaji.ac.in
Social exclusion Discrimination Environmental Science Urban biodiversity Wilderness
TITLE
CITED BY
YEAR
Urban wilderness in and around Kolhapur municipal corporation limits
N Desai, J Samant
Indian Journal of Applied Research 6 (5), 173-178
2
2016
Perception of Local People on Urban Wilderness habitats in Kolhapur City
JS Nilisha Desai
International Journal of Scientific Research 5 (II), 272-275
2
2016
A study of problems faced by Women Entrepreneurs with special reference to Self Help Groups in the city of Kolhapur.
MN Desai, MA Gaikwad
Ethno-historical Identity Issue of De-notified Tribe in Kolhapur
MNP Desai, MU Gaikwad
Need of Sustainable Development for Inclusive Growth of Marginalised Communities: A
Cited by
All
Since 2013
Citations 4 4
h-index 2 2
i10-index 0 0
2016 2017 2018
Co-authors
EDIT
No co-authors
010103.pdf
2013-2014 _ IJAAR....pdf
2013-2014 _ IJAAR.pdf
Criterion VI.docx
Criterion V final.docx
Show all
4:03 PM

Journal of Modern Applied Statistical Methods
November 2014, Vol. 13, No. 2, 131-150.

ISSN 1538 – 9472

Robust Winsorized Shrinkage Estimators for Linear Regression Model

Nileshkumar H. Jadhav
D.R.K College of Commerce
Kolhapur, India

Dattatraya N. Kashid
Shivaji University
Kolhapur, India

In multiple linear regression, the ordinary least squares estimator is very sensitive to the presence of multicollinearity and outliers in the response variable. To handle these problems in the data, Winsorized shrinkage estimators are proposed and the performance of these estimators is evaluated through mean square error sense.

Keywords: Multicollinearity, outliers, contaminated normal error, Winsorization, mean square error, multiple linear regression

Introduction

In the multiple linear regression model

$$Y = X\beta + \varepsilon, \tag{1}$$

Y is an vector of n observations on the response variable, X is an $n \times p$ matrix of independent variables known as regressor variables, β is a $p \times 1$ vector of unknown regression parameters and ε is an $n \times 1$ vector of unobserved random errors. Classically, it is assumed that the ε_i, $i = 1, 2, ..., n$, are independent and identically normally distributed with zero mean and constant variance σ^2.

It is well known that when the normality assumption holds, the ordinary least squares (OLS) estimator becomes a maximum likelihood estimator and the best linear unbiased estimator of the unknown regression parameters and has the smallest variance in the class of all linear unbiased estimators. However, the real life data often may not satisfy these assumptions and the violation of assumptions dramatically affects the OLS estimation and consequently the prediction based on the OLS estimator. In the literature, the effect of violation of assumptions has been

Nileshkumar H. Jadhav is an Assistant Professor of Statistics at the D.R.K. College of Commerce. Email at n.nil08@gmail.com. Dr. Kashid is a Professor in the Department of Statistics. Email him at dnk_stats@unishivaji.ac.in.

Statistical Methodology 19 (2014) 44–59

Contents lists available at ScienceDirect

Statistical Methodology

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

Subset selection in multiple linear regression in the presence of outlier and multicollinearity

Nileshkumar H. Jadhav [a,*], Dattatraya N. Kashid [b], Subhash R. Kulkarni [c]

[a] *Department of Community Medicine, Krishna Institute of Medical Sciences Deemed University, Malkapur, Karad, 415110, Maharashtra, India*

[b] *Department of Statistics, Shivaji University, Kolhapur, 416004, India*

[c] *[24]7 Inc., Innovations Labs, Bangalore, India*

ARTICLE INFO

Article history:
Received 23 June 2012
Received in revised form
4 February 2014
Accepted 6 February 2014

Keywords:
Multiple linear regression
Subset selection
Outlier
Multicollinearity
Jackknifed ridge *M*-estimator

ABSTRACT

Various subset selection methods are based on the least squares parameter estimation method. The performance of these methods is not reasonably well in the presence of outlier or multicollinearity or both. Few subset selection methods based on the *M*-estimator are available in the literature for outlier data. Very few subset selection methods account the problem of multicollinearity with ridge regression estimator.

In this article, we develop a generalized version of S_p statistic based on the jackknifed ridge *M*-estimator for subset selection in the presence of outlier and multicollinearity. We establish the equivalence of this statistic with the existing C_p, S_p and R_p statistics. The performance of the proposed method is illustrated through some numerical examples and the correct model selection ability is evaluated using simulation study.

1. Introduction

Consider the multiple linear regression model

$$Y = X\beta + \varepsilon, \tag{1.1}$$

* Corresponding author. Tel.: +91 9595840635.
E-mail addresses: n.nil08@gmail.com (N.H. Jadhav), dnkashid_in@yahoo.com (D.N. Kashid).

http://dx.doi.org/10.1016/j.stamet.2014.02.002
1572-3127/

where Y is a vector of n observations on the response variable, X is an $n \times k$ matrix of n observations on $(k-1)$ regressor variables with 1's in the first column, $\beta = (\beta_0, \beta_1, \ldots, \beta_{k-1})'$ is a vector of k unknown regression parameters and ε is an unknown random error assumed to follow normal distribution with zero mean and constant variance σ^2. Without loss of generality, we assume that the regressor variables are standardized in such a way that $X'X$ is in the form of a correlation matrix.

In the literature, various subset selection methods based on the least squares (LS) estimator are available like Mallows' C_p [13], stepwise selection methods. The Mallows' C_p is one of the most popular subset selection methods. It is defined as

$$C_p = \frac{RSS_p}{\sigma^2} - (n - 2p), \tag{1.2}$$

where RSS_p is the residual sum of squares of the subset model based on $(p-1)$ regressor variables, σ^2 is the error variance and is replaced by its suitable estimate $\hat{\sigma}^2 = \left(Y - X\hat{\beta}_{LS}\right)' (Y - X\hat{\beta}_{LS})/(n-k)$, $\hat{\beta}_{LS}$ is the LS estimator of β of the full model based on $(k-1)$ regressor variables.

It is well known that, the C_p statistic is based on the LS estimator and the LS estimator is very sensitive to the presence of outliers or the violation of the assumption of normality on the error variable (see Huber [9]). In the past three decades, many robust parameter estimation methods as well as subset selection methods have been devised. For instance, Ronchetti [17] proposed robust version of AIC called RAIC, Ronchetti and Statudte [18] proposed robust version of Mallows' C_p called RC_p, Sommer and Huggins [19] proposed RT_p criterion based on Wald test statistic, Kim and Hwang [12] defined $C_{p(k)}$ method, Kashid and Kulkarni [11] proposed an S_p criterion which is a more general criterion for subset selection in the presence of outlier in the data. The S_p criterion is operationally simple to implement as compared to the other robust subset selection methods; it is defined as

$$S_p = \frac{\sum_{i=1}^{n} \left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - (k - 2p), \tag{1.3}$$

where $\hat{Y}_{ik}$ and $\hat{Y}_{ip}$ are the predicted values of Y_i based on the full model and the subset model respectively. The unknown σ is replaced by its suitable estimate based on the full model as $\hat{\sigma} = 1.4826$ median$|r_i - \text{median}(r_i)|$, where r_i is the ith residual.

The presence of multicollinearity is also one of the most serious and frequently encountered problems in multiple linear regression. Due to the presence of multicollinearity, the variance of the LS estimator gets inflated and consequently, the LS estimator becomes unstable and gives misleading results. To overcome such a problem, Hoerl and Kennard [5,6] proposed the ordinary ridge regression (ORR) estimator. Recently, Dorugade and Kashid [3] proposed R_p statistic for subset selection based on the ORR estimator of β. It is defined as

$$R_p = \frac{\sum_{i=1}^{n} \left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - \text{tr}\left(H_R' H_R\right) + \text{tr}\left(H_{RA}' H_{RA}\right) + p, \tag{1.4}$$

where σ^2 is the error variance and is replaced by its suitable estimate $\hat{\sigma}^2 = (Y - X\hat{\beta}_R)'(Y - X\hat{\beta}_R)/(n-k)$ and $\hat{\beta}_R$ is the ORR estimator of β based on the full model. The matrix $H_R = X(X'X + rI)^{-1}X'$, $H_{RA} = X_A(X_A'X_A + r_AI)^{-1}X_A'$, r and r_A are the biasing constants known as ridge parameters. Note that, the above S_p and the R_p statistics are equivalent to Mallows' C_p when the LS estimator is used. Though the C_p, S_p and R_p Statistics are used for correct subset selection in different situations, the subset selection procedure of these three statistics is same and it is given as follows.

Subset selection procedure based on C_p, S_p and R_p statistics

Step I. Compute the value of statistic for all possible subset models.

Step II. Select a subset of minimum size, for which the value of the statistic is close to 'p' or plot the values of statistic vs. 'p' for all possible subset models and select the subset which is closer to the line 'statistic $= p$'.

Many researchers have pointed out that the M-estimator is a better alternative to the LS estimator in the presence of outliers (see Brikes and Dodge [1]) and the ORR estimator performs better in the presence of multicollinearity (see [5–7]). Brikes and Dodge [1], Montgomery et al. [16] have given description of these methods in the context of parameter estimation. However, these methods give misleading results when outlier and multicollinearity occur simultaneously in the data (see Jadhav and Kashid [10]).

To overcome the problem of simultaneous occurrence of outlier and multicollinearity, very recently, Jadhav and Kashid [10] proposed an estimator known as Jackknifed Ridge M- (JRM) estimator. They showed that, the performance of the JRM estimator is better in the mean square error sense when outliers and multicollinearity present in the data.

In this article, we have proposed a generalized S_p criterion, called as GS_p criterion for subset selection based on the JRM estimator when outlier and multicollinearity occurs simultaneously in the data.

The rest of the article is organized as follows. In Section 2, the effect of presence of multicollinearity and outliers on the existing subset selection criteria is demonstrated. Section 3 briefly introduces the various estimators which are used in this article. In Section 4, a motivation to propose a new subset selection criterion is presented and a subset selection criterion based on the JRM estimator is defined. Some results and the equivalence of GS_p statistic with C_p, R_p and S_p statistics are presented in Section 5. In Section 6, simulated data sets are considered to illustrate the performance of the proposed method. Also, a correct model selection ability of the GS_p statistic and the performance of various robust estimates of σ^2 are presented in Section 6. Finally, the article ends with some concluding remarks.

2. The problem

This section illustrates the problem of outlier and multicollinearity from the viewpoint of subset selection. The purpose of this section is to highlight the effect of simultaneous occurrence of outlier and multicollinearity on the subset selection criteria based on the LS estimator (C_p), M-estimator (S_p) and ORR estimator (R_p).

A simulation design given by McDonald and Galarneau [15] is used to introduce multicollinearity in the regressor variables as follows.

$$x_{ij} = \left(1 - \rho^2\right)^{1/2} z_{ij} + \rho z_{i(l+1)} \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, l \tag{2.1}$$

where z_{ij}'s are independent standard normal pseudo random numbers and ρ^2 is the correlation between any two regressor variables. Here, $l = 4$ and $\rho = 0.999$ are considered to generate $n = 20$ observations on the response variable Y using the regression model

$$Y_i = 10 + 3x_{i1} + 5x_{i2} + 0x_{i3} + 0x_{i4} + \varepsilon_i, \quad i = 1, 2, \ldots, 20,$$

where ε_i are independent and identically normally distributed with mean 0 and variance 0.25. A single outlier observation is introduced in the response variable corresponding to the maximum absolute residual value. The actual value of the response variable $Y_{13} = 1.5363$ is changed to $Y_{13} = 30.7260$. How the outlier observation is introduced in the response variable is given in Example 6.1 of Section 6. To identify the severity of the multicollinearity, the variance inflation factor (VIF) is considered (see Marquardt [14], Montgomery et al. [16]). For this data, the VIF for each term are 388.5831, 772.2531, 688.8542 and 296.6157. These VIFs indicate the presence of severe multicollinearity in the data. We compute the values of C_p, S_p and R_p statistics and are reported in Table 1. Also, we plot the values of C_p, S_p and R_p statistics of all possible subset models in Fig. 1.

From Table 1, it is observed that the criteria C_p, S_p and R_p select wrong subset models. Consequently, in Fig. 1, the values of C_p, S_p and R_p statistics are close to p for wrong subset models. It is clear that, when both outlier and multicollinearity present in the data, the C_p, S_p and R_p statistics fail to select the correct subset model or there is ambiguity concerns to the selection of correct model. This study indicates that subset selection method based on the LS estimator, M-estimator and ORR estimator fails to select the correct model when the outlier and multicollinearity occurs simultaneously in the data.

Table 1
Values of C_p, S_p and R_p statistics for all subset models.

Regressors in subset model	C_p	S_p	R_p
X_1	2.1903	5.4235	2.3993
X_2	2.7501	1.0843	2.7413
X_3	2.4824	5.0971	2.6124
X_4	1.9387	14.6250	2.2086
$X_1\,X_2$	2.9131	1.0914	2.9373
$X_1\,X_3$	3.9301	3.2280	3.0330
$X_1\,X_4$	3.8417	5.4157	2.8594
$X_2\,X_3$	3.9397	2.5536	3.1816
$X_2\,X_4$	1.7539	2.6440	2.5774
$X_3\,X_4$	3.1544	5.5966	2.8219
$X_1\,X_2\,X_3$	4.6727	3.0192	4.0344
$X_1\,X_2\,X_4$	3.0811	3.0597	3.7427
$X_1\,X_3\,X_4$	5.0529	4.9692	3.8366
$X_2\,X_3\,X_4$	3.5144	4.3136	3.7644
$X_1\,X_2\,X_3\,X_4$	5.0000	5.0000	5.0000

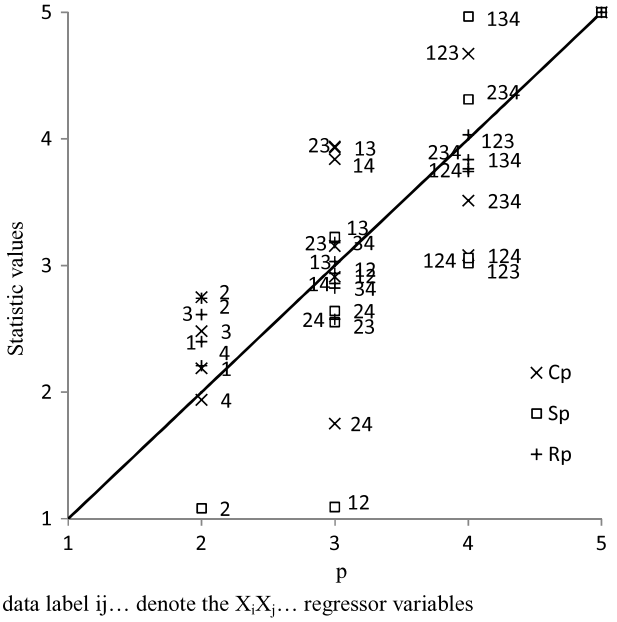


Fig. 1. Values of C_p, S_p and R_p statistics versus p.

3. The estimators

In the multiple linear regression, an important task is to estimate the unknown regression parameters β using an appropriate method of estimation. In this section, some of the existing estimation methods of β are briefly discussed as follows.

Least squares (LS) estimator

For the multiple linear regression model given in Eq. (1.1), the LS estimator of the unknown regression parameters β is defined as

$$\hat{\beta}_{LS} = \left(X'X\right)^{-1} X'Y. \tag{3.1}$$

Any standard textbook of regression like Draper and Smith [4], Montgomery et al. [16] give detailed description about the LS estimator.

M-estimator

To handle the problem of outliers, Huber [8] proposed the M-estimator for β. It is obtained by minimizing a sum of the function of the scaled residuals

$$\sum_{i=1}^{n} \rho\left(\frac{Y_i - X_i'\beta}{s}\right), \tag{3.2}$$

where $\rho : R \to R^+$ robust criterion function (Montgomery et al. [16]) and s is a scale parameter which is replaced by its suitable estimate. To minimize Eq. (3.2), differentiate Eq. (3.2) partially with respect to each parameter and equate it to zero, we get k nonlinear equations of the form

$$\sum_{i=1}^{n} \psi\left(\frac{Y_i - X_i'\beta}{s}\right) x_{ij}, \quad j = 0,\ 1,\ 2, \ldots, k-1, \tag{3.3}$$

where $\psi(\cdot)$ is partial derivative of $\rho(\cdot)$ and x_{ij} is jth entry in the ith row of matrix X with $x_{i0} = 1$. Solution to these k equations is obtained by iterative reweighted least squares method (see Draper and Smith [4]). At convergence, M-estimator may be given as

$$\hat{\beta}_M = \left(X'WX\right)^{-1} X'W\,Y, \tag{3.4}$$

where W is diagonal matrix of weight with ith diagonal element, $W_i = \psi\left(\frac{Y_i - X_i'\beta}{s}\right) / \left(\frac{Y_i - X_i'\beta}{s}\right)$.

Ordinary Ridge Regression (ORR) estimator

To overcome the problem of multicollinearity, various biased estimators are available in the literature. The ORR estimator proposed by Hoerl and Kennard [5,6] is one of the most popular biased estimators. It is defined as

$$\hat{\beta}_R = \left(X'X + rI\right)^{-1} X'\mathrm{Y}, \tag{3.5}$$

where r is a ridge parameter.

Jackknifed Ridge M (JRM) estimator

Jadhav and Kashid [10] proposed the JRM estimator of β to combat the simultaneous occurrence of outlier and multicollinearity in the data. It is defined as

$$\begin{aligned}\hat{\beta}_{JRM} &= \left[I - r^2 Q'\left(X'X + rI\right)^{-2} Q\right]\hat{\beta}_M \\ &= R\hat{\beta}_M,\end{aligned} \tag{3.6}$$

where $R = \left[I - r^2 Q'\left(X'X + rI\right)^{-2} Q\right]$, Q is the matrix of eigenvectors of $X'X$, $\hat{\beta}_M$ is the M-estimator of the unknown parameters β and r is the ridge parameter. The performance of this estimator is better in the MSE sense as compared to the LS estimator, M-estimator, and the ORR estimator when both outliers and multicollinearity present in the data.

Selection of ridge parameter r

To implement the ORR estimator, we need to obtain the value of ridge parameter r. Various methods are available in the literature for determining a ridge parameter r. Among these, the choice of the ridge parameter proposed by Hoerl et al. [7] is widely used to obtain the ORR estimator. It is given by

$$r = \frac{(k-1)\,\hat{\sigma}^2}{\hat{\beta}_{LS}'\hat{\beta}_{LS}}, \tag{3.7}$$

where $\hat{\sigma}^2 = \left(Y'Y - \hat{\beta}_{LS}'X'Y\right)/(n-k)$ (see Dorugade and Kashid [3]). Jadhav and Kashid [10] obtained the JRM estimator by replacing the values of r by $\tilde{r}$ and it is given by

$$\tilde{r} = \frac{(k-1)\,s^2}{\hat{\beta}_M'\hat{\beta}_M}, \tag{3.8}$$

where $s = 1.4826\,\text{median}|r_i - \text{median}(r_i)|$ and r_i is ith residual obtained using M-estimator. Based on the JRM estimator, we propose a generalized version of S_p statistic for subset selection in the presence of outlier and multicollinearity.

4. Proposed method

Consider the multiple linear regression model given in Eq. (1.1). Then the vector of predicted values of Y based on the JRM estimator of β is

$$\begin{aligned}\hat{Y}_k &= X\hat{\beta}_{JRM} \\ &= HY \end{aligned} \tag{4.1}$$

where $H = XR\left(X'WX\right)^{-1}X'W$ is the prediction matrix based on the full model. The full model is the one which contains all $(k-1)$ regressor variables.

The model given in Eq. (1.1) can be written as

$$Y = X_A\beta_A + X_B\beta_B + \varepsilon \tag{4.2}$$

where X and β are partitioned as $X = [X_A : X_B]$ and $\beta' = \left[\beta'_A : \beta'_B\right]$. The matrix X_A is of order $n \times p$ with 1s in the first column and the matrix X_B is of order $n \times (k-p)$. β_A and β_B are the vectors of parameters of order $p \times 1$ and $(k-p) \times 1$ respectively.

Consider the subset model based on the $(p-1)$ regressor variables

$$Y = X_A\beta_A + \varepsilon. \tag{4.3}$$

Suppose, $\hat{\beta}_{AJRM}$ be the JRM estimator of β_A based on the subset model given in Eq. (4.3), then the vector of predicted values of Y based on the JRM estimator is

$$\begin{aligned}\hat{Y}_p &= X_A\hat{\beta}_{AJRM} \\ &= H_1Y \end{aligned} \tag{4.4}$$

where $H_1 = X_AR_A\left(X'_AW_AX_A\right)^{-1}X'_AW_A$ denote the prediction matrix for the subset model based on the JRM estimator.

Based on the two predicted vectors $\hat{Y}_k$ and $\hat{Y}_p$ of Y, we propose a generalized version of S_p statistic. The main objective to define new criterion is that to select a subset model of size p $(\leq k)$ such that it will predict the response variable 'as accurate as' the full model.

4.1. Motivation

The motivation of the proposed method is similar to the S_p criterion proposed by Kashid and Kulkarni [11]. Let $\hat{Y}_{ik}$ and $\hat{Y}_{ip}$ be the predicted values of Y_i based on the JRM estimator for the full model and the subset model respectively. The quantity $\frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2}{\sigma^2}$ is small when $\hat{Y}_{ik}$ is close to $\hat{Y}_{ip}$ that is, the prediction based on the model with $(p-1)$ regressors is as accurate as that based on the model with $(k-1)$ regressors. When the quantity $\frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2}{\sigma^2}$ is large then $\hat{Y}_{ik}$ is far away from $\hat{Y}_{ip}$ and it implies that, prediction based on the model with $(p-1)$ regressors may not be as 'accurate' as that based on the model with $(k-1)$ regressors. Using the same logic, we propose the following criterion for subset selection.

4.2. Definition of GS_p

The proposed generalized S_p criterion is denoted by GS_p and is defined as

$$GS_p = \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2}{\sigma^2} - \text{tr}\left[\left(H-H_1\right)'\left(H-H_1\right)\right] + p \tag{4.5}$$

where σ^2 is the unknown error variance, replaced by its suitable estimate based on the JRM estimator, H and H_1 are the prediction matrices defined in the beginning of Section 4. The purpose of subtracting $\operatorname{tr}\left[(H-H_1)'(H-H_1)\right]-p$ from the first term $\frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2}{\sigma^2}$ is simply to compare the GS_p statistic with the dimensions of the subset model for the selection of a correct or an adequate model. A subset model is said to be correct or adequate, if the prediction based on the subset model is as accurate as that based on full model.

Below we discuss some results which are useful for implementing the proposed method. Here, note that the prediction matrix changes with respect to the change in the estimator of unknown regression parameters and the choice of the estimator is based on the nature of the data.

5. Some results

In this section, we present some results to support the use of the proposed criterion to select the correct subset model. Also, we have derived the equivalence of the proposed GS_p statistic with the C_p, S_p and R_p statistics.

Result 5.1. If the subset model is adequate then,

$$E\left(\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2\right)\cong\sigma^2\operatorname{tr}\left[(H-H_1)'(H-H_1)\right]. \tag{5.1}$$

Proof. Consider, $\hat{Y}_{ik}$ and $\hat{Y}_{ip}$ be the ith predicted values of Y based on the JRM estimator for the full model and the subset model respectively. Then, we can write,

$$\begin{aligned}\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2&=\left(\hat{Y}_k-\hat{Y}_p\right)'\left(\hat{Y}_k-\hat{Y}_p\right)\\&=(HY-H_1Y)'(HY-H_1Y)\\&=Y'(H-H_1)'(H-H_1)Y.\end{aligned} \tag{5.2}$$

Now,

$$E\left[\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2\right]=E\left[Y'(H-H_1)'(H-H_1)Y\right].$$

Since $\left[(H-H_1)'(H-H_1)\right]$ is a symmetric matrix, then using the properties of quadratic form, the expected value of the quadratic equation given in Eq. (5.2) can be written as,

$$E\left[\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2\right]=\sigma^2\operatorname{tr}\left[(H-H_1)'(H-H_1)\right]+\beta'X'\left[(H-H_1)'(H-H_1)\right]X\beta.$$

But, when the subset model is appropriate, the quantity $\beta'X'\left[(H-H_1)'(H-H_1)\right]X\beta\cong 0$, hence,

$$E\left[\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2\right]\cong\sigma^2\operatorname{tr}\left[(H-H_1)'(H-H_1)\right].$$

Result 5.2. If the subset model is adequate then

$$E\left[GS_p\right]\cong p. \tag{5.3}$$

Proof. When the subset model is adequate, from Result 5.1, we can write,

$$E\left[\frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2}{\sigma^2}\right] \cong \operatorname{tr}\left[(H-H_1)'(H-H_1)\right].$$

Therefore,

$$\begin{aligned} E\left[GS_p\right] &\cong \left[\operatorname{tr}\left[(H-H_1)'(H-H_1)\right]\right]-\left[\operatorname{tr}\left[(H-H_1)'(H-H_1)\right]\right]+p \\ &= p. \end{aligned} \tag{5.4}$$

Result 5.3.

$$\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2 = RSS_p - RSS_k - 2Y'(H-H'H-H_1'+H_1'H)Y. \tag{5.5}$$

Proof. We can write,

$$\begin{aligned} \sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2 &= \left(\hat{Y}_k-\hat{Y}_p\right)'\left(\hat{Y}_k-\hat{Y}_p\right) \\ &= \left(\hat{Y}_k-Y+Y-\hat{Y}_p\right)'\left(\hat{Y}_k-Y+Y-\hat{Y}_p\right) \\ &= \left((Y-\hat{Y}_p)-(Y-\hat{Y}_k)\right)'\left((Y-\hat{Y}_p)-(Y-\hat{Y}_k)\right) \\ &= \left(Y-\hat{Y}_p\right)'\left(Y-\hat{Y}_p\right)+\left(Y-\hat{Y}_k\right)'\left(Y-\hat{Y}_k\right)-2\left(Y-\hat{Y}_p\right)'\left(Y-\hat{Y}_k\right) \\ &= \left(Y-\hat{Y}_p\right)'\left(Y-\hat{Y}_p\right)+\left(Y-\hat{Y}_k\right)'\left(Y-\hat{Y}_k\right) \\ &\quad -2(Y-\hat{Y}_k+\hat{Y}_k-\hat{Y}_p)'\left(Y-\hat{Y}_k\right) \\ &= RSS_p+RSS_k-2RSS_k-2(\hat{Y}_k-\hat{Y}_p)'\left(Y-\hat{Y}_k\right) \\ &= RSS_p-RSS_k-2Y'(H-H'H-H_1+H_1'H)Y. \end{aligned} \tag{5.6}$$

The quantity in Eq. (5.6) depends on the estimator of β. It varies according to the choice of the estimator. The following results discuss the equivalence of the GS_p statistic with the C_p, S_p and R_p statistics.

Result 5.4. When the LS estimator $\left(\hat{\beta}_{LS}\right)$ of β is used to obtain the predicted values, then the GS_p statistic reduces to the C_p statistic.

Proof. Suppose the LS estimator $\left(\hat{\beta}_{LS}\right)$ of β is used to obtain the predicted values of the response variable, then the prediction matrices H and H_1 becomes $H = X\left(X'X\right)^{-1}X'$ and $H_1 = X_A\left(X_A'X_A\right)^{-1}X_A'$ respectively. Also, H and H_1 matrices are symmetric and idempotent. When the subset model is adequate, then the last quantity $Y'(H-H'H-H_1+H_1'H)Y$ of Eq. (5.6) becomes zero and the GS_p statistic becomes

$$\begin{aligned} GS_p &= \frac{RSS_p-RSS_k}{\sigma^2}-\operatorname{tr}\left[(H-H_1)'(H-H_1)\right]+p \\ &= \frac{RSS_p-RSS_k}{\sigma^2}-\operatorname{tr}\left[H'H-H'H_1-H_1'H+H_1'H_1\right]+p \\ &= \frac{RSS_p-RSS_k}{\sigma^2}-\operatorname{tr}\left(H'H\right)+\operatorname{tr}\left(H'H_1\right)+\operatorname{tr}\left(H_1'H\right)-\operatorname{tr}\left(H_1'H_1\right)+p. \end{aligned}$$

Using the idempotent property of H and H_1 matrices, we can write

$$\begin{aligned} GS_p &= \frac{RSS_p}{\sigma^2} - \frac{RSS_k}{\sigma^2} - \mathrm{tr}(H) + \mathrm{tr}\,(H_1) + \mathrm{tr}\left(H_1'\right) - \mathrm{tr}\,(H_1) + p \\ &= \frac{RSS_p}{\sigma^2} - \frac{RSS_k}{\sigma^2} - k + p + p - p + p \\ &= \frac{RSS_p}{\sigma^2} - (n-k) - (k-2p) \quad \left(\text{where } \sigma^2 \text{ is replaced by } \frac{RSS_k}{n-k}\right) \\ &= \frac{RSS_p}{\sigma^2} - (n-2p) \\ &= C_p. \end{aligned} \tag{5.7}$$

Thus, the GS_p statistic reduces to the C_p when the LS estimator $\hat{\beta}_{LS}$ of β is used.

Result 5.5. When the M-estimator $\left(\hat{\beta}_M\right)$ of β is used to obtain the predicted values, then the GS_p statistic reduces to the S_p statistic.

Proof. Let $\hat{\beta}_M$ be the M-estimator of β used to predict the response variable, then the prediction matrix H reduces to $H = X(X'WX)^{-1}X'W$ and H_1 reduces to $H_1 = X_A(X_A'W_AX_A)^{-1}X_A'W_A$.

Hence,

$$GS_p = \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - \mathrm{tr}\left(H'H\right) + \mathrm{tr}\left(H'H_1\right) + \mathrm{tr}\left(H_1'H\right) - \mathrm{tr}\left(H_1'H_1\right) + p.$$

The matrices H and H_1 are idempotent matrices and are not exact but close to symmetric matrices. Hence,

$$\begin{aligned} GS_p &\cong \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - \mathrm{tr}(H) + \mathrm{tr}\,(H_1) + \mathrm{tr}\left(H_1'\right) - \mathrm{tr}\,(H_1) + p \\ &\cong \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - k + p + p - p + p \\ &= \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - (k-2p) \\ &= S_p. \end{aligned} \tag{5.8}$$

Thus, the GS_p statistic is equivalent to the S_p statistic when the M-estimator of β is used.

Result 5.6. When the ORR estimator $\left(\hat{\beta}_R\right)$ of β is used to obtain the predicted values, then the GS_p statistic reduces to the R_p statistic.

Proof. Suppose, $\hat{\beta}_R$ is used to obtain the predicted values of the response variable, then matrix H and H_1 become $H_R = X(X'X + rI)^{-1}X'$ and $H_{RA} = X_A(X_A'X_A + r_AI)^{-1}X_A'$.

Hence, the GS_p statistic becomes,

$$\begin{aligned} GS_p &= \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - \mathrm{tr}\left[\left(H_R - H_{RA}\right)'\left(H_R - H_{RA}\right)\right] + p \\ &= \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - \mathrm{tr}\left[H_R'H_R - H_R'H_{RA} - H_{RA}'H_R + H_{RA}'H_{RA}\right] + p. \end{aligned}$$

We decompose H_R into sum of H_{RA} and H_{RB} such that, $H_{RA}H_{RB} = 0$ (see Chatterji and Hadi [2], Dorugade and Kashid [4]). Using this decomposition, we can write

$$\begin{aligned} GS_p &\cong \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - \operatorname{tr}\left[H_R'H_R - H_R'H_{RA} - H_{RA}'H_R + H_{RA}'H_{RA}\right] + p \\ &\cong \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - \operatorname{tr}\left[H_R'H_R - H_{RA}'H_{RA}\right] + p \\ &= \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} - \operatorname{tr}\left(H_R'H_R\right) + \operatorname{tr}(H_{RA}'H_{RA}) + p \\ &= R_p. \end{aligned} \tag{5.9}$$

Subset selection procedure based on the GS_p statistic

Using the Result 5.3, the subset selection procedure based on the GS_p statistic is given as follows:

Step I. Compute the value of the GS_p statistic for all possible subset models.

Step II. Select a subset of minimum size, for which the value of the GS_p statistic is close to 'p'.

In the following section, we study the correct subset selection performance of the C_p, S_p, R_p and GS_p statistics.

6. Simulation study

A simulation study is carried out to illustrate the performance of proposed method. A simulation study is divided into three subsections. Section 6.1 illustrate the performance of the C_p, S_p, R_p and GS_p criteria through numerical examples for all combinations of absence and presence of outlier and multicollinearity. A correct model selection ability of these criteria is evaluated in Section 6.2. Also, the various choices of the estimator of σ^2 are considered in Section 6.3 and their performance is studied through numerical example.

For the computation of M-estimator, Huber's robust criterion function is used. The form of $\rho(\cdot)$ is

$$\rho(z) = \begin{cases} \dfrac{1}{2}z^2 & |z| \le t \\ |z|\, t - \dfrac{1}{2}t^2 & |z| > t \end{cases}$$

where z is scaled residuals and $t = 1.345$.

6.1. Numerical examples

In this subsection, we have considered four cases: 1. Clean data, 2. Data with outlier, 3. Data with multicollinearity and 4. Data with outlier and multicollinearity. Example 6.1 illustrate the performance of the C_p, S_p, R_p and GS_p statistics for first two cases. However, Example 6.2 consider the last two cases to study the correct subset selection performance of the C_p, S_p, R_p and GS_p statistics.

Example 6.1. Here, we have considered the following regression model to generate $n = 20$ observations on the response variable Y as

$$Y_i = 1 + 2X_{i1} + 3X_{i2} + 0X_{i3} + 0X_{i4} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where $X_{ij} \sim U(0, 1)$, $i = 1, 2, \ldots, n$, $j = 1, 2, 3, 4$ and $\varepsilon_i \sim N(0, 0.25)$. A method to introduce outlier in the response variable suggested by Jadhav and Kashid [10] is implemented. A single outlier observation is introduced in the response variable corresponding to largest absolute residual by

Table 2
Values of C_p, S_p, R_p and GS_p statistics for all subset models.

Regressors in the model	Clean data				With one outlier data			
	C_p	S_p	R_p	GS_p	C_p	S_p	R_p	GS_p
X_1	41.4566	88.8442	38.9205	89.5066	6.4856	88.9771	3.9139	89.2078
X_2	21.8913	59.5440	20.1670	59.8542	3.7951	63.7876	2.8164	63.7516
X_3	55.3685	127.1283	57.0563	128.0798	4.5724	129.3113	3.0506	129.9719
X_4	59.6552	143.6816	94.6994	144.0413	6.8399	143.4294	4.1313	143.5348
$X_1\,X_2$	2.7568	3.2767	2.9584	3.1723	4.7240	3.3099	3.5026	2.9617
$X_1\,X_3$	42.1485	86.6213	38.7884	87.3354	6.2909	88.7644	3.9701	88.9678
$X_1\,X_4$	43.2062	91.3802	39.9829	92.1708	7.1991	91.0399	4.5070	91.3607
$X_2\,X_3$	19.6384	48.3448	17.5928	48.4336	2.1995	53.3889	2.5820	53.1513
$X_2\,X_4$	19.9072	57.8829	18.1727	57.8715	4.6941	65.2842	3.6406	64.9878
$X_3\,X_4$	54.2756	125.7806	51.6828	126.7200	6.1055	127.9752	3.9853	128.6459
$X_1\,X_2\,X_3$	4.3852	3.6169	4.3647	3.5583	4.0004	3.6296	3.8108	3.5170
$X_1\,X_2\,X_4$	3.6821	5.1481	3.8975	5.0931	5.0321	5.3201	4.3798	5.1078
$X_1\,X_3\,X_4$	43.5963	89.3993	39.5378	90.3715	7.5577	90.4398	4.8287	90.5849
$X_2\,X_3\,X_4$	16.6383	44.2775	14.8272	44.3727	3.4836	51.8037	3.824	51.2683
$X_1\,X_2\,X_3\,X_4$	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000

multiplying the actual value of Y by twenty (largest absolute residual corresponds to Y_{19} and its actual value and outlier value after multiplying by 20 is 4.73744 and 94.7488 respectively). The values of the C_p, S_p, R_p and GS_p statistics are computed for all possible subset models and are presented in Table 2.

It is clear that, the values of the C_p, S_p, R_p and GS_p statistics for subset $\{X_1, X_2\}$ are close to $p(=3)$ for clean data. This indicates that, all four methods agree upon the importance of two regressor variables X_1 and X_2 and select the correct subset model. For outlier case, it seems that the C_p and R_p statistics select the wrong subset or more than one subset models. These both criteria fail to select the correct subset model. As a contrast, we can see that the S_p and GS_p statistics select the same subset which is selected for clean data. Therefore, the S_p and GS_p statistics are immune to the presence of outlier in the data.

Example 6.2. In this example, the simulation design given in Eq. (2.1) is used to introduce multicollinearity in the regressor variables. The degree of multicollinearity is set to $\rho = 0.999$ and $n = 30$ observations are generated on the response variable using the following regression model

$$Y_i = 15 + 5x_{i1} + 5x_{i2} + 0x_{i3} + 0x_{i4} + \varepsilon_i,$$

where $\varepsilon_i \sim N(0,\ 1)$, $i = 1, 2, \ldots, 30$. The same scenario used in Example 6.1 is implemented to introduce an outlier observation in the response variable. The VIFs corresponding to each regressor variable are 405.0179, 441.0012, 373.2567 and 509.0688. The values of C_p, S_p, R_p and GS_p statistics for all possible subset models are computed for without and with one outlier case and are reported in the Table 3.

Table 3 shows that in presence of multicollinearity in the data, the R_p and GS_p statistics select the subset $\{X_1, X_2\}$ as both satisfy the criterion while the C_p and S_p statistics satisfy the criterion for several subsets of different sizes and so no definite conclusion can be drawn. In other words, they fail to select the correct subset.

In the presence of outlier as well as multicollinearity, the R_p statistic selects several subsets while the GS_p statistic selects the subset correctly namely $\{X_1, X_2\}$. Thus, this study shows that the performance of GS_p statistic is better than its competitors in the presence of data anomalies like outliers and multicollinearity.

Further, to assess the performance of the GS_p statistic in the presence of several outliers, we have considered above data set (used in Example 6.2). By introducing two and three outlier observations in the response variable, the correct model selection performance of the C_p, S_p, R_p and GS_p statistics is evaluated. The results are presented in Table 4.

Table 4 indicates similar conclusions as in case of one outlier, namely, the GS_p statistic selects the correct subset model $\{X_1, X_2\}$ while C_p and S_p statistics select several subsets satisfying the criterion

Table 3
Values of C_p, S_p, R_p and GS_p for all subset models in the presence of multicollinearity.

Regressors in the model	Without outlier				With one outlier			
	C_p	S_p	R_p	GS_p	C_p	S_p	R_p	GS_p
X_1	3.0624	4.0741	4.8225	4.0673	0.0868	4.0757	1.9696	4.0402
X_2	4.7698	7.1797	5.8701	7.3400	−0.0061	7.1948	1.9213	7.3243
X_3	13.7610	16.7849	13.5605	16.9714	0.2235	16.7863	2.0500	16.8117
X_4	16.9498	21.3865	12.9645	20.2555	0.0976	21.3939	1.9813	20.1864
$X_1\,X_2$	2.7505	3.5041	3.4738	3.2757	1.7151	3.5189	2.8865	3.2695
$X_1\,X_3$	3.0103	3.1471	3.7453	3.6414	1.7867	3.1461	2.8929	3.6164
$X_1\,X_4$	4.4605	5.2065	4.3467	4.8318	2.0868	5.2072	2.7099	4.7966
$X_2\,X_3$	3.9440	4.8946	4.1756	4.7671	1.1404	4.8960	2.8873	4.7696
$X_2\,X_4$	6.2422	8.2382	5.5534	7.5530	1.7710	8.2538	2.8762	7.5054
$X_3\,X_4$	14.0713	16.7995	11.4287	15.7780	1.5284	16.7907	2.9295	15.7992
$X_1\,X_2\,X_3$	3.7495	3.8968	3.4489	3.7146	3.1025	3.8993	3.9311	3.7153
$X_1\,X_2\,X_4$	4.7441	5.4564	4.2877	4.9393	3.6141	5.4687	3.9191	4.9174
$X_1\,X_3\,X_4$	4.8445	4.9141	4.9912	5.5923	3.4331	4.9156	3.9596	5.5801
$X_2\,X_3\,X_4$	5.3831	6.2449	5.5784	6.6812	3.0479	6.2465	3.9249	6.6903
$X_1\,X_2\,X_3\,X_4$	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000

Table 4
Values of C_p, S_p, R_p and GS_p for all subset models in the presence of multicollinearity and more than one outlier.

Regressors in the model	With two outliers				With three outliers			
	C_p	S_p	R_p	GS_p	C_p	S_p	R_p	GS_p
X_1	0.2189	3.1665	2.0991	3.6478	−0.8819	3.1855	2.2393	3.6339
X_2	0.1246	5.7824	2.0488	6.8607	−0.8933	5.7982	2.2342	6.8489
X_3	0.4238	17.1435	2.2255	18.4844	−0.7931	17.1703	2.2813	18.8167
X_4	0.2749	20.3236	2.1399	20.8400	−0.7848	20.3956	2.2828	20.4596
$X_1\,X_2$	1.9256	2.7502	2.8969	2.7616	1.1066	2.7400	2.9518	2.7213
$X_1\,X_3$	1.7165	3.1903	2.9178	3.7398	1.0853	3.1907	2.9094	3.7252
$X_1\,X_4$	2.1909	4.6171	2.8028	4.5929	1.0607	4.6175	2.9481	4.5601
$X_2\,X_3$	1.0655	4.7947	2.9047	4.9666	1.0642	4.8012	2.9204	4.9436
$X_2\,X_4$	1.7747	7.2267	2.9031	7.2986	1.0137	7.2338	2.9766	7.3958
$X_3\,X_4$	1.6978	16.9648	2.9623	17.1498	1.2064	16.9686	3.0786	17.0818
$X_1\,X_2\,X_3$	3.0614	3.8247	3.9376	3.7411	3.0550	3.8260	3.9361	3.7317
$X_1\,X_2\,X_4$	3.7025	4.7580	3.9313	4.5242	3.0000	4.7454	3.9879	4.4937
$X_1\,X_3\,X_4$	3.4453	5.0073	3.9738	5.6390	3.0607	5.0082	3.9730	5.7065
$X_2\,X_3\,X_4$	3.0068	6.2014	3.9417	6.7953	3.0130	6.2008	3.9974	6.7293
$X_1\,X_2\,X_3\,X_4$	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000

and thereby leading to inconclusive decision. The values of the R_p statistic corresponding to more than one subset models are close to p and it selects correct as well as wrong subset models. Hence, the performance of the GS_p statistic is again better than the rest in the presence of multicollinearity and more than one outlier observations in the data.

The graphical representation brings out the above facts clearly. For this purpose, the data set exhibiting multicollinearity and having two outliers' given in Table 4 is considered. The values of S_p, R_p and GS_p statistics corresponding to subset models are plotted in Fig. 2. Also, from Tables 3 and 4, the values of GS_p statistic for subset models corresponding to one outlier case, two outliers case and three outliers' case are obtained and plotted against p in Fig. 3. These figures clearly indicate that the GS_p statistic consistently select the correct subset model for the problem of multicollinearity with different number of outliers in the response variable.

6.2. Model selection ability

In this subsection, the correct model selection ability of the C_p, S_p, R_p and GS_p statistics is studied. A simulation design given by McDonald and Galarneau [15] in Eq. (2.1) is used to induce

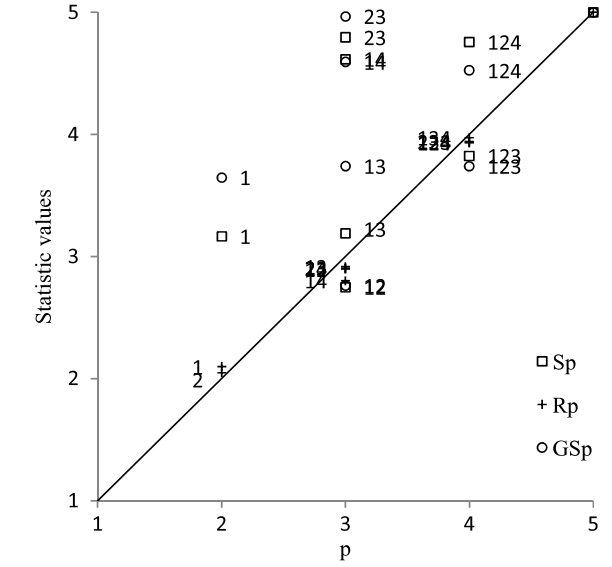


Fig. 2. Values of S_p, R_p and GS_p statistics versus p.

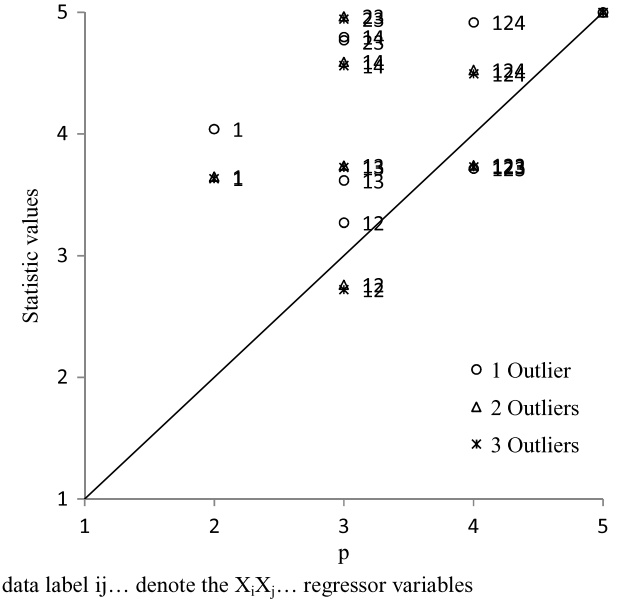


Fig. 3. Values of GS_p statistic versus p.

multicollinearity in X_1 and X_2 regressor variables of the following regression models.

$$\text{Model I} \quad Y_i = 5 + 4X_{i1} + 0X_{i2} + 3X_{i3} + 0X_{i4} + \varepsilon_i, \quad i = 1, 2, \ldots, 30$$

$$\text{Model II} \quad Y_i = 5 + 2X_{i1} + 0X_{i2} + 4X_{i3} + 0X_{i4} + 3X_{i5} + \varepsilon_i, \quad i = 1, 2, \ldots, 50.$$

Table 5
Model selection ability of C_p, S_p, R_p and GS_p statistics.

	ρ	Without outlier				With one outlier			
		0.6	0.7	0.8	0.9	0.6	0.7	0.8	0.9
Model I	$\sigma^2 = 0.25$								
	C_p	72	81	74	71	23	24	20	18
	S_p	63	66	63	62	63	68	62	64
	R_p	72	79	72	71	25	23	20	21
	GS_p	63	66	64	63	63	68	64	65
	$\sigma^2 = 1$								
	C_p	70	74	74	71	17	19	19	18
	S_p	61	63	64	62	65	60	63	59
	R_p	69	74	73	70	18	21	22	17
	GS_p	62	63	64	63	65	62	64	61
Model II	$\sigma^2 = 0.25$								
	C_p	81	74	76	79	16	24	21	20
	S_p	75	67	69	72	59	65	70	63
	R_p	81	74	75	77	17	29	21	20
	GS_p	76	68	69	73	62	66	72	66
	$\sigma^2 = 1$								
	C_p	63	74	74	79	22	21	19	19
	S_p	62	67	62	72	57	63	63	63
	R_p	61	74	74	79	24	25	18	23
	GS_p	63	69	62	74	62	65	64	67

The remaining regressor variables in Model I (X_3 and X_4) and Model II (X_3 and X_5) are generated from standard normal distribution. To make Model II more complicated, regressor variable X_4 is taken as a product of X_2 and X_3 regressor variable. The error variable ε_i, $i = 1, 2, \ldots, n$ is generated from normal distribution with mean 0 and variance $\sigma^2 = 0.25$ and $\sigma^2 = 1$. The different degrees of multicollinearity are achieved by setting $\rho = 0.6, 0.7, 0.8$ and 0.9. The scenario used in Example 6.1 of Section 6 is followed to introduce outlier observation in the response variable.

The simulation experiment is replicated 100 times for each model, for all combinations of degree of multicollinearity (ρ), error variance (σ^2) and with and without outlier case. The values of the C_p, S_p, R_p and GS_p statistics are computed for all possible subset models. The number of times that the C_p, S_p, R_p and GS_p statistics selects the correct subset model is counted and reported in the Table 5.

Table 5 shows that, for without outlier case, the frequency of correct model selection of the C_p and R_p statistics is larger than that of the S_p and GS_p statistics. But, for single outlier case, the correct model selection ability of the GS_p statistic is uniformly larger than that of the C_p and R_p statistics for both model I and model II. For one outlier case, the percentage of correct subset model selection by the GS_p statistic is larger than that of the S_p statistic when the value of ρ is large.

6.3. Choice of the estimator of σ^2

In the computation of the GS_p statistic, we use a scale parameter σ^2. Since it is unknown, we need to use its suitable estimator. Various estimators of σ^2 are available in the literature. For examining the performance of the GS_p criterion, we use four different types of estimators of σ^2 which are based on the JRM estimator. Let r_i be the ith residual based on the JRM estimator of β and it is defined as,

$$r_i = Y_i - X_i' \hat{\beta}_{JRM}, \quad i = 1, 2, \ldots, n.$$

We consider the following estimator of σ^2.

1. $\hat{\sigma}_1^2 = [1.4826\,\text{Median}\ |r_i - \text{Median}\,(r_i)|]^2$
2. $\hat{\sigma}_2^2 = [1.4826\,\text{Median}\ |r_i|]^2$
3. $\hat{\sigma}_3^2 = \sum_{i=1}^{n} r_i^2 w_i^2 / (n-k)$
4. $\hat{\sigma}_4^2 = \sum_{i=1}^{n} r_i^2 w_i^2 / \left(\sum_{i=1}^{n} |r_i w_i|\right)$.

Table 6
Values of GS_p statistic for all subset models with different estimates of σ^2.

Regressors in the model	With one outlier				With two outliers			
	$\hat{\sigma}_1^2$ (0.9881)[a]	$\hat{\sigma}_2^2$ (0.8478)	$\hat{\sigma}_3^2$ (0.8112)	$\hat{\sigma}_4^2$ (0.9949)	$\hat{\sigma}_1^2$ (0.9881)	$\hat{\sigma}_2^2$ (0.8478)	$\hat{\sigma}_3^2$ (0.8050)	$\hat{\sigma}_4^2$ (0.9901)
X_1	337.5378	393.5145	411.3049	335.2256	336.6605	392.5125	413.4319	336.0047
X_2	102.7412	119.8496	125.2869	102.0346	103.7496	121.0292	127.5012	103.5467
X_3	615.8711	717.9162	750.3480	611.6561	615.6479	717.6772	755.8924	614.4499
X_4	124.4055	145.1017	151.6793	123.5507	125.0403	145.8491	153.6430	124.7960
$X_1\,X_2$	2.9559	3.2151	3.2974	2.9452	2.9414	3.2006	3.2977	2.9384
$X_1\,X_3$	314.6453	366.4917	382.9694	312.5038	313.0681	364.6737	384.0026	312.4621
$X_1\,X_4$	12.9300	14.8421	15.4498	12.8510	12.9644	14.8887	15.6094	12.9418
$X_2\,X_3$	43.3745	50.3237	52.5323	43.0875	43.6117	50.6038	53.2227	43.5296
$X_2\,X_4$	103.6550	120.5901	125.9723	102.9555	104.5139	121.5935	127.9907	104.3133
$X_3\,X_4$	11.7165	13.4261	13.9695	11.6459	11.6709	13.3807	14.0210	11.6508
$X_1\,X_2\,X_3$	3.5522	3.5621	3.5653	3.5518	3.5441	3.5539	3.5576	3.5440
$X_1\,X_2\,X_4$	3.8178	3.8903	3.9134	3.8148	3.8174	3.8897	3.9168	3.8165
$X_1\,X_3\,X_4$	6.9981	7.5776	7.7618	6.9741	7.0097	7.5940	7.8129	7.0029
$X_2\,X_3\,X_4$	11.7398	13.1290	13.5705	11.6825	11.7104	13.0941	13.6124	11.6941
$X_1\,X_2\,X_3\,X_4$	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000

[a] Figures in parenthesis indicate the estimate of $\sigma_i^2,\ i = 1, 2, 3, 4$.

The performance of these estimators is illustrated by using a simulated example. Here, a random sample of size $n = 30$ is generated from $N_4(0,\ \Sigma)$, on X_1, X_2, X_3 and X_4, where,

$$\sum = \begin{bmatrix} 1 & 0.531 & -0.850 & -0.531 \\ 0.531 & 1 & -0.401 & -0.972 \\ -0.850 & -0.401 & 1 & 0.288 \\ -0.531 & -0.972 & 0.288 & 1 \end{bmatrix}.$$

A regression model considered to generate $n = 30$ observations on the response variable is given by

$$Y_i = 5 + 2X_{i1} + 4X_{i2} + 0X_{i3} + 0X_{i4} + \varepsilon_i, \quad i = 1,\ 2, \ldots, 30,$$

where $\varepsilon_i \sim N(0,\ 1)$. The VIFs of each term are 31.0458, 182.9686, 33.1639 and 210.6130. One and two outlier observations are introduced in the response variable using the same procedure given in Example 6.1. Based on the simulated data, the values of the GS_p statistic for one outlier and two outlier observations case with four different estimators of σ^2 are obtained and presented in Table 6.

From Table 6, it is clear that the GS_p statistic select the same subset $\{X_1, X_2\}$ in the presence of multicollinearity with one outlier and two outliers case for all estimators of σ^2. The values of $\hat{\sigma}_1^2$ and $\hat{\sigma}_4^2$ are close to true value of σ^2 as compare to the values of $\hat{\sigma}_2^2$ and $\hat{\sigma}_3^2$. Consequently, the value of GS_p statistic corresponding to the correct subset model using $\hat{\sigma}_1^2$ and $\hat{\sigma}_4^2$ are close to p as compare to that of the $\hat{\sigma}_2^2$ and $\hat{\sigma}_3^2$.

7. Conclusion

We have developed a subset selection procedure based on the JRM estimator of the unknown regression parameters. This method works well in subset selection for clean data or in presence of only outlier or only multicollinearity or both outlier and multicollinearity. Also, the performance of the proposed method is evaluated for the presence of more than one outlier observations and multicollinearity in the data. The correct model selection ability of the proposed method is also obtained. It reveals that, the performance of the proposed method is considerably better as compare to other existing methods when the outlier observations and multicollinearity occur simultaneously in the data.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments and constructive suggestions which substantially improve the quality of the manuscript. This research was supported by the University Grants Commission, New Delhi, India under Major Research Project Scheme.

References

[1] D. Brikes, Y. Dodge, Alternative Methods of Regression, John Wiley and Sons, New York, 1993.
[2] S. Chatterji, A.S. Hadi, Sensitive Analysis in Linear Regression, John Wiley and Sons, New York, 1988.
[3] A.V. Dorugade, D.N. Kashid, Variable selection in linear regression based on ridge estimator, J. Stat. Comput. Simul. 80 (11) (2010) 1211–1224.
[4] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley and Sons, New York, 2003.
[5] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55–67.
[6] A.E. Hoerl, R.W. Kennard, Ridge regression: applications to nonorthogonal problems, Technometrics 12 (1970) 69–82.
[7] A.E. Hoerl, R.W. Kennard, K.F. Baldwin, Ridge regression: some simulations, Commun. Stat. 4 (1975) 105–123.
[8] P.J. Huber, Robust Regression: asymptotics, conjectures, and Monte Carlo, Ann. Statist. 1 (5) (1973) 799–821.
[9] P.J. Huber, Robust Statistics, John Wiley and Sons, New York, 1981.
[10] N.H. Jadhav, D.N. Kashid, A jackknifed ridge M-estimator for regression model with multicollinearity and outliers, J. Stat. Theory Pract. 5 (4) (2011) 659–673.
[11] D.N. Kashid, S.R. Kulkarni, More general criterion for subset selection in multiple linear regressions, Comm. Statist. Theory Methods 31 (5) (2002) 795–811.
[12] C. Kim, S. Hwang, Influential subsets on the variable selection, Comm. Statist. Theory Methods 29 (2) (2000) 335–347.
[13] C.L. Mallows, Some comments on C_p, Technometrics 15 (1973) 661–675.
[14] D.W. Marquardt, Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation, Technometrics 12 (1970) 591–612.
[15] G.C. McDonald, D.I. Galarneau, A Monte Carlo evaluation of some ridge-type estimators, J. Amer. Statist. Assoc. 70 (350) (1975) 407–416.
[16] D.C. Montgomery, E.A. Peck, G.G. Vining, Introduction to Linear Regression Analysis, John Wiley and Sons, New York, 2006.
[17] E. Ronchetti, Robust model selection in regression, Statist. Probab. Lett. 3 (1985) 21–23.
[18] E.M. Ronchetti, R.G. Statudte, A robust version of Mallow's C_p, J. Amer. Statist. Assoc. 89 (426) (1994) 550–559.
[19] S. Sommer, R.M. Huggins, Variables selection using the Wald test and a robust C_p, J. R. Stat. Soc. Ser. C. Appl. Stat. 45 (1) (1996) 15–29.

Communications in Statistics—Theory and Methods, 43: 3135–3147, 2014

ISSN: 0361-0926 print / 1532-415X online
DOI: 10.1080/03610926.2012.694548

Taylor & Francis
Taylor & Francis Group

The Steady-State Performance of Cumulative Count of a Conforming Control Chart Based On Runs Rules

S. K. KHILARE AND D. T. SHIRKE

Department of Statistics, Shivaji University, Kolhapur, India

Cumulative count of conforming control chart is usually used to monitor fraction non-conforming in high-yield processes. In this article, we propose m-of-m control chart based on cumulative count of conforming units for high-yield processes. The steady-state properties of the m-of-m control chart are investigated. We compare performance of the m-of-m control chart with control chart based on cumulative count of conforming units. We present Markov chain model of the m-of-m control chart to evaluate average run length, standard deviation of run length and quartiles.

Keywords High-yield processes; Average run length; Standard deviation of run length; Markov chain; Steady-state and percentiles.

Mathematics Subject Classification Primary 60G35; Secondary 93E10.

1. Introduction

Fraction non conforming of modern manufacturing processes is usually very low at parts per million. Such a process is known as high-yield or high-quality process. The Shewhart *p* chart is not suitable for monitoring the high-yield processes, since for a large subgroup size the number of non conforming units in a subgroup is assumed to be approximately normal. When the fraction nonconforming of units is very small, normal approximation becomes incorrect. Goh (1987) showed that the use of Shewhart *p* chart in high yield processes results in high false alarm rates and consequently it fails to detect process improvement. In order to provide adequate statistical process control technique for high-yield processes, Goh (1987) proposed the cumulative count of conforming (CCC) control chart based on geometric distribution as an alternative to the *p*chart. The concept of cumulative count of conforming items is first introduced by Calvin (1983) and its advantages explored by many researchers; see Xie and Goh (1992, 1993, 2003), Kaminsky et al. (1992), Nelson (1994), Xie et al. (1995, 1999), Kuralmani et al. (2002), Steiner et al. (2004), and Chen et al. (2010). The approach of setting of control limits of the CCC control chart and the CCC-*r* chart are investigated by Xie et al. (2000) and Chen and Cheng (2010). The concept

Received December 16, 2011; Accepted May 7, 2012.
Address correspondence to S. K. Khilare, Department of Statistics, Shivaji University, Kolhapur, India 416004; E-mail: shashi.khilare@gmail.com
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.

Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v7n2p228

Reliability estimation of k-unit series system based on progressively censored data
By Potdar K.G., Shirke D.T.

Published: 14 October 2014

Electronic Journal of Applied Statistical Analysis
Vol. 07, Issue 02, 2014, 228-253
DOI: 10.1285/i20705948v7n2p228

Reliability estimation of k-unit series system based on progressively censored data

K. G. Potdar*[a] and D. T. Shirke[b]

[a] *Department of Statistics, Ajara Mahavidyalaya, Ajara, Dist-Kolhapur, Maharashtra, India - 416505.*
[b] *Department of Statistics, Shivaji University, Kolhapur, Dist-Kolhapur, Maharashtra, India - 416004.*

Published: 14 October 2014

In this article, we consider a k-unit series system with component lifetime distribution to be a member of the scale family of distributions. We discuss estimation of the scale parameter and estimation of reliability function of the family based on progressively Type-II censored sample. The maximum likelihood estimator (MLE) of the scale parameter is derived using Expectation-Maximization (EM) algorithm and is used to estimate reliability function. Confidence intervals are constructed using asymptotic distribution of MLE. β -expectation tolerance interval for lifetime of the system is obtained. We consider half-logistic distribution as a member of the scale family and study performance of the MLE, reliability estimate and confidence interval using simulation experiments. Illustration through real data example is provided.

keywords: Progressively Type-II censoring, EM algorithm, MLE, confidence interval, coverage probability, reliability, β-expectation tolerance interval, half-logistic distribution.

1 Introduction

In industrial phenomenon series systems are widely used. Electric, automobile as well as in chemical industry various units are connected in series. Here system is working if all

*Corresponding author: potdarkiran.stat@gmail.com.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

units in system are working. If any one unit is failed then system fails. Thus, system life is smaller than unit life. Life testing under series system is more costly, because failure of one unit reflects in system failure. Therefore, we use censoring criteria, in that; we remove some working systems without observing its failure time. The unobserved failure time data are called censored data.

Broadly censoring is classified into two types; Type-I and Type-II censoring. Type-I censoring depends on time. In this type, an experiment continues up to a pre-determined time T. Units having failure time after time T are not observed. Here, failure time will be known exactly only if it is less than T. For example, if n units are placed on test, but decision is made to terminate the test at time T, then failure times will be known exactly only for those units that fail before time T. In Type-I censoring, the number of exact failure times observed is random.

Type-II censoring scheme is often used in life testing experiment. In this scheme only m units in a random sample of size $n(m < n)$ are observed. Progressive Type-II censoring is a generalization of Type-II censoring. In progressive censoring scheme, the number m and R_1, R_2,...., R_m are fixed prior to the test and $\sum_{i=1}^{m} R_i = n - m$. At the first failure R_1 units are randomly removed from remaining $n - 1$ units. At the second failure, R_2 units are randomly removed from remaining $n - 2 - R_1$ units and so on. At the m^{th} failure all remaining R_m units are removed. Here, we observe failure time of m units and remaining $n - m$ units are removed at different stages of experiment. In conventional Type-II censoring scheme $R_1 = R_2 = ... = R_{m-1} = 0$ and $R_m = n - m$. In this article, the progressive Type-II censoring scheme is considered.

Many authors studied progressive Type-II censoring scheme for various lifetime distributions. Cohen (1963) introduced progressive Type-II censoring. Mann (1969) and Mann (1971) considered Weibull distribution with progressive censoring. Balakrishnan and Asgharzadeh (2005), Balakrishnan et al. (2003) and Balakrishnan et al. (2004) discussed inference for half-logistic, Gaussian and extreme value distribution under progressive Type-II censoring scheme respectively. Ng (2005) studied parameter estimation for modified Weibull distribution under progressive Type-II censoring. Balakrishnan and Aggarwala (2000) gave details about progressive censoring. Balakrishnan (2007) studied various distributions and inferential methods for progressively censored data. Pradhan (2007) considered point and interval estimation of a k-unit parallel system based on progressive Type-II censoring scheme with exponential distribution as the component life distribution.

Kim and Han (2010) discussed half-logistic distribution for Type-II progressively censored samples. Iliopoulos and Balakrishnan (2011) studied likelihood inference for Laplace distribution based on progressively Type-II censored samples. Asgharzadeh and Valiollahi (2011) considered estimation of the scale parameter of the Lomax distribution under progressive censoring scheme. Krishna and Malik (2012), Krishna and Kumar (2011) and Krishna and Kumar (2013) studied reliability estimation in Maxwell, Lindley and generalized inverted exponential distribution with progressively Type-II censored data. Recently, Potdar and Shirke (2014) discussed inference for the scale parameter of lifetime distribution of k-unit parallel system based on progressively Type-II censored data. Potdar and Shirke (2012) studied inference for the distribution of a k-unit par-

allel system with exponential distribution as the component life distribution based on Type-II progressively censored sample. Potdar and Shirke (2013a) discussed inference for the parameters of generalized inverted family of distributions. Potdar and Shirke (2013b) studied reliability estimation for the distribution of a k-unit parallel system when Rayleigh distribution as component lifetime distribution.

Dempster et al. (1977) introduced expectation-maximization (EM) algorithm. They presented maximum likelihood estimation for incomplete data. McLachlan and Krishnan (2007) gave more details about EM algorithm. Little and Rubin (2002) have discussed EM algorithm for exponential family of distributions. Pradhan and Kundu (2009) used EM algorithm to estimate parameters of generalized exponential distribution under progressive Type-II censoring scheme. Ng et al. (2002) used EM algorithm to estimate parameters of lognormal and Weibull distributions under Type-II censoring scheme. In this article, we use EM algorithm for estimation of the parameters of a k-unit series system based on progressive Type-II censoring scheme when unit lifetime distribution belongs to the scale family. Parameter estimation is based on the lifetimes of the system. We assume that n units are put on test and failure times of $\sum_{i=1}^{m} R_i = n - m.$ units are censored. Failure times of these censored units are unknown. We consider this data as missing and use EM algorithm to compute MLE. We use idea of missing information principle of Louis (1982). Asymptotic normal distribution of MLE is used to construct confidence interval for the scale parameter. We also discuss tolerance interval for the lifetime of the system, on the lines of Kumbhar and Shirke (2004).

The present work is different than the work reported by Pradhan (2007) in many aspects. The first thing is that we consider scale family of distributions and exponential distribution considered by Pradhan (2007) is a member of the family. Further, we obtain MLE using EM algorithm instead of using Newton-Raphson method. We use Newton-Raphson method within EM algorithm. Pradhan (2007) has considered only parameter estimation, while we consider inference of parameter as well as reliability function. We use missing information principle to compute Fisher information. We illustrate use of the results developed with half-logistic distribution, which is a member of scale family. Number of schemes that we consider are 30, which include schemes with small sample sizes.

In Section 2, we introduce the model and obtain MLE for the scale parameter and reliability function. We also provide an expression for Fisher information. Asymptotic confidence interval for the scale parameter based on MLE, log-MLE and confidence interval for the reliability function is discussed in the same section. Section 3 provides β-expectation tolerance interval for the lifetime of a k-unit series system based on progressively censored data. In Section 4, we consider the half-logistic distribution as a member of the scale family and discuss MLE, reliability function, confidence intervals and tolerance intervals. Performance of the MLE and confidence intervals of scale parameter and reliability function of half-logistic distribution is investigated using simulations. Results of simulation study have been reported in Section 5. Real data application is discussed in Section 6. Conclusions are presented in Section 7.

2 Model and Estimation of the Scale Parameter

Let $\mathbb{G}_\lambda$ be a scale family of lifetime distributions where λ is the parameter of the interest. Consider a k-unit series system with independent and identically distributed units having lifetimes $X_1, X_2, ..., X_k$ of k units. That is, X_i is the lifetime of the i^{th} unit having cumulative distribution function (cdf) $G\left(\frac{x_i}{\lambda}\right)$. The lifetime of the system is $X = Min.(X_1, X_2, ..., X_k)$. The cdf of X is

$$F(x;\lambda) = 1 - \left[1 - G\left(\frac{x}{\lambda}\right)\right]^k \qquad x \geq 0,\ \lambda > 0.$$

The probability density function (pdf) of X is

$$f(x;\lambda) = \frac{k}{\lambda} g\left(\frac{x}{\lambda}\right)\left[1 - G\left(\frac{x}{\lambda}\right)\right]^{k-1} \qquad x \geq 0,\ \lambda > 0.$$

where $g(.)$ is the pdf of X_i when $\lambda = 1$.

2.1 Maximum Likelihood Estimation

Suppose n k-unit series systems are under test and we observe failure times of m systems under progressive type-II censoring. Let $(R_1, R_2, ..., R_m)$ be a progressive censoring scheme.

The likelihood function for the observed data is

$$L(\lambda|\underline{x}) = C\prod_{i=1}^{m} f(x_{(i)};\lambda)\left[1 - F(x_{(i)};\lambda)\right]^{R_i},$$

$$\text{where } C = n\prod_{j=1}^{m-1}\left(n - j - \sum_{i=1}^{j} R_i\right).$$

$$L(\lambda|\underline{x}) = C\prod_{i=1}^{m}\frac{k}{\lambda} g\left(\frac{x_{(i)}}{\lambda}\right)\left[1 - G\left(\frac{x_{(i)}}{\lambda}\right)\right]^{k-1}\left[1 - G\left(\frac{x_{(i)}}{\lambda}\right)\right]^{kR_i},$$

Suppose $x_1, x_2, ..., x_m$ is the observed data and $z_1, z_2, ..., z_m$ is the censored data. We note that z_i is a vector with R_i elements, which is not observable for $i = 1, 2, ..., m$. The censored data $Z = (z_1, z_2, ..., z_m)$ can be considered as missing data.
$X = (x_1, x_2, ..., x_m)$ is observed data. $W = (X, Z)$ is the complete data set. Then complete log-likelihood function is

$$L_c = nlog(k) - nlog(\lambda) + \sum_{i=1}^{m} log\left[g\left(\frac{x_i}{\lambda}\right)\right] + (k-1)\sum_{i=1}^{m} log\left[1 - G\left(\frac{x_i}{\lambda}\right)\right]$$

$$+ \sum_{i=1}^{m}\sum_{j=1}^{R_i} log\left[g\left(\frac{z_{ij}}{\lambda}\right)\right] + (k-1)\sum_{i=1}^{m}\sum_{j=1}^{R_i} log\left[1 - G\left(\frac{z_{ij}}{\lambda}\right)\right]. \qquad (1)$$

In order to obtain MLE of λ, we use EM algorithm due to Dempster et al. (1977). For the E step in EM algorithm we take Expectation of Z_{ij}. The derivative of L_c with respect to λ is taken for the M step, where

$$\frac{dL_c}{d\lambda} = -\frac{n}{\lambda} - \frac{1}{\lambda^2}\sum_{i=1}^{m}\frac{x_i g'\left(\frac{x_i}{\lambda}\right)}{g\left(\frac{x_i}{\lambda}\right)} + \frac{(k-1)}{\lambda^2}\sum_{i=1}^{m}\frac{x_i G'\left(\frac{x_i}{\lambda}\right)}{1-G\left(\frac{x_i}{\lambda}\right)}$$

$$-\frac{1}{\lambda^2}\sum_{i=1}^{m} R_i a(x_i,k,\lambda^0) + \frac{(k-1)}{\lambda^2}\sum_{i=1}^{m} R_i b(x_i,k,\lambda^0). \tag{2}$$

$$\text{where } a(x_i,k,\lambda) = E\left[\frac{Z_{ij}g'\left(\frac{Z_{ij}}{\lambda}\right)}{g\left(\frac{Z_{ij}}{\lambda}\right)}\,\middle|\, Z_{ij} > x_i\right] = \int_{x_i}^{\infty}\frac{z g'\left(\frac{z}{\lambda}\right)}{g\left(\frac{z}{\lambda}\right)}\,\frac{f(z;\lambda)}{1-F(x_i;\lambda)}dz,$$

$$\text{and } b(x_i,k,\lambda) = E\left[\frac{Z_{ij}G'\left(\frac{Z_{ij}}{\lambda}\right)}{1-G\left(\frac{Z_{ij}}{\lambda}\right)}\,\middle|\, Z_{ij} > x_i\right] = \int_{x_i}^{\infty}\frac{z G'\left(\frac{z}{\lambda}\right)}{1-G\left(\frac{z}{\lambda}\right)}\,\frac{f(z;\lambda)}{1-F(x_i;\lambda)}dz.$$

We have to solve equation $\frac{dL_c}{d\lambda} = 0$ to obtain λ^1 as the solution. But this equation does not have solution in the closed form. Therefore we use Newton-Raphson method and compute λ^1. By using λ^1, we compute $a(x_i,k,\lambda^1)$ and $b(x_i,k,\lambda^1)$. This ends M-step. We continue this procedure until convergence takes place.

In Newton-Raphson method, we have to choose initial value of λ. We use least square estimate. Ng (2005) discussed estimation of model parameters of modified Weibull distribution based on progressively Type-II censored data where the empirical distribution function is computed as (see Meeker and Escobar (1998))

$$\hat{F}(x_i) = 1 - \prod_{j=1}^{i}(1-\hat{p}_j),$$

with

$$\hat{p}_j = \frac{1}{n - \sum_{k=2}^{j} R_{k-1} - j + 1} \qquad \text{for } j = 1,2,.....,m.$$

The estimate of the parameter can be obtained by using least square fit of simple linear regression.

$$y_i = \beta x_i \quad \text{with} \quad \beta = \frac{1}{\lambda}$$

$$y_i = G^{-1}\left[1 - \frac{\left[1-\hat{F}(x_{i-1})\right]^{1/k} + \left[1-\hat{F}(x_i)\right]^{1/k}}{2}\right] \text{ for } i = 1,2,.....,m.$$

$$\hat{F}(x_0) = 0,$$

The least square estimates of λ is given by

$$\hat{\lambda}_0 = \frac{\sum_{i=1}^{m} x_i^2}{\sum_{i=1}^{m} x_i \; y_i},$$

We use $\hat{\lambda}_0$ as an initial value of λ to obtain the MLE $\hat{\lambda}_n$ using Newton-Raphson method. Reliability function at time t is

$$R(t) = \left[1 - G\left(\frac{t}{\lambda}\right)\right]^k \qquad t \geq 0, \; \lambda > 0.$$

The Maximum likelihood estimator of $R(t)$ is

$$\hat{R}_n(t) = \left[1 - G\left(\frac{t}{\hat{\lambda}_n}\right)\right]^k \qquad t \geq 0.$$

2.2 Fisher Information

We compute observed Fisher information using the idea of missing information principle of Louis (1982).

Thus, observed information = complete information - missing information.

$$I_x(\lambda) = I_w(\lambda) - I_{w|x}(\lambda),$$

where the complete information $= I_w(\lambda) = -E\left[\frac{d^2L}{d\lambda^2}\right]$ and L is the log-likelihood function based on all n observations. We obtain $I_w(\lambda)$ and $I_{w|x}(\lambda)$ in the following.

Now,

$$L = nlog(k) - nlog(\lambda) + \sum_{i=1}^{n} log\left[g\left(\frac{x_i}{\lambda}\right)\right] + (k-1)\sum_{i=1}^{n} log\left[1 - G\left(\frac{x_i}{\lambda}\right)\right]. \qquad (3)$$

and

$$\frac{dL}{d\lambda} = -\frac{n}{\lambda} - \frac{1}{\lambda^2}\sum_{i=1}^{n} \frac{x_i g'\left(\frac{x_i}{\lambda}\right)}{g\left(\frac{x_i}{\lambda}\right)} + \frac{(k-1)}{\lambda^2}\sum_{i=1}^{n} \frac{x_i G'\left(\frac{x_i}{\lambda}\right)}{1 - G\left(\frac{x_i}{\lambda}\right)}.$$

$$\frac{d^2L}{d\lambda^2} = \frac{n}{\lambda^2} + \frac{1}{\lambda^4}\sum_{i=1}^{n} \frac{x_i^2 g\left(\frac{x_i}{\lambda}\right) g''\left(\frac{x_i}{\lambda}\right) - x_i^2\left[g'\left(\frac{x_i}{\lambda}\right)\right]^2 + 2\lambda x_i g\left(\frac{x_i}{\lambda}\right) g'\left(\frac{x_i}{\lambda}\right)}{\left[g\left(\frac{x_i}{\lambda}\right)\right]^2}$$

$$-\frac{(k-1)}{\lambda^4}\sum_{i=1}^{n} \frac{x_i^2\left[1 - G\left(\frac{x_i}{\lambda}\right)\right] G''\left(\frac{x_i}{\lambda}\right) + x_i^2\left[G'\left(\frac{x_i}{\lambda}\right)\right]^2 + 2\lambda x_i\left[1 - G\left(\frac{x_i}{\lambda}\right)\right] G'\left(\frac{x_i}{\lambda}\right)}{\left[1 - G\left(\frac{x_i}{\lambda}\right)\right]^2}.$$

Complete information is given by

$$I_w(\lambda) = -\frac{n}{\lambda^2} - \frac{1}{\lambda^4}\sum_{i=1}^{n} E\left[\frac{X_i^2 g\left(\frac{X_i}{\lambda}\right) g''\left(\frac{X_i}{\lambda}\right) - X_i^2\left[g'\left(\frac{X_i}{\lambda}\right)\right]^2 + 2\lambda X_i g\left(\frac{X_i}{\lambda}\right) g'\left(\frac{X_i}{\lambda}\right)}{\left[g\left(\frac{X_i}{\lambda}\right)\right]^2}\right]$$

$$+\frac{(k-1)}{\lambda^4}\sum_{i=1}^{n} E\left[\frac{X_i^2\left[1-G\left(\frac{X_i}{\lambda}\right)\right] G''\left(\frac{X_i}{\lambda}\right) + X_i^2\left[G'\left(\frac{X_i}{\lambda}\right)\right]^2 + 2\lambda X_i\left[1-G\left(\frac{X_i}{\lambda}\right)\right] G'\left(\frac{X_i}{\lambda}\right)}{\left[1-G\left(\frac{X_i}{\lambda}\right)\right]^2}\right]. \tag{4}$$

Missing information is given by

$$I_{w|x}(\lambda) = \sum_{i=1}^{m} R_i I_{w|x}^{(i)}(\lambda) = -\sum_{i=1}^{m}\sum_{j=1}^{R_i} E_{Z|X}\left[\frac{d^2 log\left(f(Z_{ij}|x_i,\lambda)\right)}{d\lambda^2}\right]$$

Consider

$$f_{Z|X}(z_{ij}|x_i,\lambda) = \frac{f(z_{ij};\lambda)}{1-F(x_i,\lambda)} = \frac{\frac{k}{\lambda} g\left(\frac{z_{ij}}{\lambda}\right)\left[1-G\left(\frac{z_{ij}}{\lambda}\right)\right]^{k-1}}{\left[1-G\left(\frac{x_i}{\lambda}\right)\right]^k}.$$

Therefore,

$$log(f) = log(k) - log(\lambda) + log\left[g\left(\frac{z_{ij}}{\lambda}\right)\right] + (k-1) log\left[1-G\left(\frac{z_{ij}}{\lambda}\right)\right] - k log\left[1-G\left(\frac{x_i}{\lambda}\right)\right].$$

$$\frac{dlogf}{d\lambda} = -\frac{1}{\lambda} - \frac{z_{ij} g'\left(\frac{z_{ij}}{\lambda}\right)}{\lambda^2 g\left(\frac{z_{ij}}{\lambda}\right)} + \frac{(k-1) z_{ij} G'\left(\frac{z_{ij}}{\lambda}\right)}{\lambda^2\left[1-G\left(\frac{z_{ij}}{\lambda}\right)\right]} - \frac{k x_i G'\left(\frac{x_i}{\lambda}\right)}{\lambda^2\left[1-G\left(\frac{x_i}{\lambda}\right)\right]}.$$

and

$$\frac{d^2 logf}{d\lambda^2} = \frac{1}{\lambda^2} + \frac{z_{ij}^2 g\left(\frac{z_{ij}}{\lambda}\right) g''\left(\frac{z_{ij}}{\lambda}\right) - z_{ij}^2\left[g'\left(\frac{z_{ij}}{\lambda}\right)\right]^2 + 2\lambda z_{ij} g\left(\frac{z_{ij}}{\lambda}\right) g'\left(\frac{z_{ij}}{\lambda}\right)}{\lambda^4\left[g\left(\frac{z_{ij}}{\lambda}\right)\right]^2}$$

$$-\frac{(k-1)\left\{z_{ij}^2\left[1-G\left(\frac{z_{ij}}{\lambda}\right)\right] G''\left(\frac{z_{ij}}{\lambda}\right) + z_{ij}^2\left[G'\left(\frac{z_{ij}}{\lambda}\right)\right]^2 + 2\lambda z_{ij}\left[1-G\left(\frac{z_{ij}}{\lambda}\right)\right] G'\left(\frac{z_{ij}}{\lambda}\right)\right\}}{\lambda^4\left[1-G\left(\frac{z_{ij}}{\lambda}\right)\right]^2}$$

$$+\frac{k\left\{x_i^2\left[1-G\left(\frac{x_i}{\lambda}\right)\right] G''\left(\frac{x_i}{\lambda}\right) + x_i^2\left[G'\left(\frac{x_i}{\lambda}\right)\right]^2 + 2\lambda x_i G'\left(\frac{x_i}{\lambda}\right)\left[1-G\left(\frac{x_i}{\lambda}\right)\right]\right\}}{\lambda^4\left[1-G\left(\frac{x_i}{\lambda}\right)\right]^2}.$$

Hence, missing information is

$$I_{w|x}(\lambda) = \sum_{i=1}^{m} R_i I_{w|x}^{(i)}(\lambda) = -\sum_{i=1}^{m}\sum_{j=1}^{R_i} E_{Z|X}\left[\frac{d^2 log\left(f(Z_{ij}|x_i,\lambda)\right)}{d\lambda^2}\right]$$

$$= -\frac{n-m}{\lambda^2}$$
$$-\frac{1}{\lambda^4}\sum_{i=1}^{m}\sum_{j=1}^{R_i} E\left[\frac{Z_{ij}^2 g\left(\frac{Z_{ij}}{\lambda}\right) g''\left(\frac{Z_{ij}}{\lambda}\right) - Z_{ij}^2\left[g'\left(\frac{Z_{ij}}{\lambda}\right)\right]^2 + 2\lambda Z_{ij} g\left(\frac{Z_{ij}}{\lambda}\right) g'\left(\frac{Z_{ij}}{\lambda}\right)}{\left[g\left(\frac{Z_{ij}}{\lambda}\right)\right]^2}\right]$$
$$-\frac{(k-1)}{\lambda^4}\sum_{i=1}^{m}\sum_{j=1}^{R_i} E\left[\frac{Z_{ij}^2\left[1-G\left(\frac{Z_{ij}}{\lambda}\right)\right] G''\left(\frac{Z_{ij}}{\lambda}\right) + Z_{ij}^2\left[G'\left(\frac{Z_{ij}}{\lambda}\right)\right]^2}{\left[1-G\left(\frac{Z_{ij}}{\lambda}\right)\right]^2}\right]$$
$$-\frac{2(k-1)}{\lambda^3}\sum_{i=1}^{m}\sum_{j=1}^{R_i} E\left[\frac{Z_{ij}\left[1-G\left(\frac{Z_{ij}}{\lambda}\right)\right] G'\left(\frac{Z_{ij}}{\lambda}\right)}{\left[1-G\left(\frac{Z_{ij}}{\lambda}\right)\right]^2}\right]$$
$$+\frac{k}{\lambda^4}\sum_{i=1}^{m} R_i\left[\frac{x_i^2\left[1-G\left(\frac{x_i}{\lambda}\right)\right] G''\left(\frac{x_i}{\lambda}\right) + x_i^2\left[G'\left(\frac{x_i}{\lambda}\right)\right]^2 + 2\lambda x_i G'\left(\frac{x_i}{\lambda}\right)\left[1-G\left(\frac{x_i}{\lambda}\right)\right]}{\left[1-G\left(\frac{x_i}{\lambda}\right)\right]^2}\right]. \quad (5)$$

Using expressions in equations (4) and (5) we obtain observed Fisher information.

2.3 Confidence Intervals

By using asymptotic normal distribution of MLE $\hat{\lambda}_n$, we construct confidence interval for λ. Let $\hat{\sigma}^2(\hat{\lambda_n}) = \frac{1}{I(\hat{\lambda_n})}$ is the estimated variance of $\hat{\lambda}_n$. Therefore, $100(1-\alpha)\%$ asymptotic confidence interval for λ is given by

$$\left(\hat{\lambda}_n - \tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\hat{\lambda}_n)}, \quad \hat{\lambda}_n + \tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\hat{\lambda}_n)}\right), \quad (6)$$

where $\tau_{\alpha/2}$ is the upper $100(\alpha/2)^{th}$ percentile of standard normal distribution.

Meeker and Escobar (1998) reported that the asymptotic confidence interval for λ can be computed using $log(\hat{\lambda}_n)$. An approximate $100(1-\alpha)\%$ confidence interval for $log(\lambda)$ is given by

$$\left(log(\hat{\lambda}_n) - \tau_{\alpha/2}\sqrt{\hat{\sigma}^2(log(\hat{\lambda}_n))}, \quad log(\hat{\lambda}_n) + \tau_{\alpha/2}\sqrt{\hat{\sigma}^2(log(\hat{\lambda}_n))}\right),$$

where $\hat{\sigma}^2(log(\hat{\lambda_n}))$ is the estimated variance of $log(\hat{\lambda}_n)$ which is approximated by $\hat{\sigma}^2(log(\hat{\lambda_n})) \approx \frac{\hat{\sigma}^2(\hat{\lambda_n})}{\hat{\lambda_n}^2}$. Hence, an approximate $100(1-\alpha)\%$ confidence interval for λ is given by

$$\left(\hat{\lambda}_n\, e^{\left(-\frac{\tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\hat{\lambda}_n)}}{\hat{\lambda}_n}\right)}, \quad \hat{\lambda_n}\, e^{\left(\frac{\tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\hat{\lambda}_n)}}{\hat{\lambda}_n}\right)}\right). \quad (7)$$

Let $\hat{R}_n$ is the MLE of reliability function $R(t)$ and $\sigma^2(\hat{R}_n)$ is the variance of $\hat{R}_n$, where

$$\hat{\sigma}^2(\hat{R}_n) \approx \frac{k^2t^2}{\hat{\lambda}_n^4}\left[1-G\left(\frac{t}{\hat{\lambda}_n}\right)\right]^{2(k-1)}\left[G'\left(\frac{t}{\hat{\lambda}_n}\right)\right]^2\hat{\sigma}^2(\hat{\lambda}_n)$$

Therefore, $100(1-\alpha)\%$ asymptotic confidence interval for $R(t)$ is given by

$$\left(\hat{R}_n-\tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\hat{R}_n)},\quad \hat{R}_n+\tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\hat{R}_n)}\right), \tag{8}$$

3 Tolerance Intervals

Kumbhar and Shirke (2004) derived the expression for β-expectation tolerance interval for the lifetime distribution of a k-unit parallel system with component life as exponential distribution. They investigated the performance of the tolerance interval based on complete data. We study the performance of the tolerance interval for the lifetime distribution of a k-unit series system based on progressively Type-II censored data for the scale family of distributions. Let $l_\beta(\lambda)$ be the lower quantile of order β of the cdf $F(x;\lambda)$. Then, we have

$$l_\beta(\lambda) = \lambda G^{-1}\left[1-(1-\beta)^{1/k}\right].$$

Thus, an upper β-expectation tolerance interval for $F(x;\lambda)$ is obtained by

$$I_\beta = (0, l_\beta(\lambda)).$$

The maximum likelihood estimator of $l_\beta(\lambda)$ is given by

$$l_\beta(\hat{\lambda}_n) = \hat{\lambda}_n\, G^{-1}\left[1-(1-\beta)^{1/k}\right],$$

yielding an approximate β- expectation tolerance interval as

$$\hat{I}_\beta = \left(0,\ l_\beta(\hat{\lambda}_n)\right).$$

The expectation of $\hat{I}_\beta$ can be obtained approximately using the approach suggested by Atwood (1984) and given as,

$$E\left[F(I_\beta(\hat{\lambda}_n);\lambda)\right] \approx \beta - 0.5\, F_{02}\, \sigma^2(\hat{\lambda}_n) + \frac{F_{01}\, \sigma^2(\hat{\lambda}_n)\, F_{11}}{F_{10}}, \tag{9}$$

$$\text{where } F_{10} = \frac{dF}{dx},\quad F_{01} = \frac{dF}{d\lambda},\quad F_{11} = \frac{d^2F}{dxd\lambda},\quad F_{02} = \frac{d^2F}{d\lambda^2},$$

$$F_{10} = \frac{k}{\lambda}\left[1-G\left(\frac{x}{\lambda}\right)\right]^{k-1} g\left(\frac{x}{\lambda}\right),\quad F_{01} = -\frac{kx}{\lambda^2}\left[1-G\left(\frac{x}{\lambda}\right)\right]^{k-1} G'\left(\frac{x}{\lambda}\right),$$

$$F_{11} = -\frac{k}{\lambda^3}\left[1-G\left(\frac{x}{\lambda}\right)\right]^{k-2}\times$$

$$\left\{x\left[1-G\left(\frac{x}{\lambda}\right)\right]g'\left(\frac{x}{\lambda}\right)+x(k-1)G'\left(\frac{x}{\lambda}\right)g\left(\frac{x}{\lambda}\right)+\lambda\left[1-G\left(\frac{x}{\lambda}\right)\right]g\left(\frac{x}{\lambda}\right)\right\},$$

$$F_{02}=\frac{kx}{\lambda^4}\left[1-G\left(\frac{x}{\lambda}\right)\right]^{k-2}\times$$

$$\left\{x\left[1-G\left(\frac{x}{\lambda}\right)\right]G''\left(\frac{x}{\lambda}\right)-x(k-1)\left[G'\left(\frac{x}{\lambda}\right)\right]^2+2\lambda\left[1-G\left(\frac{x}{\lambda}\right)\right]G'\left(\frac{x}{\lambda}\right)\right\}.$$

The derivatives of F are evaluated at $x=l_\beta(\lambda)$ with $\lambda=\hat{\lambda}_n$. Instead of the actual value of $\sigma^2(\hat{\lambda}_n)$ we use estimate of it.

4 Application to Half-Logistic Distribution

Consider a member of the scale family of distributions, namely half-logistic distribution with scale parameter λ. The cdf of X is

$$F(x;\lambda)=1-\left[\frac{2e^{-x/\lambda}}{1+e^{-x/\lambda}}\right]^k \qquad x\geq 0,\ \lambda>0.$$

The pdf of X is

$$f(x;\lambda)=\frac{k}{\lambda}\frac{2^k e^{-kx/\lambda}}{\left(1+e^{-x/\lambda}\right)^{k+1}} \qquad x\geq 0,\ \lambda>0.$$

4.1 Maximum Likelihood Estimation

The complete log-likelihood function for half-logistic distribution with scale parameter λ from equation (1) is

$$L_c=nlog(k)-nlog(\lambda)+\sum_{i=1}^{m}log\left[\frac{2e^{-x_i/\lambda}}{\left(1+e^{-x_i/\lambda}\right)^2}\right]+(k-1)\sum_{i=1}^{m}log\left[\frac{2e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}\right]$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{R_i}log\left[\frac{2e^{-z_{ij}/\lambda}}{\left(1+e^{-z_{ij}/\lambda}\right)^2}\right]+(k-1)\sum_{i=1}^{m}\sum_{j=1}^{R_i}log\left[\frac{2e^{-z_{ij}/\lambda}}{1+e^{-z_{ij}/\lambda}}\right]. \tag{10}$$

In order to obtain MLE of λ, we use EM algorithm due to Dempster et al. (1977). For the E step in EM algorithm we take Expectation of Z_{ij}. The derivative of L_c with respect to λ is taken for the M step, where

$$\frac{dL_c}{d\lambda}=-\frac{n}{\lambda}+\frac{k}{\lambda^2}\sum_{i=1}^{m}x_i-\frac{(k+1)}{\lambda^2}\sum_{i=1}^{m}\frac{x_i e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}+\frac{k}{\lambda^2}\sum_{i=1}^{m}R_i a(x_i,k,\lambda^0)$$

$$-\frac{(k+1)}{\lambda^2}\sum_{i=1}^{m}R_i b(x_i,k,\lambda^0). \tag{11}$$

$$\text{where } a(x_i,k,\lambda) = E\left(Z_{ij}\right) \text{and } b(x_i,k,\lambda) = E\left[\frac{Z_{ij}e^{-Z_{ij}/\lambda}}{1+e^{-Z_{ij}/\lambda}}\right].$$

To solve this equation, we use Newton-Raphson method.
Reliability function at time t is

$$R(t) = \left[\frac{2e^{-t/\lambda}}{1+e^{-t/\lambda}}\right]^k \qquad t \geq 0,\ \lambda > 0.$$

The Maximum likelihood estimate of $R(t)$ is

$$\hat{R}_n(t) = \left[\frac{2e^{-t/\hat{\lambda}_n}}{1+e^{-t/\hat{\lambda}_n}}\right]^k \qquad t \geq 0.$$

4.2 Fisher Information

The observed information = complete information - missing information.

$$I_x(\lambda) = I_w(\lambda) - I_{w|x}(\lambda),$$

Consider log-likelihood function for n observations is

$$L = nlog(k) - nlog(\lambda) + \sum_{i=1}^{n} log\left[\frac{2e^{-x_i/\lambda}}{\left(1+e^{-x_i/\lambda}\right)^2}\right] + (k-1)\sum_{i=1}^{n} log\left[\frac{2e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}\right]. \quad (12)$$

Then complete information is

$$I_w(\lambda) = -E\left[\frac{d^2L}{d\lambda^2}\right] = -\frac{n}{\lambda^2} + \frac{2k}{\lambda^3}\sum_{i=1}^{n} E\left[X_i\right] + \frac{(k+1)}{\lambda^4}\sum_{i=1}^{n} E\left[\frac{X_i^2 e^{-X_i/\lambda}}{(1+e^{-X_i/\lambda})^2}\right]$$

$$-\frac{2(k+1)}{\lambda^3}\sum_{i=1}^{n} E\left[\frac{X_i e^{-X_i/\lambda}}{1+e^{-X_i/\lambda}}\right]. \quad (13)$$

and missing information is given by

$$I_{w|x}(\lambda) = \sum_{i=1}^{m} R_i I_{w|x}^{(i)}(\lambda) = -\sum_{i=1}^{m}\sum_{j=1}^{R_i} E_{Z|X}\left[\frac{d^2 log\left(f(Z_{ij}|x_i,\lambda)\right)}{d\lambda^2}\right]$$

$$= -\frac{n-m}{\lambda^2} + \frac{2k}{\lambda^3}\sum_{i=1}^{m}\sum_{j=1}^{R_i} E\left[Z_{ij}\right] + \frac{(k+1)}{\lambda^4}\sum_{i=1}^{m}\sum_{j=1}^{R_i} E\left[\frac{Z_{ij}^2 e^{-Z_{ij}/\lambda}}{(1+e^{-Z_{ij}/\lambda})^2}\right]$$

$$-\frac{2(k+1)}{\lambda^3}\sum_{i=1}^{m}\sum_{j=1}^{R_i} E\left[\frac{Z_{ij}e^{-Z_{ij}/\lambda}}{1+e^{-Z_{ij}/\lambda}}\right] - \frac{k}{\lambda^4}\sum_{i=1}^{m}\left[\frac{R_i x_i^2 e^{-x_i/\lambda}}{(1+e^{-x_i/\lambda})^2}\right]$$

$$+\frac{2k}{\lambda^3}\sum_{i=1}^{m}\left[\frac{R_i x_i e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}\right]-\frac{2k}{\lambda^3}\sum_{i=1}^{m}R_i x_i. \tag{14}$$

4.3 Confidence Interval and Tolerance Interval

Using equations (6) - (8) with $\hat{\sigma}^2(\hat{\lambda}_n)=\frac{1}{I_x(\hat{\lambda}_n)}$ and

$$\sigma^2(\hat{R}_n(t)) \approx \left[\frac{kt}{\hat{\lambda}_n^2}\frac{\left(2e^{-t/\hat{\lambda}_n}\right)^k}{\left(1-e^{-t/\hat{\lambda}_n}\right)^{k+1}}\right]^2 \sigma^2(\hat{\lambda}_n)$$

we construct confidence intervals for scale parameter and reliability function.

Let $l_\beta(\lambda)$ be the lower quantile of order β of the cdf $F(x;\lambda)$. Then, we have

$$l_\beta(\lambda)=\lambda log\left[\frac{2-(1-\beta)^{1/k}}{(1-\beta)^{1/k}}\right],$$

Thus, an upper β-expectation Tolerance Interval for $F(x;\lambda)$ is obtained by

$$I_\beta=(0,l_\beta(\lambda)).$$

The maximum likelihood estimator of $l_\beta(\lambda)$ is given by

$$l_\beta(\hat{\lambda}_n)=\hat{\lambda}_n log\left[\frac{2-(1-\beta)^{1/k}}{(1-\beta)^{1/k}}\right],$$

yielding an approximate β- expectation tolerance interval as

$$\hat{I}_\beta=\left(0,\ l_\beta(\hat{\lambda}_n)\right).$$

The expectation of $\hat{I}_\beta$ can be obtained approximately using the approach suggested and given as,

$$E\left[F(I_\beta(\hat{\lambda}_n);\lambda)\right] \approx \beta - 0.5\ F_{02}\ \sigma^2(\hat{\lambda}_n)+\frac{F_{01}\ \sigma^2(\hat{\lambda}_n)\ F_{11}}{F_{10}}, \tag{15}$$

$$\text{where } F_{10}=\frac{k2^k}{\lambda}\frac{\left(e^{-x/\lambda}\right)^k}{\left(1+e^{-x/\lambda}\right)^{k+1}},\qquad F_{01}=-\frac{kx2^k}{\lambda^2}\frac{\left(e^{-x/\lambda}\right)^k}{\left(1+e^{-x/\lambda}\right)^{k+1}},$$

$$F_{11}=\frac{k2^k}{\lambda^3}\frac{\left(e^{-x/\lambda}\right)^k}{\left(1+e^{-x/\lambda}\right)^{k+2}}\left[(kx-\lambda)-e^{-x/\lambda}(x+\lambda)\right],$$

$$\text{and }\ F_{02}=-\frac{kx2^k}{\lambda^4}\frac{\left(e^{-x/\lambda}\right)^k}{\left(1+e^{-x/\lambda}\right)^{k+2}}\left[(kx-2\lambda)-e^{-x/\lambda}(x+2\lambda)\right].$$

5 Simulation Study

A simulation study is carried out to investigate the performance of MLE, reliability estimate and confidence interval of the scale parameter of half-logistic distribution. We obtain estimate of bias and MSE for various progressively Type-II censoring scheme. Asymptotic confidence intervals based on the MLE and log-transformed MLE are compared through their confidence levels. The coverage of the β- expectation tolerance intervals is studied using simulation. Balakrishnan and Sandhu (1995) presented algorithm for sample generation from progressively Type-II censored scheme. This algorithm is used to generate progressively censored samples from half-logistic distribution of a k-unit series system.

Algorithm

1. Generate independently and identically distributed observations $(W_1, W_2, ..., W_m)$ from U(0, 1).
2. For $(R_1, R_2, ..., R_m)$ progressive Type-II censoring scheme, set $E_i = 1/(i + R_m + R_{m-1} + ... + R_{m-i+1})$ for $i = 1, 2, ..., m$.
3. Set $V_i = W_i^{E_i}$ for $i = 1, 2, ..., m$.
4. Set $U_i = 1 - V_m V_{m-1} ..... V_{m-i+1}$ for $i = 1, 2, ..., m$. Then $(U_1, U_2, ..., U_m)$ is the U(0, 1) progressively Type-II censored sample.
5. For the given value of the parameter λ, set

$$x_i = \lambda \, log \left[\frac{2 - (1 - U_i)^{1/k}}{(1 - U_i)^{1/k}} \right] \qquad \text{for } i = 1, 2, ..., m. \tag{16}$$

Then $(x_1, x_2, ..., x_m)$ is the required progressively Type-II censored sample from the distribution of a k-unit series system with half-logistic distribution as the component life distribution In Table 1 scheme (a, b) stands for $R_1 = a$ and $R_2 = b$. Similar meaning holds for schemes described through completely specified vector, while scheme $(10, 4 * 0)$ means $R_1 = 10$ and rest four $R_i s$ are zero. i.e. $R_2 = R_3 = R_4 = R_5 = 0$. A simulation was carried out for 2-unit, 3-unit and 5-unit series system (i.e. k=2, 3 and 5) with $\lambda = 1$. EM algorithm and Newton-Raphson method are used to compute MLE. For each particular progressive censoring scheme, 10,000 sets of observations were generated. The bias, MSE, confidence levels with their standard errors (SE) for the corresponding confidence intervals for λ are displayed in Table 1 - 3 for k=2, 3 and 5 respectively. The bias, MSE, confidence levels with their SE for the confidence intervals for reliability function are displayed in Table 4 - 6 for k=2, 3 and 5 respectively. The simulated mean coverage and the estimated expectation of the tolerance interval are given in Table 7 - 9. ($^+$MSE and SE are given in parenthesis.)

Table 1: Bias, MSE$^+$, Confidence levels and its SE$^+$ for MLE (k=2)

n	m	Scheme No.	Scheme	Bias and MSE	Level and SE-MLE 90%	Level and SE-MLE 95%	Level and SE-log(MLE) 90%	Level and SE-log(MLE) 95%
5	2	[1]	(3,0)	-0.0708	0.7481	0.7802	0.8411	0.8883
				(0.3379)	(0.0377)	(0.0343)	(0.0267)	(0.0198)
		[2]	(0,3)	-0.0654	0.7525	0.7824	0.8445	0.8934
				(0.3529)	(0.0372)	(0.0341)	(0.0263)	(0.0190)
		[3]	(1,2)	-0.0694	0.7540	0.7884	0.8489	0.8907
				(0.3497)	(0.0371)	(0.0334)	(0.0257)	(0.0195)
		[4]	(2,1)	-0.0642	0.7520	0.7868	0.8427	0.8900
				(0.3559)	(0.0373)	(0.0335)	(0.0265)	(0.0196)
15	5	[5]	(10, 4*0)	-0.0248	0.8339	0.8656	0.8727	0.9263
				(0.1425)	(0.0092)	(0.0078)	(0.0074)	(0.0046)
		[6]	(4*0, 10)	-0.0196	0.8325	0.8693	0.8807	0.9313
				(0.1624)	(0.0093)	(0.0076)	(0.0070)	(0.0043)
		[7]	(2,2,2,2,2)	-0.0205	0.8315	0.8643	0.8777	0.9303
				(0.1546)	(0.0093)	(0.0078)	(0.0072)	(0.0043)
	10	[8]	(5,9*0)	-0.0121	0.8652	0.9041	0.8902	0.9401
				(0.0702)	(0.0078)	(0.0058)	(0.0065)	(0.0038)
		[9]	(9*0,5)	-0.0141	0.8680	0.9037	0.8941	0.9434
				(0.0723)	(0.0076)	(0.0058)	(0.0063)	(0.0036)
		[10]	(3,2, 8*0)	-0.0134	0.8694	0.9057	0.8894	0.9368
				(0.0713)	(0.0076)	(0.0057)	(0.0066)	(0.0039)
20	10	[11]	(10, 9*0)	-0.0117	0.8669	0.9045	0.8863	0.9391
				(0.0705)	(0.0058)	(0.0043)	(0.005)	(0.0029)
		[12]	(9*0,10)	-0.0086	0.8686	0.9069	0.8936	0.9423
				(0.0767)	(0.0057)	(0.0042)	(0.0048)	(0.0027)
25	10	[13]	(15,9*0)	-0.0167	0.8679	0.9070	0.8927	0.9398
				(0.0693)	(0.0046)	(0.0034)	(0.0038)	(0.0023)
		[14]	(9*0,15)	-0.0161	0.8613	0.8973	0.8829	0.9356
				(0.0805)	(0.0048)	(0.0037)	(0.0041)	(0.0024)
		[15]	(5,5,5,7*0)	-0.0106	0.8641	0.9033	0.8893	0.9401
				(0.0733)	(0.0047)	(0.0035)	(0.0039)	(0.0023)
	15	[16]	(10, 14*0)	-0.0099	0.8792	0.9198	0.8952	0.9455
				(0.0464)	(0.0042)	(0.003)	(0.0038)	(0.0021)
		[17]	(14*0,10)	-0.0123	0.8745	0.9160	0.8935	0.9458
				(0.0499)	(0.0044)	(0.0031)	(0.0038)	(0.0021)
30	10	[18]	(20, 9*0)	-0.0079	0.8676	0.9070	0.8889	0.9366
				(0.0725)	(0.00380	(0.0028)	(0.0033)	(0.002)
		[19]	(9*0,20)	-0.0100	0.8637	0.8994	0.8888	0.9389
				(0.0844)	(0.0039)	(0.003)	(0.0033)	(0.0019)
	15	[20]	(15, 14*0)	-0.0089	0.8745	0.9142	0.8865	0.9400
				(0.0481)	(0.0037)	(0.0026)	(0.0034)	(0.0019)
		[21]	(14*0,15)	-0.0087	0.8792	0.9171	0.8940	0.9460
				(0.0523)	(0.0035)	(0.0025)	(0.0032)	(0.0017)
		[22]	(5,5,5,12*0)	-0.0073	0.8777	0.9219	0.8960	0.9437
				(0.0474)	(0.0036)	(0.0024)	(0.0031)	(0.0018)
	20	[23]	(10, 19*0)	-0.0040	0.8859	0.9281	0.8942	0.9452
				(0.0355)	(0.0034)	(0.0022)	(0.0032)	(0.0017)
		[24]	(19*0,10)	-0.0064	0.8891	0.9287	0.8973	0.9460
				(0.0366)	(0.0033)	(0.0022)	(0.0031)	(0.0017)
		[25]	(0,5,5,17*0)	-0.0064	0.8839	0.9273	0.8946	0.9449
				(0.0356)	(0.0034)	(0.0022)	(0.0031)	(0.0017)
50	20	[26]	(30,19*0)	-0.0058	0.8827	0.9245	0.8945	0.9440
				(0.0360)	(0.0021)	(0.0014)	(0.0019)	(0.0011)
		[27]	(19*0,30)	-0.0095	0.8773	0.9218	0.8892	0.9423
				(0.0411)	(0.0022)	(0.0014)	(0.002)	(0.0011)
	35	[28]	(15,34*0)	-0.0021	0.8920	0.9350	0.8950	0.9467
				(0.0207)	(0.0019)	(0.0012)	(0.0019)	(0.0010)
		[29]	(34*0,15)	-0.0054	0.8920	0.9346	0.8980	0.9473
				(0.0211)	(0.0019)	(0.0012)	(0.0018)	(0.0010)
		[30]	(5,5,5,32*0)	-0.0044	0.8898	0.9342	0.8962	0.9444
				(0.0205)	(0.0020)	(0.0012)	(0.0019)	(0.0011)

Table 2: Bias, MSE$^+$, Confidence levels and its SE$^+$ for MLE (k=3)

n	m	Scheme No.	Scheme	Bias and MSE	Level and SE-MLE 90%	Level and SE-MLE 95%	Level and SE-log(MLE) 90%	Level and SE-log(MLE) 95%
5	2	[1]	(3,0)	-0.0492	0.7498	0.7796	0.8368	0.8927
				(0.3704)	(0.0375)	(0.0344)	0.0273	(0.0192)
		[2]	(0,3)	-0.0535	0.7506	0.7858	0.8496	0.8980
				(0.3822)	(0.0374)	(0.0337)	(0.0256)	(0.0183)
		[3]	(1,2)	-0.0356	0.7606	0.7934	0.8535	0.9016
				(0.3921)	(0.0364	(0.0328)	(0.0250)	(0.0177)
		[4]	(2,1)	-0.0535	0.7549	0.7849	0.8443	0.8912
				(0.3774)	(0.0370)	(0.0338	(0.0263)	(0.0194)
15	5	[5]	(10, 4*0)	-0.0265	0.8251	0.8630	0.8742	0.9228
				(0.1503)	(0.0096)	(0.0079)	(0.0073)	(0.0047)
		[6]	(4*0, 10)	-0.0210	0.8300	0.8662	0.8787	0.9272
				(0.1705)	(0.0094)	(0.0077)	(0.0071)	(0.0045)
		[7]	(2,2,2,2,2)	-0.0271	0.8284	0.8612	0.8767	0.9246
				(0.1635)	(0.0095)	(0.0080)	(0.0072)	(0.0046)
	10	[8]	(5,9*0)	-0.0107	0.8658	0.9070	0.8922	0.9408
				(0.0733)	(0.0077)	(0.0056)	(0.0064)	(0.0037)
		[9]	(9*0,5)	-0.0103	0.8657	0.9024	0.8868	0.9396
				(0.0794)	(0.0078	(0.0059)	(0.0067)	(0.0038)
		[10]	(3,2, 8*0)	-0.0117	0.8685	0.9042	0.8905	0.9390
				(0.0719)	(0.0076)	(0.0058)	(0.0065)	(0.0038)
20	10	[11]	(10, 9*0)	-0.0136	0.8676	0.9055	0.8905	0.9426
				(0.0720)	(0.0057)	(0.0043)	(0.0049)	(0.0027)
		[12]	(9*0,10)	-0.0120	0.8653	0.9043	0.8924	0.9421
				(0.0818)	(0.0058)	(0.0043)	(0.0048)	(0.0027)
25	10	[13]	(15,9*0)	-0.0151	0.8612	0.8983	0.8815	0.9325
				(0.0756)	(0.0048)	(0.0037)	(0.0042)	(0.0025)
		[14]	(9*0,15)	-0.0098	0.8644	0.9023	0.8889	0.9385
				(0.0859)	(0.0047)	(0.0035)	(0.0040)	(0.0023)
		[15]	(5,5,5,7*0)	-0.0126	0.8639	0.9013	0.8875	0.9359
				(0.0764)	(0.0047)	(0.0036)	(0.0040)	(0.0024)
	15	[16]	(10, 14*0)	-0.0100	0.8714	0.9141	0.8881	0.9384
				(0.0493)	(0.0045)	(0.0031)	(0.0040)	(0.0023)
		[17]	(14*0,10)	-0.0098	0.8755	0.9121	0.8903	0.9407
				(0.0545)	(0.0044)	(0.0032)	(0.0039)	(0.0022)
30	10	[18]	(20, 9*0)	-0.0139	0.8649	0.9041	0.8878	0.9385
				(0.0737)	(0.0039)	(0.0029)	(0.0033)	(0.0019)
		[19]	(9*0, 20)	-0.0045	0.8666	0.9014	0.8877	0.9377
				(0.0894)	(0.0039)	(0.0030)	(0.0033)	(0.0019)
	15	[20]	(15, 14*0)	-0.0104	0.8766	0.9156	0.8893	0.9419
				(0.0493)	(0.0036)	(0.0026)	(0.0033)	(0.0018)
		[21]	(14*0,15)	-0.0091	0.8715	0.9137	0.8876	0.9379
				(0.0563)	(0.0037)	(0.0026)	(0.0033)	(0.0019)
		[22]	(5,5,5,12*0)	-0.0110	0.8767	0.9158	0.8889	0.9419
				(0.0497)	(0.0036)	(0.0026)	(0.0033)	(0.0018)
	20	[23]	(10, 19*0)	-0.0084	0.8789	0.9245	0.8937	0.9424
				(0.0369)	(0.0035)	(0.0023)	(0.0032)	(0.0018)
		[24]	(19*0,10)	-0.0052	0.8813	0.9252	0.8942	0.9428
				(0.0395)	(0.0035)	(0.0023)	(0.0032)	(0.0018)
		[25]	(0,5,5,17*0)	-0.0043	0.8831	0.9257	0.8937	0.9437
				(0.0378)	(0.0034)	(0.0023)	(0.0032)	(0.0018)
50	20	[26]	(30,19*0)	-0.0052	0.8821	0.9243	0.8894	0.9426
				(0.0375)	(0.0021)	(0.0014)	(0.0020)	(0.0011)
		[27]	(19*0,30)	-0.0060	0.8839	0.9248	0.8955	0.9459
				(0.0438)	(0.0021)	(0.0014)	(0.0019)	(0.0010)
	35	[28]	(15,34*0)	-0.0043	0.8865	0.9317	0.8919	0.9441
				(0.0212)	(0.0020)	(0.0013)	(0.0019)	(0.0011)
		[29]	(34*0,15)	-0.0025	0.8944	0.9404	0.8998	0.9473
				(0.0223)	(0.0019)	(0.0011)	(0.0018)	(0.0010)
		[30]	(5,5,5,32*0)	-0.0028	0.8896	0.9364	0.8965	0.9449
				(0.0215)	(0.0020)	(0.0012)	(0.0019)	(0.0010)

Table 3: Bias, MSE$^+$, Confidence levels and its SE$^+$ for MLE (k=5)

n	m	Scheme No.	Scheme	Bias and MSE	Level and SE (MLE) 90%	95%	Level and SE (log(MLE)) 90%	95%
5	2	[1]	(3,0)	-0.05431	0.7545	0.7878	0.8445	0.8924
				(0.3776)	(0.0370)	(0.0334)	(0.0263)	(0.0192)
		[2]	(0,3)	-0.0283	0.7489	0.7825	0.8444	0.8959
				(0.4394)	(0.0376)	(0.0340)	(0.0263)	(0.0187)
		[3]	(1,2)	-0.0329	0.7626	0.7932	0.8498	0.9024
				(0.4076)	(0.0362)	(0.0328)	(0.0255)	(0.0176)
		[4]	(2,1)	-0.0372	0.7536	0.7861	0.8441	0.8948
				(0.4153)	(0.0371)	(0.0336)	(0.0263)	(0.0188)
15	5	[5]	(10, 4*0)	-0.0191	0.8306	0.8668	0.8755	0.9279
				(0.1563)	(0.0094)	(0.0077)	(0.0073)	(0.0045)
		[6]	(4*0, 10)	-0.0097	0.8273	0.8608	0.8750	0.9271
				(0.1875)	(0.0095)	(0.0080)	(0.0073)	(0.0045)
		[7]	(2,2,2,2,2)	-0.0211	0.8252	0.8570	0.8703	0.9209
				(0.1758)	(0.0096)	(0.0082)	(0.0075)	(0.0049)
	10	[8]	(5,9*0)	-0.0138	0.8668	0.9050	0.8928	0.9384
				(0.0761)	(0.0077)	(0.0057)	(0.0064)	(0.0039)
		[9]	(9*0,5)	-0.0107	0.8629	0.8959	0.8868	0.9375
				(0.0842)	(0.0079)	(0.0062)	(0.0067)	(0.0039)
		[10]	(3,2, 8*0)	-0.0143	0.8562	0.8966	0.8827	0.9324
				(0.0801)	(0.0082)	(0.0062)	(0.0069)	(0.0042)
20	10	[11]	(10, 9*0)	-0.0128	0.8641	0.9033	0.8893	0.9378
				(0.0783)	(0.0059)	(0.0044)	(0.0049)	(0.0029)
		[12]	(9*0,10)	-0.0093	0.8680	0.9027	0.8897	0.9413
				(0.0870)	(0.0057)	(0.0044)	(0.0049)	(0.0028)
25	10	[13]	(15,9*0)	-0.0134	0.8651	0.9032	0.8893	0.9365
				(0.0777)	(0.0047)	(0.0035)	(0.0039)	(0.0024)
		[14]	(9*0,15)	-0.0133	0.8682	0.9025	0.8927	0.9419
				(0.0870)	(0.0046)	(0.0035)	(0.0038)	(0.0022)
		[15]	(5,5,5,7*0)	-0.0079	0.8670	0.9058	0.8930	0.9400
				(0.0797)	(0.0046)	(0.0034)	(0.0038)	(0.0023)
	15	[16]	(10, 14*0)	-0.0110	0.8777	0.9171	0.8914	0.9409
				(0.0515)	(0.0043)	(0.0030)	(0.0039)	(0.0022)
		[17]	(14*0,10)	-0.0093	0.8750	0.9158	0.8923	0.9426
				(0.0580)	(0.0044)	(0.0031)	(0.0038)	(0.0022)
30	10	[18]	(20, 9*0)	-0.0138	0.8602	0.8968	0.8843	0.9362
				(0.0791)	(0.0040)	(0.0031)	(0.0034)	(0.0020)
		[19]	(9*0,20)	-0.0064	0.8660	0.9018	0.8886	0.9375
				(0.0920)	(0.0039)	(0.0030)	(0.0033)	(0.0020)
	15	[20]	(15, 14*0)	-0.0097	0.8782	0.9188	0.8932	0.9419
				(0.0517)	(0.0036)	(0.0025)	(0.0032)	(0.0018)
		[21]	(14*0,15)	-0.0022	0.8819	0.9234	0.8991	0.9468
				(0.0578)	(0.0035)	(0.0024)	(0.0030)	(0.0017)
		[22]	(5,5,5,12*0)	-0.0095	0.8808	0.9204	0.8950	0.9427
				(0.0517)	(0.0035)	(0.0024)	(0.0031)	(0.0018)
	20	[23]	(10, 19*0)	-0.0066	0.8864	0.9239	0.8936	0.9458
				(0.0389)	(0.0034)	(0.0023)	(0.0032)	(0.0017)
		[24]	(19*0,10)	-0.0071	0.8796	0.9226	0.8955	0.9445
				(0.0424)	(0.0035)	(0.0024)	(0.0031)	(0.0017)
		[25]	(0,5,5,17*0)	-0.0067	0.8841	0.9262	0.8961	0.9423
				(0.0391)	(0.0034)	(0.0023)	(0.0031)	(0.0018)
50	20	[26]	(30,19*0)	-0.0117	0.8801	0.9221	0.8947	0.9436
				(0.0389)	(0.0021)	(0.0014)	(0.0019)	0.0011
		[27]	(19*0,30)	-0.0057	0.8840	0.9238	0.8939	0.9455
				(0.0447)	(0.0021)	(0.0014)	(0.0019)	(0.0010)
	35	[28]	(15,34*0)	-0.0059	0.8806	0.9316	0.8891	0.9434
				(0.0228)	(0.0021)	(0.0013)	(0.0020)	(0.0011)
		[29]	(34*0,15)	-0.0030	0.8936	0.9378	0.8979	0.9468
				(0.0242)	(0.0019)	(0.0012)	(0.0018)	(0.0010)
		[30]	(5,5,5,32*0)	-0.0022	0.8887	0.9416	0.9035	0.9511
				(0.0219)	(0.0020)	(0.0011)	(0.0017)	(0.0009)

Table 4: Bias, MSE^+, Confidence levels and its SE^+ for R(t) (k=2)

n	m	Scheme No.	Scheme	Bias and MSE	Level and SE (MLE) 90%	95%
5	2	[1]	(3,0)	-0.1108	0.7909	0.8309
				(0.0660)	(0.0331)	(0.0281)
		[2]	(0,3)	-0.1142	0.7909	0.8350
				(0.0677)	(0.0331)	-0.0276
		[3]	(1,2)	-0.1088	0.7963	0.8391
				(0.0657)	(0.0324)	(0.0270)
		[4]	(2,1)	-0.1182	0.7861	0.8267
				(0.0687)	(0.0336)	(0.0287)
15	5	[5]	(10, 4*0)	-0.0472	0.8634	0.9116
				(0.0246)	(0.0079)	(0.0054)
		[6]	(4*0, 10)	0.0541	0.8580	0.9067
				(0.0282)	(0.0081)	(0.0056)
		[7]	(2,2,2,2,2)	-0.0499	0.8599	0.9077
				(0.0270)	(0.0080)	(0.0056)
	10	[8]	(5,9*0)	-0.0229	0.8826	0.9372
				(0.0108)	(0.0069)	(0.0039)
		[9]	(9*0,5)	-0.0260	0.8846	0.9308
				(0.0117)	(0.0068)	(0.0043)
		[10]	(3,2, 8*0)	-0.0226	0.8859	0.9337
				(0.0108)	(0.0067)	(0.0041)
20	10	[11]	(10, 9*0)	-0.0260	0.8811	0.9316
				(0.0113)	(0.0052)	(0.0032)
		[12]	(9*0,10)	-0.0276	0.8884	0.9380
				(0.0120)	(0.0050)	(0.0029)
25	10	[13]	(15,9*0)	-0.0239	0.8864	0.9359
				(0.0109)	(0.0040)	(0.0024)
		[14]	(9*0,15)	-0.0279	0.8795	0.9312
				(0.0130)	(0.0042)	(0.0026)
		[15]	(5,5,5,7*0)	-0.0237	0.8847	0.9345
				(0.0110)	(0.0041)	(0.0024)
	15	[16]	(10, 14*0)	-0.0157	0.8886	0.9409
				(0.0069)	(0.0040)	(0.0022)
		[17]	(14*0,10)	-0.0152	0.8943	0.9425
				(0.0071)	(0.0038)	(0.0022)
30	10	[18]	(20, 9*0)	-0.0269	0.8783	0.9258
				(0.0116)	(0.0036)	(0.0023)
		[19]	(9*0,20)	-0.0261	0.8726	0.9253
				(0.0136)	(0.0037)	(0.0023)
	15	[20]	(15, 14*0)	-0.0158	0.8897	0.9394
				(0.0069)	(0.0033)	(0.0019)
		[21]	(14*0,15)	-0.0191	0.8841	0.9373
				(0.0081)	(0.0034)	(0.0020)
		[22]	(5,5,5,12*0)	-0.0155	0.8921	0.9422
				(0.0069)	(0.0032)	(0.0018)
	20	[23]	(10, 19*0)	-0.0107	0.8969	0.9454
				(0.0049)	(0.0031)	(0.0017)
		[24]	(19*0,10)	-0.0142	0.8935	0.9431
				(0.0054)	(0.0032)	(0.0018)
		[25]	(0,5,5,17*0)	-0.0133	0.8944	0.9457
				(0.0050)	(0.0031)	(0.0017)
50	20	[26]	(30,19*0)	-0.0119	0.8903	0.9410
				(0.0051)	(0.0020)	(0.0011)
		[27]	(19*0,30)	-0.0159	0.8906	0.9390
				(0.0060)	(0.0019)	(0.0011)
	35	[28]	(15,34*0)	-0.0069	0.8969	0.9446
				(0.0028)	(0.0018)	(0.0010)
		[29]	(34*0,15)	-0.0076	0.8934	0.9479
				(0.0029)	(0.0019)	(0.0010)
		[30]	(5,5,5,32*0)	-0.0069	0.8924	0.9440
				(0.0028)	(0.0019)	(0.0011)

Table 5: Bias, MSE$^+$, Confidence levels and its SE$^+$ for R(t) (k=3)

n	m	Scheme No.	Scheme	Bias and MSE	Level and SE (MLE) 90%	95%
5	2	[1]	(3,0)	-0.0947	0.7412	0.7829
				(0.0570)	(0.0384)	(0.0340)
		[2]	(0,3)	-0.0880	0.7470	0.7823
				(0.0578)	(0.0378)	(0.0341)
		[3]	(1,2)	-0.0886	0.7510	0.7878
				(0.0570)	(0.0374)	(0.0334)
		[4]	(2,1)	-0.0896	0.7463	0.7854
				(0.0570)	(0.0379)	(0.0337)
15	5	[5]	(10, 4*0)	-0.0435	0.8423	0.8914
				(0.0254)	(0.0089)	(0.0065)
		[6]	(4*0, 10)	-0.0455	0.8302	0.8784
				(0.0287)	(0.0094)	(0.0071)
		[7]	(2,2,2,2,2)	-0.0456	0.8334	0.8828
				(0.0275)	(0.0093)	(0.0069)
	10	[8]	(5,9*0)	-0.0247	0.8723	0.9193
				(0.0128)	(0.0074)	(0.0049)
		[9]	(9*0,5)	-0.0247	0.8706	0.9166
				(0.0136)	(0.0075)	(0.0051)
		[10]	(3,2, 8*0)	-0.0228	0.8657	0.9189
				(0.0129)	(0.0078)	(0.0050)
20	10	[11]	(10, 9*0)	-0.0229	0.8650	0.9164
				(0.0129)	(0.0058)	(0.0038)
		[12]	(9*0,10)	-0.0244	0.8691	0.9199
				(0.0140)	(0.0057)	(0.0037)
25	10	[13]	(15,9*0)	-0.0234	0.8638	0.9140
				(0.0130)	(0.0047)	(0.0031)
		[14]	(9*0,15)	-0.0244	0.8675	0.9146
				(0.0146)	(0.0046)	(0.0031)
		[15]	(5,5,5,7*0)	-0.0228	0.8703	0.9161
				(0.0131)	(0.0045	(0.0031)
	15	[16]	(10, 14*0)	-0.0149	0.8772	0.9282
				(0.0085)	(0.0043)	(0.0027)
		[17]	(14*0,10)	-0.0174	0.8764	0.9290
				(0.0093)	(0.0043)	nn(0.0026)
30	10	[18]	(20, 9*0)	-0.0240	0.8693	0.9219
				(0.0127)	(0.0038)	(0.0024)
		[19]	(9*0,20)	-0.0249	0.8606	0.9133
				(0.0151)	(0.0040)	(0.0026)
	15	[20]	(15, 14*0)	-0.015	0.8795	0.9286
				(0.0085)	(0.0035)	(0.0022)
		[21]	(14*0,15)	-0.0157	0.8836	0.9325
				(0.0092)	(0.0034)	(0.0021)
		[22]	(5,5,5,12*0)	-0.0166	0.8776	0.9279
				(0.0087)	(0.0036)	(0.0022)
	20	[23]	(10, 19*0)	-0.0115	0.8863	0.9389
				(0.0062)	(0.0034)	(0.0019)
		[24]	(19*0,10)	-0.0129	0.8887	0.9385
				(0.0066)	(0.0033)	(0.0019)
		[25]	(0,5,5,17*0)	-0.0127	0.8783	0.9295
				(0.0064)	(0.0036)	(0.0022)
50	20	[26]	(30,19*0)	-0.0106	0.8860	0.9359
				(0.0064)	(0.0020)	(0.0012)
		[27]	(19*0,30)	-0.0127	0.8809	0.9323
				(0.0074)	(0.0021)	(0.0013)
	35	[28]	(15,34*0)	-0.0078	0.8872	0.9387
				(0.0036)	(0.0020)	(0.0012)
		[29]	(34*0,15)	-0.0072	0.8872	0.9408
				(0.0038)	(0.0020)	(0.0011)
		[30]	(5,5,5,32*0)	-0.0065	0.8909	0.9392
				(0.0035)	(0.0019)	(0.0011)

Table 6: Bias, MSE^+, Confidence levels and its SE^+ for R(t) (k=5)

n	m	Scheme No.	Scheme	Bias and MSE	Level and SE (MLE) 90%	95%
5	2	[1]	(3,0)	-0.0348 (0.0340)	0.6993 (0.0421)	0.7390 (0.0390
		[2]	(0,3)	-0.0363 (0.0347)	0.7000 (0.0420)	0.7319 (0.0392)
		[3]	(1,2)	-0.0341 (0.0349)	0.7020 (0.0418)	0.7344 (0.0390)
		[4]	(2,1)	-0.0341 (0.0343)	0.7028 (0.0418)	0.7364 (0.0388)
15	5	[5]	(10, 4*0)	-0.0203 (0.0176)	0.8093 (0.0103)	0.8515 (0.0084)
		[6]	(4*0, 10)	-0.0165 (0.0197)	0.8024 (0.0106)	0.8437 (0.0088)
		[7]	(2,2,2,2,2)	-0.0193 (0.0198)	0.7933 (0.0109)	0.8396 (0.0090)
	10	[8]	(5,9*0)	-0.0126 (0.0100)	0.8493 (0.0085)	0.8907 (0.0065)
		[9]	(9*0,5)	-0.0133 (0.0109)	0.8360 (0.0091)	0.8808 (0.0070)
		[10]	(3,2, 8*0)	-0.0117 (0.0101)	0.8493 (0.0085)	0.8956 (0.0062)
20	10	[11]	(10, 9*0)	-0.0124 (0.0104)	0.8436 (0.0066)	0.8871 (0.0050)
		[12]	(9*0,10)	-0.0107 (0.0115)	0.8366 (0.0068)	0.8852 (0.0051)
25	10	[13]	(15,9*0)	-0.0133 (0.0102)	0.8450 (0.0052)	0.8931 (0.0038)
		[14]	(9*0,15)	-0.0116 (0.0112)	0.8465 (0.0052)	0.8919 (0.0039)
		[15]	(5,5,5,7*0)	-0.0123 (0.0105)	0.8414 (0.0053)	0.8900 (0.0039)
	15	[16]	(10, 14*0)	-0.0089 (0.0069)	0.8705 (0.0045)	0.9134 (0.0032)
		[17]	(14*0,10)	-0.0092 (0.0078)	0.8586 (0.0049)	0.9025 (0.0035)
30	10	[18]	(20, 9*0)	-0.0120 (0.0101)	0.8492 (0.0043)	0.8970 (0.0031)
		[19]	(9*0,20)	-0.0106 (0.0116)	0.8432 (0.0044)	0.8864 (0.0034)
	15	[20]	(15, 14*0)	-0.0095 (0.0070)	0.8617 (0.0040)	0.9108 (0.0027)
		[21]	(14*0,15)	-0.0087 0.0079	0.8570 (0.0041)	0.9054 (0.0029)
		[22]	(5,5,5,12*0)	-0.0069 (0.0072)	0.8596 (0.0040)	0.9079 (0.0028)
	20	[23]	(10, 19*0)	-0.0063 (0.0054)	0.8729 (0.0037)	0.9186 (0.0025)
		[24]	(19*0,10)	-0.0069 (0.0058)	0.8705 (0.0038)	0.9182 (0.0025)
		[25]	(0,5,5,17*0)	-0.0065 (0.0055)	0.8717 (0.0037)	0.9168 (0.0025)
50	20	[26]	(30,19*0)	-0.0077 (0.0054)	0.8712 (0.0022)	0.9178 (0.0015)
		[27]	(19*0,30)	-0.0069 (0.0063)	0.8663 (0.0023)	0.9142 (0.0016)
	35	[28]	(15,34*0)	-0.0045 (0.0031)	0.8849 (0.0020)	0.9355 (0.0012)
		[29]	(34*0,15)	-0.0046 (0.0033)	0.8855 (0.0020)	0.9315 (0.0013)
		[30]	(5,5,5,32*0)	-0.0035 (0.0034)	0.8828 (0.0020)	0.9311 (0.0013)

Table 7: Simulated mean and estimated expectation of the approximate β- expectation tolerance interval for k=2

n	m	Scheme No.	Scheme	Simulated Mean			Estimated Expectation		
				90%	95%	99%	90%	95%	99%
5	2	[1]	(3,0)	0.7630	0.8175	0.8874	0.7916	0.8792	0.9685
		[2]	(0,3)	0.7612	0.8151	0.8843	0.7835	0.8738	0.9669
		[3]	(1,2)	0.7619	0.8163	0.8860	0.7851	0.8749	0.9672
		[4]	(2,1)	0.7628	0.8172	0.8869	0.7879	0.8767	0.9677
15	5	[5]	(10, 4*0)	0.8430	0.8975	0.9564	0.8584	0.9228	0.9817
		[6]	(4*0, 10)	0.8392	0.8935	0.9530	0.8518	0.9185	0.9804
		[7]	(2,2,2,2,2)	0.8407	0.8949	0.9542	0.8540	0.9199	0.9809
	10	[8]	(5,9*0)	0.8717	0.9247	0.9758	0.8798	0.9368	0.9860
		[9]	(9*0,5)	0.8700	0.9232	0.9748	0.8786	0.936	0.9857
		[10]	(3,2, 8*0)	0.8710	0.9242	0.9755	0.8797	0.9367	0.9860
20	10	[11]	(10, 9*0)	0.8716	0.9246	0.9757	0.8797	0.9367	0.9860
		[12]	(9*0,10)	0.8704	0.9232	0.9746	0.8774	0.9352	0.9855
25	10	[13]	(15,9*0)	0.8706	0.9240	0.9755	0.8796	0.9367	0.9859
		[14]	(9*0,15)	0.8668	0.9203	0.9729	0.8765	0.9347	0.9853
		[15]	(5,5,5,7*0)	0.8711	0.9240	0.9752	0.8791	0.9364	0.9859
	15	[16]	(10, 14*0)	0.8806	0.9330	0.9810	0.8865	0.9412	0.9873
		[17]	(14*0,10)	0.8787	0.9313	0.9801	0.8854	0.9405	0.9871
30	10	[18]	(20, 9*0)	0.8721	0.9248	0.9756	0.8796	0.9366	0.9859
		[19]	(9*0,20)	0.8676	0.9206	0.9729	0.8759	0.9342	0.9852
	15	[20]	(15, 14*0)	0.8803	0.9326	0.9807	0.8865	0.9412	0.9873
		[21]	(14*0,15)	0.8789	0.9313	0.9799	0.8849	0.9401	0.9870
		[22]	(5,5,5,12*0)	0.8811	0.9333	0.9811	0.8863	0.9411	0.9873
	20	[23]	(10, 19*0)	0.8862	0.9378	0.9836	0.8899	0.9434	0.9880
		[24]	(19*0,10)	0.8851	0.9370	0.9832	0.8893	0.9430	0.9879
		[25]	(0,5,5,17*0)	0.8855	0.9373	0.9834	0.8898	0.9434	0.9880
50	20	[26]	(30,19*0)	0.8855	0.9373	0.9834	0.8899	0.9434	0.9880
		[27]	(19*0,30)	0.8826	0.9348	0.9821	0.8882	0.9423	0.9877
	35	[28]	(15,34*0)	0.8920	0.9430	0.9865	0.8943	0.9462	0.9889
		[29]	(34*0,15)	0.8909	0.9422	0.9862	0.8939	0.9460	0.9888
		[30]	(5,5,5,32*0)	0.8914	0.9426	0.9863	0.8942	0.9462	0.9889

Table 8: Simulated mean and estimated expectation of the approximate β- expectation tolerance interval for k=3

n	m	Scheme No.	Scheme	Simulated Mean			Estimated Expectation		
				90%	95%	99%	90%	95%	99%
5	2	[1]	(3,0)	0.7696	0.8229	0.8900	0.7898	0.8772	0.9674
		[2]	(0,3)	0.7648	0.8184	0.8865	0.7807	0.8712	0.9655
		[3]	(1,2)	0.7713	0.8244	0.8915	0.7822	0.8722	0.9658
		[4]	(2,1)	0.7671	0.8209	0.8890	0.7849	0.8740	0.9664
15	5	[5]	(10, 4*0)	0.8423	0.8965	0.9550	0.8577	0.9221	0.9813
		[6]	(4*0, 10)	0.8386	0.8926	0.9516	0.8508	0.9175	0.9799
		[7]	(2,2,2,2,2)	0.8381	0.8923	0.9516	0.8528	0.9188	0.9803
	10	[8]	(5,9*0)	0.8724	0.9250	0.9755	0.8794	0.9364	0.9858
		[9]	(9*0,5)	0.8705	0.9232	0.9743	0.8778	0.9353	0.9854
		[10]	(3,2, 8*0)	0.8724	0.9250	0.9754	0.8793	0.9363	0.9858
20	10	[11]	(10, 9*0)	0.8718	0.9247	0.9754	0.8793	0.9364	0.9858
		[12]	(9*0,10)	0.8692	0.9221	0.9735	0.8766	0.9346	0.9852
25	10	[13]	(15,9*0)	0.8701	0.9231	0.9744	0.8793	0.9363	0.9857
		[14]	(9*0,15)	0.8687	0.9214	0.9730	0.8759	0.9341	0.9850
		[15]	(5,5,5,7*0)	0.8707	0.9235	0.9745	0.8788	0.936	0.9856
	15	[16]	(10, 14*0)	0.8803	0.9325	0.9803	0.8863	0.9409	0.9872
		[17]	(14*0,10)	0.8787	0.9310	0.9794	0.8849	0.9400	0.9869
30	10	[18]	(20, 9*0)	0.8711	0.9240	0.9749	0.8792	0.9363	0.9857
		[19]	(9*0,20)	0.8694	0.9218	0.9730	0.8753	0.9337	0.9849
	15	[20]	(15, 14*0)	0.8802	0.9324	0.9803	0.8863	0.9409	0.9872
		[21]	(14*0,15)	0.8784	0.9307	0.9792	0.8844	0.9397	0.9868
		[22]	(5,5,5,12*0)	0.8800	0.9323	0.9803	0.8861	0.9408	0.9871
	20	[23]	(10, 19*0)	0.8851	0.9369	0.9830	0.8898	0.9432	0.9879
		[24]	(19*0,10)	0.8852	0.9368	0.9828	0.8889	0.9427	0.9877
		[25]	(0,5,5,17*0)	0.8859	0.9375	0.9832	0.8897	0.9432	0.9879
50	20	[26]	(30,19*0)	0.8857	0.9373	0.9831	0.8897	0.9432	0.9879
		[27]	(19*0,30)	0.8835	0.9353	0.9820	0.8879	0.9420	0.9875
	35	[28]	(15,34*0)	0.8916	0.9426	0.9862	0.8942	0.9461	0.9888
		[29]	(34*0,15)	0.8917	0.9427	0.9862	0.8937	0.9458	0.9887
		[30]	(5,5,5,32*0)	0.8919	0.9428	0.9863	0.8941	0.9461	0.9888

Table 9: Simulated mean and estimated expectation of the approximate β- expectation tolerance interval for k=5

n	m	Scheme No.	Scheme	Simulated Mean			Estimated Expectation		
				90%	95%	99%	90%	95%	99%
5	2	[1]	(3,0)	0.7710	0.8250	0.8919	0.7884	0.8759	0.9665
		[2]	(0,3)	0.7688	0.8220	0.8888	0.7799	0.8702	0.9647
		[3]	(1,2)	0.7735	0.8267	0.8928	0.7811	0.8710	0.9650
		[4]	(2,1)	0.7706	0.824	0.8907	0.7834	0.8726	0.9655
15	5	[5]	(10, 4*0)	0.8459	0.8997	0.9568	0.8569	0.9214	0.9809
		[6]	(4*0, 10)	0.8414	0.8950	0.9528	0.8508	0.9173	0.9797
		[7]	(2,2,2,2,2)	0.8399	0.8939	0.9521	0.8523	0.9183	0.9800
	10	[8]	(5,9*0)	0.8722	0.9250	0.9754	0.8789	0.9360	0.9856
		[9]	(9*0,5)	0.8704	0.9231	0.9738	0.8772	0.9348	0.9852
		[10]	(3,2, 8*0)	0.8706	0.9235	0.9744	0.8789	0.9360	0.9856
20	10	[11]	(10, 9*0)	0.8715	0.9242	0.9747	0.8789	0.9360	0.9856
		[12]	(9*0,10)	0.8704	0.9231	0.9738	0.8762	0.9342	0.9850
25	10	[13]	(15,9*0)	0.8715	0.9243	0.9748	0.8788	0.9359	0.9855
		[14]	(9*0,15)	0.8689	0.9219	0.9732	0.8757	0.9339	0.9849
		[15]	(5,5,5,7*0)	0.8728	0.9253	0.9753	0.8784	0.9357	0.9855
	15	[16]	(10, 14*0)	0.8806	0.9328	0.9804	0.8860	0.9407	0.9871
		[17]	(14*0,10)	0.8792	0.9314	0.9795	0.8845	0.9397	0.9867
30	10	[18]	(20, 9*0)	0.8710	0.9238	0.9745	0.8788	0.9359	0.9855
		[19]	(9*0,20)	0.8699	0.9225	0.9734	0.8753	0.9336	0.9848
	15	[20]	(15, 14*0)	0.8808	0.9329	0.9804	0.8859	0.9407	0.9870
		[21]	(14*0,15)	0.8813	0.9330	0.9802	0.8841	0.9395	0.9867
		[22]	(5,5,5,12*0)	0.8810	0.9331	0.9806	0.8858	0.9406	0.9870
	20	[23]	(10, 19*0)	0.8858	0.9374	0.9831	0.8895	0.9430	0.9878
		[24]	(19*0,10)	0.8846	0.9363	0.9824	0.8886	0.9424	0.9876
		[25]	(0,5,5,17*0)	0.8857	0.9373	0.9831	0.8894	0.9430	0.9878
50	20	[26]	(30,19*0)	0.8843	0.9363	0.9826	0.8895	0.9430	0.9878
		[27]	(19*0,30)	0.8843	0.9360	0.9822	0.8878	0.9419	0.9874
	35	[28]	(15,34*0)	0.8911	0.9423	0.9860	0.8940	0.9460	0.9887
		[29]	(34*0,15)	0.8915	0.9425	0.9860	0.8935	0.9457	0.9886
		[30]	(5,5,5,32*0)	0.8924	0.9432	0.9864	0.8940	0.9460	0.9887

6 Real Data Application

Consider following real data which represents failure times, for a specific type of electrical insulation that was subjected to a continuously increasing voltage stress given by Lawless (2011).

12.3, 21.8, 24.4, 28.6, 43.2, 46.9, 70.7, 75.3, 95.5, 98.1, 138.6, 151.9.

According to Balakrishnan and Chan (1992), half-logistic distribution satisfactory fit to this data. We consider this data as outcome for lifetime for two unit series system. We use this data with three censoring schemes as (2,0,0,0), (0,0,0,2) and (1,1,0,0). We obtain reliability estimate for time period t=1. MLE of reliability estimate and its MSE is given in Table 10. We construct confidence interval based on MLE. These 90% and 95% confidence intervals and their lengths are presented in same Table.

Table 10: Bias, MSE^{+}, Confidence intervals and its length for R(t)

n	m	Scheme	Bias and MSE	90% C. I. and its length	95% C. I. and its length
6	4	(2,0,0,0)	-0.0084	(0.9689, 0.9970)	(0.9665, 0.9970)
			(0.000002)	0.0281	0.0305
		(0,0,0,2)	-0.0011	(0.9811, 0.9969)	(0.9796, 0.9984)
			(0.000075)	0.0158	0.0188
		(1,1,0,0)	-0.0049	(0.9747, 0.9957)	(0.9737, 0.9977)
			(0.00024)	0.021	0.024

Method of MLE using EM algorithm and confidence interval based on MLE of reliability function gives best performance for real data. Bias is small in case of conventional censoring scheme whereas MSE is small in case of progressive censoring scheme. Length of confidence interval is small in case of conventional censoring scheme.

7 Conclusion and Discussion

Simulation study results indicate that, the bias, MSE of the MLE and reliability estimate decrease with increase in sample size n and increase in the effective sample size m. Same trend is observed in case of SE of confidence level of confidence intervals. The MSE is relatively smaller for progressive Type-II censoring scheme as compared with conventional Type-II censoring scheme. Confidence levels of confidence interval using log-transformed MLE are better than the confidence levels of confidence interval using MLE. SE for confidence levels of confidence intervals using log-transformed MLE is

smaller than SE for confidence levels of confidence intervals using MLE. Confidence levels of confidence intervals of reliability function are better for large sample size.

β-expectation tolerance interval shows good results. As sample size n and effective sample size m increases the estimated expectation and simulated mean approaches to nominal coverage. Estimated expectation and simulated mean have better coverage for progressive Type-II censoring scheme than conventional Type-II censoring scheme, for small sample size. As number of units in system (k) increases the simulated mean decreases, but estimated expectation increases.

EM algorithm method works well for small sample size and for smaller effective sample size. Overall both conventional Type-II censoring scheme and progressive Type-II censoring scheme give better results. The MSE of progressive Type-II censoring method is smaller than the MSE of conventional censoring method, while bias, confidence interval and β-expectation tolerance interval perform equally good for both the methods. The results reported in this paper can also be applied when k is replaced by any known positive real number.

Aknowledgment

The authors are grateful to the editor and the learned referee for making constructive and valuable comments which have significantly improved the contents of this article.

References

Asgharzadeh, A. and Valiollahi, R. (2011). Estimation of the scale parameter of the lomax distribution under progressive censoring. *International Journal of Statistics & Economics*, 6(S11):37–48.

Atwood, C. L. (1984). Approximate tolerance intervals, based on maximum likelihood estimates. *Journal of the American Statistical Association*, 79(386):459–465.

Balakrishnan, N. (2007). Progressive censoring methodology: an appraisal. *Test*, 16(2):211–259.

Balakrishnan, N. and Aggarwala, R. (2000). *Progressive censoring: theory, methods, and applications*. Springer.

Balakrishnan, N. and Asgharzadeh, A. (2005). Inference for the scaled half-logistic distribution based on progressively type-ii censored samples. *Communications in Statistics—Theory and Methods*, 34(1):73–87.

Balakrishnan, N. and Chan, P. (1992). Estimation for the scaled half logistic distribution under type ii censoring. *Computational statistics & data analysis*, 13(2):123–141.

Balakrishnan, N., Kannan, N., Lin, C., and Wu, S. (2004). Inference for the extreme value distribution under progressive type-ii censoring. *Journal of Statistical Computation and Simulation*, 74(1):25–45.

Balakrishnan, N., Kannan, N., Lin, C.-T., and Ng, H. T. (2003). Point and interval

estimation for gaussian distribution, based on progressively type-ii censored samples. *Reliability, IEEE Transactions on*, 52(1):90–95.

Balakrishnan, N. and Sandhu, R. (1995). A simple simulational algorithm for generating progressive type-ii censored samples. *The American Statistician*, 49(2):229–230.

Cohen, A. C. (1963). Progressively censored samples in life testing. *Technometrics*, 5(3):327–339.

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 (Methodological)*, pages 1–38.

Iliopoulos, G. and Balakrishnan, N. (2011). Exact likelihood inference for laplace distribution based on type-ii censored samples. *Journal of Statistical Planning and Inference*, 141(3):1224–1239.

Kim, C. and Han, K. (2010). Estimation of the scale parameter of the half-logistic distribution under progressively type ii censored sample. *Statistical Papers*, 51(2):375–387.

Krishna, H. and Kumar, K. (2011). Reliability estimation in lindley distribution with progressively type ii right censored sample. *Mathematics and Computers in Simulation*, 82(2):281–294.

Krishna, H. and Kumar, K. (2013). Reliability estimation in generalized inverted exponential distribution with progressively type ii censored sample. *Journal of Statistical Computation and Simulation*, 83(6):1007–1019.

Krishna, H. and Malik, M. (2012). Reliability estimation in maxwell distribution with progressively type-ii censored data. *Journal of Statistical Computation and Simulation*, 82(4):623–641.

Kumbhar, R. and Shirke, D. (2004). Tolerance limits for lifetime distribution of k-unit parallel system. *Journal of Statistical Computation and Simulation*, 74(3):201–213.

Lawless, J. F. (2011). *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons.

Little, R. J. and Rubin, D. B. (2002). Statistical analysis with missing data.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233.

Mann, N. R. (1969). Exact three-order-statistic confidence bounds on reliable life for a weibull model with progressive censoring. *Journal of the American Statistical Association*, 64(325):306–315.

Mann, N. R. (1971). Best linear invariant estimation for weibull parameters under progressive censoring. *Technometrics*, 13(3):521–533.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical methods for reliability data*, volume 314. John Wiley & Sons.

Ng, H. (2005). Parameter estimation for a modified weibull distribution, for progressively type-ii censored samples. *Reliability, IEEE Transactions on*, 54(3):374–380.

Ng, H., Chan, P., and Balakrishnan, N. (2002). Estimation of parameters from progressively censored data using em algorithm. *Computational Statistics & Data Analysis*, 39(4):371–386.

Potdar, K. and Shirke, D. (2012). Inference for the distribution of a k-unit parallel system with exponential distribution as the component life distribution based on type-ii progressively censored sample. *Int. J. Agricult. Stat. Sci*, 8(2):503–517.

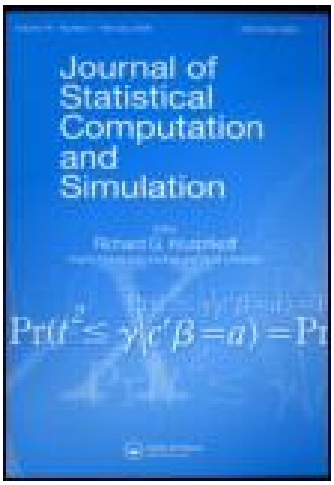
Potdar, K. and Shirke, D. (2013a). Inference for the parameters of generalized inverted family of distributions. In *ProdStat Forum*, volume 6, pages 18–28.

Potdar, K. and Shirke, D. (2013b). Reliability estimation for the distribution of a k-unit parallel system with rayleigh distribution as the component life distribution. In *International Journal of Engineering Research and Technology*, volume 2. ESRSA Publications.

Potdar, K. and Shirke, D. (2014). Inference for the scale parameter of lifetime distribution of k-unit parallel system based on progressively censored data. *Journal of Statistical Computation and Simulation*, 84(1):171–185.

Pradhan, B. (2007). Point and interval estimation for the lifetime distribution of a k-unit parallel system based on progressively type-ii censored data. *Economic Quality Control*, 22(2):175–186.

Pradhan, B. and Kundu, D. (2009). On progressively censored generalized exponential distribution. *Test*, 18(3):497–515.



Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Inference for the scale parameter of lifetime distribution of k-unit parallel system based on progressively censored data

K. G. Potdar[a] & D. T. Shirke[b]

[a] Department of Statistics, Ajara Mahavidyalaya, Ajara, Kolhapur, Maharashtra 416505, India

[b] Department of Statistics, Shivaji University, Kolhapur 416004, India

Published online: 06 Jul 2012.

To cite this article: K. G. Potdar & D. T. Shirke (2014) Inference for the scale parameter of lifetime distribution of k-unit parallel system based on progressively censored data, Journal of Statistical Computation and Simulation, 84:1, 171-185, DOI: 10.1080/00949655.2012.700314

To link to this article: http://dx.doi.org/10.1080/00949655.2012.700314

Journal of Statistical Computation and Simulation, 2014
Vol. 84, No. 1, 171–185, http://dx.doi.org/10.1080/00949655.2012.700314

Taylor & Francis
Taylor & Francis Group

Inference for the scale parameter of lifetime distribution of *k*-unit parallel system based on progressively censored data

K.G. Potdar[a]* and D.T. Shirke[b]

[a]*Department of Statistics, Ajara Mahavidyalaya, Ajara, Kolhapur, Maharashtra 416505, India;*
[b]*Department of Statistics, Shivaji University, Kolhapur 416004, India*

(*Received 10 February 2012; final version received 1 June 2012*)

In this paper, inference for the scale parameter of lifetime distribution of a k-unit parallel system is provided. Lifetime distribution of each unit of the system is assumed to be a member of a scale family of distributions. Maximum likelihood estimator (MLE) and confidence intervals for the scale parameter based on progressively Type-II censored sample are obtained. A β-expectation tolerance interval for the lifetime of the system is obtained. As a member of the scale family, half-logistic distribution is considered and the performance of the MLE, confidence intervals and tolerance intervals are studied using simulation.

Keywords: progressively Type-II censoring; EM algorithm; MLE; confidence interval; coverage probability; β-expectation tolerance interval; half-logistic distribution

Mathematics Subject Classifications: 62N02; 62F10; 62F25

1. Introduction

In life testing experiments, certain units are put on test and we observe failure time for each of these units. Sometimes it is impossible to observe failure times of all the units or we have to terminate the experiment at some specified time. In such cases, failure times for some of the units may not be observed. The unobserved failure time data are called censored data. Broadly, censoring is classified into two types: Type-I and Type-II censoring. Type-I censoring depends on time. An experiment continues up to a pre-determined time T. Units having failure time after time T are not observed. Here, failure time will be known only if it is exactly less than T. For example, if 'n' units are placed on test and the test is terminated at time T, the failure times will be known only for those units that fail before time T. In Type-I censoring, the number of exact failure times observed is random.

Type-II censoring scheme is often used in life testing experiment. Only m units in a random sample of size n ($m < n$) are observed. Progressive Type-II censoring is a generalization of Type-II censoring. In progressive censoring scheme, the number 'm' and $R_1, R_2, \ldots, R_m$ are fixed prior to the test and $\sum_{i=1}^{m} R_i = n - m$. At the first failure, R_1 units are randomly removed from remaining

*Corresponding author. Email: potdarkiran.stat@gmail.com

$n-1$ units. At the second failure, R_2 units are randomly removed from remaining $n-2-R_1$ units and so on. At the mth failure, all remaining R_m units are removed. Here, we observe failure time of 'm' units and remaining $n-m$ units are removed at different stages of an experiment. In conventional Type-II censoring scheme $R_1 = R_2 = \cdots R_{m-1} = 0$ and $R_m = n-m$. In this paper, the progressive Type-II censoring scheme is considered.

Many authors studied progressive Type-II censoring scheme for various lifetime distributions. Cohen [1] introduced progressive Type-II censoring. Mann [2,3] considered the Weibull distribution with progressive censoring. Balakrishnan *et al.* [4–6] discussed inference for half-logistic, Gaussian and extreme value distribution under progressive Type-II censoring scheme, respectively. Ng [7] studied parameter estimation for the modified Weibull distribution under progressively Type-II censoring.

Balakrishnan and Aggarwala [8] described details about progressive censoring. Balakrishnan [9] studied various distributions and inferential methods for progressively censored data. Pradhan [10] considered point and interval estimation of a k-unit parallel system based on progressive Type-II censoring scheme with exponential distribution as the lifetime distribution of each unit. Kim and Han [11] discussed half-logistic distribution for Type-II progressively censored sample. Recently Iliopoulos and Balakrishnan [12] studied likelihood inference for Laplace distribution based on progressively Type-II censored sample.

Dempster *et al*. [13] introduced the expectation maximization (EM) algorithm. They presented maximum likelihood estimation for incomplete data. Mclachlan and Krishnan [14] introduced more details about the EM algorithm. Little and Rubin [15] discussed EM algorithm for exponential family of distributions. Pradhan and Kundu [16] used the EM algorithm to estimate parameters of generalized exponential distribution under progressive Type-II censoring scheme. Ng *et al*. [17] used the EM algorithm to estimate parameters of lognormal and Weibull distributions under the Type-II censoring scheme. In this paper, the EM algorithm is used for the estimation of the parameters of a k-unit parallel system based on the progressive Type-II censoring scheme when lifetime distribution of each unit belongs to the scale family.

Parameter estimation is based on the lifetimes of the system. We assume that n units are put on test and failure times of $\sum_{i=1}^{m} R_i = n-m$ units are censored. Failure times of these censored units are unknown. These data are considered as missing and the EM algorithm is used to compute the maximum likelihood estimator (MLE). We used idea of missing information principle of Louis [18]. Asymptotic normal distribution of the MLE is used to construct confidence interval for the scale parameter. We also discussed tolerance interval for the lifetime of system, on the lines of Kumbhar and Shirke [19].

In Section 2, we introduced the model and obtained the MLE for the scale parameter. We also provided an expression for Fisher information. Asymptotic confidence interval for the scale parameter is discussed in the same section. Section 3 provides β-expectation tolerance interval for the lifetime of a k-unit parallel system based on progressively censored data. In Section 4, the half-logistic distribution is considered as a member of the scale family. The MLE, confidence intervals and the tolerance intervals are studied. The performance of the MLE and confidence intervals for the scale parameter of half-logistic distribution is investigated using simulations. Results of the simulation study have been reported in Section 5. Conclusions are presented in Section 6.

2. Model and estimation of the scale parameter

Let $\mathbb{G}_\lambda$ be a scale family of lifetime distributions, where λ is the parameter of interest. Consider k-unit parallel system with independently and identically distributed units having lifetimes $X_1, X_2, \ldots, X_k$ That is X_i is the lifetime of the ith unit having cumulative density function (cdf)

$G(x_i/\lambda)$. Lifetime of system is $X = \text{Max.}(X_1, X_2, \ldots, X_k)$. The cdf of X is

$$F(x;\lambda) = \left[G\left(\frac{x}{\lambda}\right)\right]^k, \quad \lambda > 0, \ x \geq 0.$$

The probability density function (pdf) of X is

$$f(x;\lambda) = \frac{k}{\lambda} g\left(\frac{x}{\lambda}\right)\left[G\left(\frac{x}{\lambda}\right)\right]^{k-1}, \quad \lambda > 0, \ x \geq 0.$$

where $g(\cdot)$ is the pdf of X_i when $\lambda = 1$.

2.1. *Maximum likelihood estimation*

Suppose n k-unit parallel systems are under test and we observe failure times of m systems under progressive Type-II censoring. Let $(R_1, R_2, \ldots, R_m)$ be a progressive censoring scheme.

The likelihood function for the observed data is

$$L(\lambda) = C\prod_{i=1}^{m} f(x_{(i)};\lambda)[1 - F(x_{(i)};\lambda)]^{R_i},$$

$$\text{where } C = n\prod_{j=1}^{m-1}\left(n - j - \sum_{i=1}^{j} R_i\right).$$

$$L(\lambda) = C\prod_{i=1}^{m} \frac{k}{\lambda} g\left(\frac{x_{(i)}}{\lambda}\right)\left[G\left(\frac{x_{(i)}}{\lambda}\right)\right]^{k-1}\left\{1 - \left[G\left(\frac{x_{(i)}}{\lambda}\right)\right]^k\right\}^{R_i}.$$

Suppose $x_{(1)}, x_{(2)}, \ldots, x_{(m)}$ is the observed data and $z_1, z_2, \ldots, z_m$ is the censored data. We note that z_i is a vector with R_i elements, which is not observable for $i = 1, 2, \ldots, m$. The censored data $Z = (z_1, z_2, \ldots, z_m)$ can be considered as the missing data. $X = (x_{(1)}, x_{(2)}, \ldots, x_{(m)})$ is the observed data. $W = (X, Z)$ is the complete data set. Then complete log-likelihood function is

$$L_c = n\log(k) - n\log(\lambda) + \sum_{i=1}^{m}\log\left[g\left(\frac{x_i}{\lambda}\right)\right] + (k-1)\sum_{i=1}^{m}\log\left[G\left(\frac{x_i}{\lambda}\right)\right]$$
$$+ \sum_{i=1}^{m}\sum_{j=1}^{R_i}\log\left[g\left(\frac{z_{ij}}{\lambda}\right)\right] + (k-1)\sum_{i=1}^{m}\sum_{j=1}^{R_i}\log\left[G\left(\frac{z_{ij}}{\lambda}\right)\right]. \tag{1}$$

In order to obtain the MLE of λ, we use the EM algorithm [13]. For the E step in the EM algorithm, we take expectation of Z_{ij}. The derivative of L_c with respect to λ is taken for the M step, where

$$\frac{\mathrm{d}L_c}{\mathrm{d}\lambda} = -\frac{n}{\lambda} - \frac{1}{\lambda^2}\sum_{i=1}^{m}\frac{x_i g'(x_i/\lambda)}{g(x_i/\lambda)} - \frac{(k-1)}{\lambda^2}\sum_{i=1}^{m}\frac{x_i G'(x_i/\lambda)}{G(x_i/\lambda)} - \frac{1}{\lambda^2}\sum_{i=1}^{m} R_i a(x_i, k, \lambda^0)$$
$$- \frac{(k-1)}{\lambda^2}\sum_{i=1}^{m} R_i b(x_i, k, \lambda^0), \tag{2}$$

$$\text{where } a(x_i, k, \lambda^0) = E\left(\frac{Z_{ij} g'(Z_{ij}/\lambda)}{g(Z_{ij}/\lambda)}\right) = \int_{x_i}^{\infty} \frac{z g'\left(\frac{z}{\lambda}\right)}{g\left(\frac{z}{\lambda}\right)} \frac{f(z;\lambda)}{1 - F(x_i;\lambda)}\,\mathrm{d}z$$

$$\text{and } b(x_i, k, \lambda^0) = E\left(\frac{Z_{ij} G'(Z_{ij}/\lambda)}{G(Z_{ij}/\lambda)}\right) = \int_{x_i}^{\infty} \frac{z G'\left(\frac{z}{\lambda}\right)}{G\left(\frac{z}{\lambda}\right)} \frac{f(z;\lambda)}{1 - F(x_i;\lambda)}\,\mathrm{d}z.$$

To obtain the solution λ^1, we have to solve the equation $dL_c/d\lambda = 0$. But the closed form solution does not exist for this equation. Therefore, we used the Newton–Raphson method and computed λ^1. Using this λ^1, we computed $a(x_i, k, \lambda^1)$ and $b(x_i, k, \lambda^1)$. This ended the M-step. We continued this procedure until convergence took place.

In the Newton–Raphson method, we have to choose the initial value of λ. Here, we used least-square estimate of λ as an initial value. Ng [7] discussed estimation of model parameters of the modified Weibull distribution based on progressively Type-II censored data where the empirical distribution function is computed as [20]

$$\hat{F}(x_{(i)}) = 1 - \prod_{j=1}^{i}(1 - \hat{p}_j).$$

$$\text{with} \quad \hat{p}_j = \frac{1}{n - \sum_{k=2}^{j} R_{k-1} - j + 1} \quad \text{for } j = 1, 2, \ldots, m.$$

The estimate of the parameters can be obtained by the least-square fit of simple linear regression

$$y_i = \beta x_{(i)} \quad \text{with } \beta = -\frac{1}{\lambda}.$$

$$y_i = \ln\left[1 - \frac{\hat{F}^{\frac{1}{k}}(x_{(i-1)}) + \hat{F}^{\frac{1}{k}}(x_i)}{2}\right] \quad \text{for } i = 1, 2, \ldots, m.$$

$$\hat{F}(x_{(0)}) = 0.$$

The least-square estimate of λ is given by

$$\hat{\lambda} = -\frac{\sum_{i=1}^{m} x_{(i)}^2}{\sum_{i=1}^{m} x_{(i)} y_i}.$$

Using this $\hat{\lambda}$, we obtained the MLE of λ by the Newton–Raphson method.

2.2. *Fisher information*

According to Louis [18], the observed Fisher information is given by

observed information = complete information – missing information. That is $I_x(\lambda) = I_w(\lambda) - I_{w|x}(\lambda)$, where

complete information = $I_w(\lambda) = -E[d^2L/d\lambda^2]$ and

L is the log-likelihood function based on all n observations. We obtain $I_w(\lambda)$ and $I_{w|x}(\lambda)$ in the following.

Now,

$$L = n\log(k) - n\log(\lambda) + \sum_{i=1}^{n} \log\left[g\left(\frac{x_i}{\lambda}\right)\right] + (k-1)\sum_{i=1}^{n} \log\left[G\left(\frac{x_i}{\lambda}\right)\right] \tag{3}$$

and

$$\frac{dL}{d\lambda} = -\frac{n}{\lambda} - \frac{1}{\lambda^2}\sum_{i=1}^{n} \frac{x_i g'(x_i/\lambda)}{g(x_i/\lambda)} - \frac{(k-1)}{\lambda^2}\sum_{i=1}^{n} \frac{x_i G'(x_i/\lambda)}{G(x_i/\lambda)}.$$

$$\frac{\mathrm{d}^2L}{\mathrm{d}\lambda^2} = \frac{n}{\lambda^2} + \frac{1}{\lambda^4}\sum_{i=1}^{n}\frac{x_i^2 g(x_i/\lambda)g''(x_i/\lambda) - x_i^2[g'(x_i/\lambda)]^2 + 2\lambda x_i g(x_i/\lambda)g'(x_i/\lambda)}{[g(x_i/\lambda)]^2}$$
$$+ \frac{(k-1)}{\lambda^4}\sum_{i=1}^{n}\frac{x_i^2 G(x_i/\lambda)G''(x_i/\lambda) - x_i^2[G'(x_i/\lambda)]^2 + 2\lambda x_i G(x_i/\lambda)G'(x_i/\lambda)}{[G(x_i/\lambda)]^2}.$$

Complete information is given by

$$I_w(\lambda) = -\frac{n}{\lambda^2} - \frac{1}{\lambda^4}\sum_{i=1}^{n}\mathrm{E}\left[\frac{X_i^2 g(X_i/\lambda)g''(X_i/\lambda) - X_i^2[g'(X_i/\lambda)]^2 + 2\lambda X_i g(X_i/\lambda)g'(X_i/\lambda)}{\left[g\left(\frac{X_i}{\lambda}\right)\right]^2}\right]$$
$$- \frac{(k-1)}{\lambda^4}\sum_{i=1}^{n}\mathrm{E}\left[\frac{X_i^2 G(X_i/\lambda)G''(X_i/\lambda) - X_i^2[G'(X_i/\lambda)]^2 + 2\lambda X_i G(X_i/\lambda)G'(X_i/\lambda)}{[G(X_i/\lambda)]^2}\right]. \quad (4)$$

Missing information is given by

$$I_{W|X}(\lambda) = \sum_{i=1}^{m} R_i I_{W|X}^{(i)}(\lambda) = -\sum_{i=1}^{m}\sum_{j=1}^{R_i} E_{Z|X}\left[\frac{\mathrm{d}^2\log(f(Z_{ij}|X_i,\lambda))}{\mathrm{d}\lambda^2}\right].$$

Consider

$$f_{z|X}(Z_{ij}|X_i,\lambda) = \frac{f(z_{ij};\lambda)}{1-F(x_i;\lambda)} = \frac{(k/\lambda)g\left(z_{ij}/\lambda\right)\left[G(z_{ij}/\lambda)\right]^{k-1}}{1-[G(x_i/\lambda)]^k}.$$

Therefore,

$$\log f = \log k - \log\lambda + \log\left[g\left(\frac{z_{ij}}{\lambda}\right)\right] + (k-1)\log\left[G\left(\frac{z_{ij}}{\lambda}\right)\right] - \log\left\{1-\left[G\left(\frac{x_i}{\lambda}\right)\right]^k\right\}.$$
$$\frac{\mathrm{d}\log f}{\mathrm{d}\lambda} = -\frac{1}{\lambda} - \frac{z_{ij}g'\left(\frac{z_{ij}}{\lambda}\right)}{\lambda^2 g\left(\frac{z_{ij}}{\lambda}\right)} - \frac{(k-1)z_{ij}G'(z_{ij}/\lambda)}{\lambda^2 G(z_{ij}/\lambda)} - \frac{kx_i[G(x_i/\lambda)]^{k-1}G'(x_i/\lambda)}{\lambda^2\{1-[G(x_i/\lambda)]^k\}}$$

and

$$\frac{\mathrm{d}^2\log f}{\mathrm{d}\lambda^2} = \frac{1}{\lambda^2} + \frac{z_{ij}^2 g(z_{ij}/\lambda)g''(z_{ij}/\lambda) - z_{ij}^2[g'(z_{ij}/\lambda)]^2 + 2\lambda z_{i,j} g(z_{ij}/\lambda)g'(z_{ij}/\lambda)}{\lambda^4[g(z_{ij}/\lambda)]^2}$$
$$+ \frac{(k-1)\{z_{ij}^2 G(z_{ij}/\lambda)G''(z_{ij}/\lambda) - z_{ij}^2[G'(z_{ij}/\lambda)]^2 + 2\lambda z_{ij}G(z_{ij}/\lambda)G'(z_{ij}/\lambda)\}}{\lambda^4[G(z_{ij}/\lambda)]^2}$$
$$+ \frac{kx_i^2\{1-[G(x_i/\lambda)]^k\}[G(x_i/\lambda)]^{k-2}\{G(x_i/\lambda)G''(x_i/\lambda) + (k-1)[G'(x_i/\lambda)]^2\}}{\lambda^4\{1-[G(x_i/\lambda)]^k\}^2}$$
$$+ \frac{k^2x_i^2[G(x_i/\lambda)]^{2k-2}[G'(x_i/\lambda)]^2 + 2\lambda kx_i[G(x_i/\lambda)]^{k-1}G'(x_i/\lambda)\{1-[G(x_i/\lambda)]^k\}}{\lambda^4\{1-[G(x_i/\lambda)]^k\}^2}.$$

Hence, missing information is

$$I_{W|X}(\lambda) = \sum_{i=1}^{m} R_i I^{(i)}_{W|X}(\lambda) = -\sum_{i=1}^{m}\sum_{j=1}^{R_i} E_{Z|X}\left[\frac{d^2 \log(f(Z_{ij}|X_i,\lambda))}{d\lambda^2}\right].$$

$$= -\frac{n-m}{\lambda^2}$$

$$- \frac{1}{\lambda^4}\sum_{i=1}^{m}\sum_{j=1}^{R_i}\left\{E\left[\frac{Z_{ij}^2 g(Z_{ij}/\lambda)g''(Z_{ij}/\lambda) - Z_{ij}^2[g'(Z_{ij}/\lambda)]^2 + 2\lambda Z_{ij} g(Z_{ij}/\lambda)g'(Z_{ij}/\lambda)}{[g(Z_{ij}/\lambda)]^2}\right]\right.$$

$$+ (k-1)E\left[\frac{Z_{ij}^2 G(Z_{ij}/\lambda)G''(Z_{ij}/\lambda) - z_{ij}^2[G'(Z_{ij}/\lambda)]^2 + 2\lambda Z_{ij} G(Z_{ij}/\lambda)G'(Z_{ij}/\lambda)}{[G(Z_{ij}/\lambda)]^2}\right]$$

$$+ \frac{kx_i^2\{1-[G(x_i/\lambda)]^k\}[G(x_i/\lambda)]^{k-2}\{G(x_i/\lambda)G''(x_i/\lambda) + (k-1)[G'(x_i/\lambda)]^2\}}{\{1-[G(x_i/\lambda)]^k\}^2}$$

$$\left. + \frac{k^2 x_i^2[G(x_i/\lambda)]^{2k-2}[G'(x_i/\lambda)]^2 + 2\lambda k x_i[G(x_i/\lambda)]^{k-1}G'(x_i/\lambda)\{1-[G(x_i/\lambda)]^k\}}{\{1-[G(x_i/\lambda)]^k\}^2}\right\}. \tag{5}$$

By using expressions in Equations (4) and (5), we obtained observed Fisher information.

2.3. *Confidence intervals*

Using asymptotic normal distribution of the MLE, confidence interval for λ is constructed. Let $\hat{\lambda}_n$ be the MLE of λ and $\hat{\sigma}^2(\hat{\lambda}_n) = 1/l_x(\hat{\lambda}_n)$ be the estimated variance of $\hat{\lambda}_n$. Therefore, $100(1-\alpha)\%$ asymptotic confidence interval for λ is given by

$$\left(\hat{\lambda}_n - \tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\hat{\lambda}_n)},\quad \hat{\lambda}_n + \tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\hat{\lambda}_n)}\right). \tag{6}$$

where $\tau_{\alpha/2}$ is the upper $100(\alpha/2)$th percentile of the standard normal distribution.

Meeker and Escobar [20] reported that the asymptotic confidence interval for λ can be computed using $\log(\hat{\lambda}_n)$. An approximate $100(1-\alpha)\%$ confidence interval for $\log(\lambda)$ is $\left(\log(\hat{\lambda}_n) - \tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\log(\hat{\lambda}_n))}, \log(\hat{\lambda}_n) + \tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\log(\hat{\lambda}_n))}\right)$, where $\hat{\sigma}^2(\log(\hat{\lambda}_n))$ is the estimated variance of $\log(\hat{\lambda}_n)$, which is approximated by $\hat{\sigma}^2(\log(\hat{\lambda}_n)) \approx \hat{\sigma}^2(\hat{\lambda}_n)/\hat{\lambda}_n^2$ Hence, an approximate $100(1-\alpha)\%$ confidence interval for λ is

$$\left(\hat{\lambda}_n e^{\left(-\frac{\tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\hat{\lambda}_n)}}{\hat{\lambda}_n}\right)},\quad \hat{\lambda}_n e^{\left(\frac{\tau_{\alpha/2}\sqrt{\hat{\sigma}^2(\hat{\lambda}_n)}}{\hat{\lambda}_n}\right)}\right). \tag{7}$$

3. Tolerance intervals

Kumbhar and Shirke [19] derived the expression for β-expectation tolerance interval for the lifetime distribution of a k-unit parallel system when the lifetime distribution of each unit is exponential. They investigated the performance of the tolerance interval based on complete data. Pradhan [10] studied the performance of the tolerance interval based on progressively Type-II censored data from the exponential distribution. The performance of the tolerance interval based

on progressively Type-II censored data for the scale family of distributions is studied. Let $l_\beta(\lambda)$ be the lower quantile of order β of the distribution function $F(x;\lambda)$. Then, we have

$$l_\beta(\lambda) = \lambda G^{-1}(\beta^{1/k}).$$

Thus, an upper β-expectation tolerance interval for $F(x;\lambda)$ is obtained by

$$I_\beta = (0, l_\beta(\lambda)).$$

The maximum likelihood estimate of $l_\beta(\lambda)$ is given by

$$l_\beta(\hat{\lambda}_n) = \hat{\lambda}_n G^{-1}(\beta^{1/k})$$

yielding an approximate β-expectation tolerance interval

$$\hat{I}_\beta = (0, l_\beta(\hat{\lambda}_n)).$$

The expectation of $\hat{I}_\beta$ can be obtained approximately using the approach suggested by Atwood [21] and is given as

$$E[F(l_\beta(\hat{\lambda}_n);\lambda)] \approx \beta - 0.5F_{02}\sigma^2(\hat{\lambda}_n) + \frac{F_{01}\sigma^2(\hat{\lambda}_n)F_{11}}{F_{10}}, \tag{8}$$

where $F_{10} = \dfrac{\partial F}{\partial x}$, $F_{01} = \dfrac{\partial F}{\partial \lambda}$, $F_{11} = \dfrac{\partial^2 F}{\partial x \partial \lambda}$, $F_{02} = \dfrac{\partial^2 F}{\partial \lambda^2}$,

$$F_{10} = \frac{k}{\lambda}\left[G\left(\frac{x}{\lambda}\right)\right]^{k-1} g\left(\frac{x}{\lambda}\right),$$

$$F_{01} = -\frac{kx}{\lambda^2}\left[G\left(\frac{x}{\lambda}\right)\right]^{k-1} G'\left(\frac{x}{\lambda}\right),$$

$$F_{11} = -\frac{k}{\lambda^3}\left[G\left(\frac{x}{\lambda}\right)\right]^{k-2}\left\{xG\left(\frac{x}{\lambda}\right)g'\left(\frac{x}{\lambda}\right) + x(k-1)G'\left(\frac{x}{\lambda}\right)g\left(\frac{x}{\lambda}\right) + \lambda G\left(\frac{x}{\lambda}\right)g\left(\frac{x}{\lambda}\right)\right\},$$

and $$F_{02} = \frac{kx}{\lambda^4}\left[G\left(\frac{x}{\lambda}\right)\right]^{k-2}\left\{xG\left(\frac{x}{\lambda}\right)G''\left(\frac{x}{\lambda}\right) + x(k-1)\left[G'\left(\frac{x}{\lambda}\right)\right]^2 + 2\lambda G\left(\frac{x}{\lambda}\right)G'\left(\frac{x}{\lambda}\right)\right\}.$$

The derivatives of F are evaluated at $x = l_\beta(\lambda)$ with $\lambda = \hat{\lambda}_n$. Instead of the actual value of $\sigma^2(\hat{\lambda})$, its estimate has been used.

4. Application to half-logistic distribution

Consider a member of the scale family of distributions, namely half-logistic distribution with scale parameter λ. The cdf of X is

$$F(x;\lambda) = \left[\frac{1-\mathrm{e}^{-x/\lambda}}{1+\mathrm{e}^{-x/\lambda}}\right]^k, \quad \lambda > 0, \ x \geq 0.$$

The pdf of X is

$$f(x;\lambda) = \frac{k}{\lambda}\frac{2\mathrm{e}^{-x/\lambda}}{(1+\mathrm{e}^{-x/\lambda})^2}\left[\frac{1-\mathrm{e}^{-x/\lambda}}{1+\mathrm{e}^{-x/\lambda}}\right]^{k-1}, \quad \lambda > 0, \ x \geq 0.$$

4.1. *Maximum likelihood estimation*

The complete log-likelihood function for the half-logistic distribution with scale parameter λ from Equation (1) is

$$L_c = n\log(k) - n\log(\lambda) + \sum_{i=1}^{m}\log\left(\frac{2e^{-x_i/\lambda}}{(1+e^{-x_i/\lambda})^2}\right) + (k-1)\sum_{i=1}^{m}\log\left[\frac{1-e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}\right] + \sum_{i=1}^{m}\sum_{j=1}^{R_i}\log\left(\frac{2e^{-\frac{z_{ij}}{\lambda}}}{(1+e^{-\frac{z_{ij}}{\lambda}})^2}\right) + (k-1)\sum_{i=1}^{m}\sum_{j=1}^{R_i}\log\left(\frac{1-e^{-z_{ij}/\lambda}}{1+e^{-z_{ij}/\lambda}}\right). \quad (9)$$

In order to obtain the MLE of λ, we use the EM algorithm [13]. For the E step in the EM algorithm, we take expectation of Z_{ij}. The derivative of L_c with respect to λ is taken for the M step, where

$$\frac{dL_c}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{m}\frac{x_i(1-e^{-\frac{x_i}{\lambda}})}{1+e^{-\frac{x_i}{\lambda}}} - \frac{2(k-1)}{\lambda^2}\sum_{i=1}^{m}\frac{x_i e^{-\frac{x_i}{\lambda}}}{1-e^{-\frac{2x_i}{\lambda}}} - \frac{1}{\lambda^2}\sum_{i=1}^{m}R_i a(x_i,k,\lambda^0) - \frac{2(k-1)}{\lambda^2}\sum_{i=1}^{m}R_i b(x_i,k,\lambda^0), \quad (10)$$

where $a(x_i,k,\lambda^0) = E\left(\frac{Z_{ij}(1-e^{-Z_{ij}/\lambda})}{1+e^{-Z_{ij}/\lambda}}\right)$

and $b(x_i,k,\lambda^0) = E(Z_{ij}e^{-Z_{ij}/\lambda}/(1-e^{-2Z_{ij}/\lambda}))$.

To solve this equation, we use the Newton–Raphson method.

4.2. *Fisher information*

The observed information = complete information – missing information. That is $I_x(\lambda) = I_w(\lambda) - I_{w|x}(\lambda)$.

Consider $I_w(\lambda) = -E[d^2L/d\lambda^2]$.

Log-likelihood function for n observations is

$$L = n\log(k) - n\log(\lambda) + \sum_{i=1}^{n}\log\left(\frac{2e^{-x_i/\lambda}}{(1+e^{-x_i/\lambda})^2}\right) + (k-1)\sum_{i=1}^{n}\log\left[\frac{1-e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}\right]. \quad (11)$$

$$\frac{dL}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^2}\sum_{i=1}^{m}\frac{x_i(1-e^{-x_i/\lambda})}{1+e^{-x_i/\lambda}} - \frac{2(k-1)}{\lambda^2}\sum_{i=1}^{n}\frac{x_i e^{-x_i/\lambda}}{1-e^{-2x_i/\lambda}}$$

$$\frac{d^2L}{d\lambda^2} = \frac{n}{\lambda^2} - \frac{2}{\lambda^4}\sum_{i=1}^{n}\frac{x_i^2 e^{-x_i/\lambda}}{(1+e^{-x_i/\lambda})^2} - \frac{2}{\lambda^3}\sum_{i=1}^{n}\frac{x_i(1-e^{-x_i/\lambda})}{1+e^{-x_i/\lambda}} - \frac{2(k-1)}{\lambda^4}\sum_{i=1}^{n}\frac{x_i^2 e^{-x_i/\lambda}(1+e^{-2x_i/\lambda})}{(1-e^{-x_i/\lambda})^2(1+e^{-x_i/\lambda})^2} + \frac{4(k-1)}{\lambda^3}\sum_{i=1}^{n}\frac{x_i e^{-x_i/\lambda}}{(1-e^{-x_i/\lambda})(1+e^{-x_i/\lambda})}$$

and

$$I_w(\lambda) = -\frac{n}{\lambda^2} + \frac{2}{\lambda^4}\sum_{i=1}^{n} E\left[\frac{X_i^2 e^{-X_i/\lambda}}{(1+e^{-X_i/\lambda})^2}\right] + \frac{2}{\lambda^3}\sum_{i=1}^{n} E\left[\frac{X_i(1-e^{-X_i/\lambda})}{1+e^{-X_i/\lambda}}\right] + \frac{2(k-1)}{\lambda^4}\sum_{i=1}^{n} E\left[\frac{X_i^2 e^{-X_i/\lambda}(1+e^{-2X_i/\lambda})}{(1-e^{-X_i/\lambda})^2(1+e^{-X_i/\lambda})^2}\right] - \frac{4(k-1)}{\lambda^3}\sum_{i=1}^{n} E\left[\frac{X_i e^{-X_i/\lambda}}{(1-e^{-X_i/\lambda})(1+e^{-X_i/\lambda})}\right]. \quad (12)$$

Missing information is

$$I_{W|X}(\lambda) = \sum_{i=1}^{m} R_i I_{W|X}^{(i)}(\lambda) = -\sum_{i=1}^{m}\sum_{j=1}^{R_i} E_{Z|X}\left[\frac{d^2 \log(f(Z_{ij}|X_i,\lambda))}{d\lambda^2}\right].$$

Consider

$$f_{z|X}(Z_{ij}|X_i,\lambda) = \frac{f(z_{ij};\lambda)}{1-F(x_i;\lambda)} = \frac{(k/\lambda)\frac{2e^{-z_{ij}/\lambda}}{(1+e^{-z_{ij}/\lambda})^2}[(1-e^{-z_{ij}/\lambda})/(1+e^{-z_{ij}/\lambda})]^{k-1}}{1-\left[\frac{1-e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}\right]^k}.$$

Therefore,

$$\log f = \log k - \log\lambda - \log\left(\frac{2e^{-z_{ij}/\lambda}}{(1+e^{-z_{ij}/\lambda})^2}\right) + (k-1)\log\left[\frac{1-e^{-z_{ij}/\lambda}}{1+e^{-z_{ij}/\lambda}}\right] - \log\left\{1-\left[\frac{1-e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}\right]^k\right\}.$$

$$\frac{d\log f}{d\lambda} = -\frac{1}{\lambda} + \frac{z_{ij}(1-e^{-z_{ij}/\lambda})}{\lambda^2(1+e^{-z_{ij}/\lambda})} - \frac{2(k-1)z_{ij}e^{-z_{ij}/\lambda}}{\lambda^2(1-e^{-2z_{ij}/\lambda})} - \frac{2kx_i e^{-x_i/\lambda}[1-e^{-x_i/\lambda}]^{k-1}}{\lambda^2(1+e^{-x_i/\lambda})^{(k+1)}\left\{1-\left[\frac{1-e^{-x_i/\lambda}}{1+e^{-x_i/\lambda}}\right]^k\right\}}$$

and

$$\frac{d^2\log f}{d\lambda^2} = \frac{1}{\lambda^2} - \frac{2z_{ij}^2 e^{-\frac{z_{ij}}{\lambda}}}{\lambda^4[1+e^{-\frac{z_{ij}}{\lambda}}]^2} - \frac{2z_{ij}(1-e^{-\frac{z_{ij}}{\lambda}})}{\lambda^3(1+e^{-\frac{z_{ij}}{\lambda}})} - \frac{2(k-1)z_{ij}^2 e^{-\frac{z_{ij}}{\lambda}}(1+e^{-\frac{2z_{ij}}{\lambda}})}{\lambda^4(1-e^{-\frac{z_{ij}}{\lambda}})^2(1+e^{-\frac{z_{ij}}{\lambda}})^2} + \frac{4(k-1)z_{ij}e^{-\frac{z_{ij}}{\lambda}}}{\lambda^3(1-e^{-\frac{z_{ij}}{\lambda}})(1+e^{-\frac{z_{ij}}{\lambda}})} + \frac{2kx_i^2 e^{-\frac{x_i}{\lambda}}\left[\frac{1-e^{-x_i/\lambda}}{1+e^{-\frac{x_i}{\lambda}}}\right]^{k-2}[2(k-1)e^{-\frac{x_i}{\lambda}} - (1-e^{-\frac{z_{ij}}{\lambda}})^2]}{\lambda^4(1+e^{-\frac{x_i}{\lambda}})^4\left\{1-\left[\frac{1-e^{-\frac{x_i}{\lambda}}}{1+e^{-\frac{x_i}{\lambda}}}\right]^k\right\}} + \frac{4k^2x_i^2e^{-\frac{2x_i}{\lambda}}(1-e^{-\frac{x_i}{\lambda}})^{2k-2}}{\lambda^4(1+e^{-\frac{x_i}{\lambda}})^{2k+2}\left\{1-\left[\frac{1-e^{-\frac{x_i}{\lambda}}}{1+e^{-\frac{x_i}{\lambda}}}\right]^k\right\}^2} + \frac{4kx_ie^{-\frac{x_i}{\lambda}}(1-e^{-\frac{x_i}{\lambda}})^{k-1}}{\lambda^3(1+e^{-\frac{x_i}{\lambda}})^{k+1}\left\{1-\left[\frac{1-e^{-\frac{x_i}{\lambda}}}{1+e^{-\frac{x_i}{\lambda}}}\right]^k\right\}}.$$

Also,

$$
\begin{aligned}
I_{W|X}(\lambda) &= \sum_{i=1}^{m} R_i I_{W|X}^{(i)}(\lambda), \\
&= -\frac{n-m}{\lambda^2} + \frac{2}{\lambda^4}\sum_{i=1}^{m}\sum_{j=1}^{R_i} E\left[\frac{Z_{ij}^2 \mathrm{e}^{-\frac{Z_{ij}}{\lambda}}}{(1+\mathrm{e}^{-\frac{Z_{ij}}{\lambda}})^2}\right] + \frac{2}{\lambda^3}\sum_{i=1}^{m}\sum_{j=1}^{R_i} E\left[\frac{Z_{ij}(1-\mathrm{e}^{-\frac{Z_{ij}}{\lambda}})}{1+\mathrm{e}^{-\frac{Z_{ij}}{\lambda}}}\right] \\
&\quad + \frac{2(k-1)}{\lambda^4}\sum_{i=1}^{m}\sum_{j=1}^{R_i} E\left[\frac{Z_{ij}^2 \mathrm{e}^{-\frac{Z_{ij}}{\lambda}}(1+\mathrm{e}^{-\frac{2Z_{ij}}{\lambda}})}{(1-\mathrm{e}^{-\frac{Z_{ij}}{\lambda}})^2(1+\mathrm{e}^{-\frac{Z_{ij}}{\lambda}})^2}\right] \\
&\quad - \frac{4(k-1)}{\lambda^3}\sum_{i=1}^{m}\sum_{j=1}^{R_i} E\left[\frac{Z_{ij}\mathrm{e}^{-\frac{Z_{ij}}{\lambda}}}{(1-\mathrm{e}^{-\frac{Z_{ij}}{\lambda}})(1+\mathrm{e}^{-\frac{Z_{ij}}{\lambda}})}\right] \\
&\quad - \frac{2k}{\lambda^4}\sum_{i=1}^{m} \frac{R_i x_i^2 \mathrm{e}^{-\frac{x_i}{\lambda}}\left[\frac{1-\mathrm{e}^{-\frac{x_i}{\lambda}}}{1+\mathrm{e}^{-\frac{x_i}{\lambda}}}\right]^{k-2}[2(k-1)\mathrm{e}^{-\frac{x_i}{\lambda}} - (1-\mathrm{e}^{-\frac{x_i}{\lambda}})^2]}{(1+\mathrm{e}^{-\frac{x_i}{\lambda}})^4\left\{1-\left[\frac{1-\mathrm{e}^{-\frac{x_i}{\lambda}}}{1+\mathrm{e}^{-\frac{x_i}{\lambda}}}\right]^k\right\}} \\
&\quad - \frac{4k^2}{\lambda^4}\sum_{i=1}^{m} \frac{R_i x_i^2 \mathrm{e}^{-\frac{2x_i}{\lambda}}(1-\mathrm{e}^{-\frac{x_i}{\lambda}})^{2k-2}}{(1+\mathrm{e}^{-\frac{x_i}{\lambda}})^{2k+2}\left\{1-\left[\frac{1-\mathrm{e}^{-\frac{x_i}{\lambda}}}{1+\mathrm{e}^{-\frac{x_i}{\lambda}}}\right]^k\right\}^2} \\
&\quad - \frac{4k}{\lambda^3}\sum_{i=1}^{m} \frac{R_i x_i \mathrm{e}^{-\frac{x_i}{\lambda}}(1-\mathrm{e}^{-\frac{x_i}{\lambda}})^{k-1}}{(1+\mathrm{e}^{-\frac{x_i}{\lambda}})^{k+1}\left\{1-\left[\frac{1-\mathrm{e}^{-\frac{x_i}{\lambda}}}{1+\mathrm{e}^{-\frac{x_i}{\lambda}}}\right]^k\right\}}. \qquad (13)
\end{aligned}
$$

4.3. *Tolerance interval*

Let $l_\beta(\lambda)$ be the lower quantile of order β of the cdf $F(x;\lambda)$. Then, we have

$$l_\beta(\lambda) = -\lambda \log\left(\frac{1-\beta^{1/k}}{1+\beta^{1/k}}\right).$$

Thus, an upper β-expectation tolerance interval for $F(x;\lambda)$ is obtained by

$$I_\beta = (0, l_\beta(\lambda)).$$

The MLE of $l_\beta(\lambda)$ is given by

$$l_\beta(\hat{\lambda}_n) = -\hat{\lambda}_n \log\left(\frac{1-\beta^{1/k}}{1+\beta^{1/k}}\right)$$

yielding an approximate β-expectation tolerance interval as

$$\hat{I}_\beta = (0, l_\beta(\hat{\lambda}_n)).$$

The expectation of $\hat{I}_\beta$ can be obtained approximately using the approach suggested and given as

$$E[F(l_\beta(\hat{\lambda}_n);\lambda)] \approx \beta - 0.5F_{02}\sigma^2(\hat{\lambda}_n) + \frac{F_{01}\sigma^2(\hat{\lambda}_n)F_{11}}{F_{10}}, \qquad (14)$$

Table 1. Bias, MSE[a], Confidence levels and its SE[a] for $k = 2$ and $\lambda = 1$.

N	m	Scheme no.	Scheme	Bias and MSE	Confidence level and SE (MLE)		Confidence level and SE (log MLE)	
					90%	95%	90%	95%
5	2	[1]	(3, 0)	−0.0195	0.8309	0.8645	0.8765	0.9233
				(0.1659)	(0.0281)	(0.0234)	(0.0216)	(0.0142)
		[2]	(0, 3)	−0.0319	0.8299	0.8639	0.8756	0.9199
				(0.1421)	(0.0282)	(0.0235)	(0.0218)	(0.0147)
		[3]	(1, 2)	−0.0325	0.8251	0.8625	0.8716	0.9200
				(0.1462)	(0.0289)	(0.0237)	(0.0224)	(0.0147)
		[4]	(2, 1)	−0.0309	0.8261	0.8597	0.8708	0.9218
				(0.1503)	(0.0287)	(0.0241)	(0.0225)	(0.0144)
15	5	[5]	(10, 4*0)	0.0064	0.8736	0.9114	0.8932	0.9420
				(0.0714)	(0.0074)	(0.0054)	(0.0064)	(0.0036)
		[6]	(4*0, 10)	−0.0151	0.8656	0.9075	0.8847	0.9355
				(0.0570)	(0.0078)	(0.0056)	(0.0068)	(0.0040)
		[7]	(2, 2, 2, 2, 2)	−0.0117	0.8769	0.9149	0.8946	0.9412
				(0.0579)	(0.0072)	(0.0052)	(0.0063)	(0.0037)
	10	[8]	(5, 9*0)	0.0052	0.8869	0.9297	0.8919	0.9453
				(0.0389)	(0.0067)	(0.0044)	(0.0064)	(0.0034)
		[9]	(9*0, 5)	−0.0011	0.8887	0.9298	0.8936	0.9452
				(0.0330)	(0.0066)	(0.0044)	(0.0063)	(0.0035)
		[10]	(3, 2, 8*0)	−0.0011	0.8832	0.9275	0.8931	0.9450
				(0.0381)	(0.0069)	(0.0045)	(0.0064)	(0.0035)
20	10	[11]	(10, 9*0)	0.0013	0.8913	0.9340	0.8968	0.9480
				(0.0373)	(0.0048)	(0.0031)	(0.0046)	(0.0025)
		[12]	(9*0, 10)	−0.0041	0.8830	0.9283	0.8936	0.9431
				(0.0307)	(0.0052)	(0.0033)	(0.0048)	(0.0027)
25	10	[13]	(15, 9*0)	−0.0031	0.8841	0.9294	0.8969	0.9493
				(0.0373)	(0.0041)	(0.0026)	(0.0037)	(0.0019)
		[14]	(9*0, 15)	−0.0057	0.8897	0.9318	0.8945	0.9456
				(0.0284)	(0.0039)	(0.0025)	(0.0038)	(0.0021)
		[15]	(5, 5, 5, 7*0)	−0.0001	0.8908	0.9330	0.8985	0.9484
				(0.0353)	(0.0039)	(0.0025)	(0.0036)	(0.0020)
	15	[16]	(10, 14*0)	0.0011	0.8955	0.9399	0.9021	0.9503
				(0.0250)	(0.0037)	(0.0023)	(0.0035)	(0.0019)
		[17]	(14*0, 10)	−0.0033	0.8912	0.9362	0.8971	0.9454
				(0.0209)	(0.0039)	(0.0024)	(0.0037)	(0.0021)
30	10	[18]	(20, 9*0)	0.0004	0.8900	0.9320	0.8961	0.9477
				(0.0369)	(0.0033)	(0.0021)	(0.0031)	(0.0017)
		[19]	(9*0, 20)	−0.0074	0.8855	0.9291	0.8953	0.9441
				(0.0278)	(0.0034)	(0.0022)	(0.0031)	(0.0018)
	15	[20]	(15, 14*0)	0.0001	0.8888	0.9357	0.8975	0.9440
				(0.0255)	(0.0033)	(0.0020)	(0.0031)	(0.0018)
		[21]	(14*0, 15)	−0.0063	0.8860	0.9321	0.8938	0.9444
				(0.0202)	(0.0034)	(0.0021)	(0.0032)	(0.0018)
		[22]	(5, 5, 5, 12*0)	−0.0022	0.8871	0.9354	0.8951	0.9420
				(0.0249)	(0.0033)	(0.0020)	(0.0031)	(0.0018)
	20	[23]	(10, 19*0)	0.0008	0.8964	0.9410	0.9002	0.9517
				(0.0190)	(0.0031)	(0.0019)	(0.0030)	(0.0015)
		[24]	(19*0, 10)	−0.0023	0.8940	0.9402	0.8990	0.9484
				(0.0159)	(0.0032)	(0.0019)	(0.0030)	(0.0016)
		[25]	(0, 5, 5, 17*0)	0.0015	0.8943	0.9390	0.8984	0.9450
				(0.0191)	(0.0032)	(0.0019)	(0.0030)	(0.0017)
50	20	[26]	(30, 19*0)	−0.00161	0.8941	0.9392	0.8987	0.9476
				(0.0188)	(0.0019)	(0.0011)	(0.0018)	(0.0010)
		[27]	(19*0, 30)	−0.0031	0.8914	0.9390	0.8970	0.9475
				(0.0145)	(0.0019)	(0.0011)	(0.0018)	(0.0010)
	35	[28]	(15, 34*0)	0.0005	0.8942	0.9442	0.8978	0.9496
				(0.0111)	(0.0019)	(0.0011)	(0.0018)	(0.0010)
		[29]	(34*0, 15)	−0.0017	0.8980	0.9459	0.9014	0.94800
				(0.0093)	(0.0018)	(0.0010)	(0.0018)	(0.0010)
		[30]	(5, 5, 5, 32*0)	0.0011	0.9002	0.9458	0.9039	0.9504
				(0.0106)	(0.0018)	(0.0010)	(0.0017)	(0.0009)

Note: [a]MSE and SE are given in parenthesis.

where

$$F_{10} = \frac{2k\mathrm{e}^{-x/\lambda}}{\lambda}\frac{(1-\mathrm{e}^{-x/\lambda})^{k-1}}{(1+\mathrm{e}^{-x/\lambda})^{k+1}},$$

$$F_{01} = -\frac{2kx\mathrm{e}^{-x/\lambda}}{\lambda^2}\frac{(1-\mathrm{e}^{-x/\lambda})^{k-1}}{(1+\mathrm{e}^{-x/\lambda})^{k+1}},$$

$$F_{11} = \frac{2k}{\lambda^3}\mathrm{e}^{-x/\lambda}\frac{(1-\mathrm{e}^{-x/\lambda})^{k-2}}{(1+\mathrm{e}^{-x/\lambda})^{k+2}}\{\mathrm{e}^{-2x/\lambda}(x+\lambda) - 2kx\mathrm{e}^{-x/\lambda} + x - \lambda\},$$

and

$$F_{02} = -\frac{2kx}{\lambda^4}\mathrm{e}^{-x/\lambda}\frac{(1-\mathrm{e}^{-x/\lambda})^{k-2}}{(1+\mathrm{e}^{-x/\lambda})^{k+2}}\{x - 2\lambda + \mathrm{e}^{-2x/\lambda}(x+2\lambda) - 2kx\mathrm{e}^{-x/\lambda}\}.$$

5. Simulation study

A simulation is carried out to study the performance of the MLE when the lifetime distribution of each unit follows the half-logistic distribution. Estimates of bias and the MSE for various progressively Type-II censoring scheme are obtained. Asymptotic confidence intervals based on the MLE and log-transformed MLE are compared with their confidence levels. The coverage of the β-expectation tolerance intervals is also studied. The algorithm by Balakrishnan and Sandhu [22] is used to generate progressively censored samples from half-logistic distribution of a k-unit parallel system.

- Algorithm
  1. Generate independently and identically distributed observations $(W_1, W_2, \ldots, W_m)$ from $U(0,1)$.
  2. For $(R_1, R_2, \ldots, R_m)$ censoring scheme set $E_i = 1/(i + R_m + R_{m-1} + \cdots + R_{m-i+1})$ for $i = 1, 2, \ldots, m$.
  3. Set $V_i = W_i^{E_i}$ for $i = 1, 2, \ldots, m$.
  4. Set $U_i = 1 - V_m\, V_{m-1} \ldots V_{m-i+1}$ for $i = 1, 2, \ldots, m$. Then $(U_1, U_2, \ldots, U_m)$ is the uniform (0,1) progressively Type-II censored sample.
  5. For the given value of the parameter λ, set

$$x_{(i)} = -\lambda \log\left[\frac{1-(U_i)^{1/k}}{1+(U_i)^{1/k}}\right] \quad \text{for } i = 1, 2, \ldots, m. \tag{15}$$

  $x_{(1)}, x_{(2)}, \ldots, x_{(m)}$ is the required progressively Type-II-censored sample from the distribution of a k-unit parallel system with half-logistic distribution as the distribution of each unit of the system.

In Table 1, scheme (a, b) stands for $R_1 = a$ and $R_2 = b$. A similar meaning holds for schemes described through completely specified vector, while scheme (10, 4*0) means $R_1 = 10$ and rest four R_i's are zero. That is $R_2 = R_3 = R_4 = R_5 = 0$.

Simulation is carried out for 2-unit parallel system with $\lambda = 1$. The EM algorithm and Newton–Raphson method are used to compute the MLE. For each particular progressive censoring scheme, 10,000 sets of observations are generated. The bias, the MSE, confidence levels for the corresponding approximate confidence intervals for λ along with their standard errors (SE) are displayed in Table 1. The simulated mean coverage and the estimated expectation of the tolerance interval along with their SE are given in Table 2.

Table 2. Simulated mean, estimated expectation and its SE[a] of the approximate β-expectation tolerance interval for $k = 2$ and $\lambda = 1$.

n	m	Scheme no.	Scheme	Simulated mean and SE			Estimated expectation and SE		
				90%	95%	99%	90%	95%	99%
5	2	[1]	(3, 0)	0.8058	0.8658	0.9375	0.8045	0.8755	0.9600
				(0.0962)	(0.0833)	(0.0568)	(0.0428)	(0.0334)	(0.0134)
		[2]	(0, 3)	0.8083	0.8694	0.9411	0.8163	0.8847	0.9637
				(0.0922)	(0.0792)	(0.0529)	(0.0374)	(0.0293)	(0.0118)
		[3]	(1, 2)	0.8097	0.8700	0.9409	0.8147	0.8834	0.9632
				(0.0926)	(0.0799)	(0.0539)	(0.0382)	(0.0297)	(0.0118)
		[4]	(2, 1)	0.8083	0.8690	0.9405	0.8119	0.8812	0.9623
				(0.0930)	(0.0800)	(0.0535)	(0.0395)	(0.0307)	(0.0126)
15	5	[5]	(10, 4*0)	0.8592	0.9158	0.9728	0.8594	0.9183	0.9772
				(0.0310)	(0.0237)	(0.0118)	(0.0106)	(0.0082)	(0.0037)
		[6]	(4*0, 10)	0.8623	0.9190	0.9747	0.8680	0.9250	0.9799
				(0.0288)	(0.0219)	(0.0110)	(0.0082)	(0.0063)	(0.0026)
		[7]	(2, 2, 2, 2, 2)	0.8616	0.9184	0.9745	0.8663	0.9237	0.9794
				(0.0290)	(0.0221)	(0.0106)	(0.0086)	(0.0068)	(0.0026)
	10	[8]	(5, 9*0)	0.8803	0.9339	0.9826	0.8788	0.9334	0.9833
				(0.0198)	(0.0139)	(0.0058)	(0.0052)	(0.0045)	(0.0016)
		[9]	(9*0, 5)	0.8800	0.9340	0.9829	0.8817	0.9357	0.9843
				(0.0190)	(0.0132)	(0.0052)	(0.0045)	(0.0037)	(0.0014)
		[10]	(3, 2, 8*0)	0.8793	0.9331	0.9823	0.8789	0.9335	0.9834
				(0.0202)	(0.0141)	(0.0058)	(0.0052)	(0.0045)	(0.0016)
20	10	[11]	(10, 9*0)	0.8798	0.9334	0.9824	0.8789	0.9335	0.9834
				(0.0174)	(0.0123)	(0.0050)	(0.0047)	(0.0037)	(0.0015)
		[12]	(9*0, 10)	0.8802	0.9343	0.9831	0.8830	0.9367	0.9847
				(0.0160)	(0.0113)	(0.0045)	(0.0038)	(0.0030)	(0.0012)
25	10	[13]	(15, 9*0)	0.8800	0.9337	0.9825	0.8790	0.9336	0.9834
				(0.0154)	(0.0109)	(0.0044)	(0.0042)	(0.0033)	(0.0013)
		[14]	(9*0, 15)	0.8817	0.9355	0.9836	0.8837	0.9372	0.9849
				(0.0138)	(0.0096)	(0.0038)	(0.0033)	(0.0026)	(0.0010)
		[15]	(5, 5, 5, 7*0)	0.8814	0.9347	0.9829	0.8799	0.9343	0.9837
				(0.0150)	(0.0106)	(0.0042)	(0.0040)	(0.0031)	(0.0013)
	15	[16]	(10, 14*0)	0.8860	0.9387	0.9850	0.8858	0.9389	0.9855
				(0.0123)	(0.0083)	(0.0031)	(0.0028)	(0.0022)	(0.0009)
		[17]	(14*0, 10)	0.8873	0.9400	0.9857	0.8882	0.9408	0.9863
				(0.0112)	(0.0076)	(0.0027)	(0.0024)	(0.0018)	(0.0007)
30	10	[18]	(20, 9*0)	0.8783	0.9324	0.9820	0.8791	0.9337	0.9834
				(0.0144)	(0.0102)	(0.0041)	(0.0037)	(0.0032)	(0.0012)
		[19]	(9*0, 20)	0.8813	0.9352	0.9835	0.8841	0.9376	0.9850
				(0.0126)	(0.0088)	(0.0037)	(0.0032)	(0.0026)	(0.0008)
	15	[20]	(15, 14*0)	0.8866	0.9391	0.9852	0.8858	0.9389	0.9856
				(0.0111)	(0.0075)	(0.0028)	(0.0026)	(0.0020)	(0.0008)
		[21]	(14*0, 15)	0.8868	0.9398	0.9857	0.8887	0.9412	0.9864
				(0.0100)	(0.0067)	(0.0024)	(0.0021)	(0.0016)	(0.0006)
		[22]	(5, 5, 5, 12*0)	0.8874	0.9397	0.9854	0.8861	0.9392	0.9857
				(0.0109)	(0.0074)	(0.0027)	(0.0025)	(0.0020)	(0.0008)
	20	[23]	(10, 19*0)	0.8895	0.9416	0.9864	0.8893	0.9416	0.9866
				(0.0095)	(0.0063)	(0.0022)	(0.0020)	(0.0015)	(0.0006)
		[24]	(19*0, 10)	0.8913	0.9430	0.9870	0.8909	0.9429	0.9871
				(0.0087)	(0.0057)	(0.0020)	(0.0017)	(0.0013)	(0.0005)
		[25]	(0, 5, 5, 17*0)	0.8900	0.9420	0.9865	0.8894	0.9418	0.9867
				(0.0094)	(0.0062)	(0.0022)	(0.0019)	(0.0015)	(0.0006)
50	20	[26]	(30, 19*0)	0.8893	0.9415	0.9863	0.8894	0.9417	0.9867
				(0.0074)	(0.0049)	(0.0017)	(0.0015)	(0.0012)	(0.0005)
		[27]	(19*0, 30)	0.8906	0.9426	0.9869	0.8919	0.9436	0.9874
				(0.0065)	(0.0043)	(0.0015)	(0.0012)	(0.0009)	(0.0004)
	35	[28]	(15, 34*0)	0.8939	0.9452	0.9880	0.8939	0.9452	0.9881
				(0.0053)	(0.0035)	(0.0011)	(0.0009)	(0.0006)	(0.0002)
		[29]	(34*0, 15)	0.8945	0.9457	0.9882	0.8947	0.9459	0.9883
				(0.0049)	(0.0032)	(0.0010)	(0.00003)	(0.0008)	(0.0002)
		[30]	(5, 5, 5, 32*0)	0.8942	0.9454	0.9881	0.8939	0.9452	0.9881
				(0.0053)	(0.0035)	(0.0011)	(0.0009)	(0.0006)	(0.0003)

Note: [a]SE are given in parenthesis.

6. Conclusion and discussion

Results of the simulation study reported in Table 1 indicate that bias and MSE of the MLE decrease as the sample size and the effective sample size increase. The MSE of the MLE is smaller for the conventional Type-II censoring scheme as compared with the progressively Type-II censoring scheme. The coverage performance of asymptotic confidence intervals is satisfactory. Confidence interval based on the log-transformed MLE shows better performance than the one based on the MLE. Results of the simulation study for the β-expectation tolerance interval, which are tabulated in Table 2, indicate that, the estimated expectation and simulated mean for small sample size are marginally lower than the nominal value. As the sample size increases, the performance of tolerance intervals improves. The SE of both the estimated expectation and of simulated mean coverage of the tolerance intervals decrease as the sample size increases. The SE of the estimated expectation is significantly smaller than that of the simulated mean coverage.

Estimation procedures reported in this paper are applicable for a wide class of lifetime distributions under progressively Type-II and conventional Type-II censoring schemes. The results reported in this paper can also be obtained when 'k' in the pdf is replaced by any known positive number greater than one.

Acknowledgements

The authors are grateful to the editor and the expert referee for making constructive and valuable comments that have significantly improved the contents of this article. The first author thanks the University Grants Commission, New Delhi, India, for providing Fellowship under Faculty Improvement Programme to carry out this research.

References

[1] A.C. Cohen, *Progressively censored samples in life testing*, Technometrics 5 (1963), pp. 327–329.
[2] N.R. Mann, *Exact three-order-statistic confidence bounds on reliable life for a Weibull model with progressive censoring*, J. Amer. Statist. Assoc. 64 (1969), pp. 306–315.
[3] N.R. Mann, *Best linear invariant estimation for Weibull parameters under progressive censoring*, Technometrics 13 (1971), pp. 521–533.
[4] N. Balakrishnan and A. Asgharzadeh, *Inference for the scaled half-logistic distribution based on progressively Type-II censored samples*, Commun. Statist. Theory Meth. 34 (2005), pp. 73–87.
[5] N. Balakrishnan, N. Kannan, C.T. Lin, and H.K.T. Ng, *Point and interval estimation for Gaussian distribution, based on progressively Type-II censored samples*, IEEE Trans. Reliab. 52(1) (2003), pp. 90–95.
[6] N. Balakrishnan, N. Kannan, C.T. Lin, and S.J.S. Wu, *Inference for the extreme value distribution under progressive Type-II Censoring*, J. Statist. Comput. Simul. 74 (2004), pp. 25–45.
[7] H.K.T. Ng, *Parameter estimation for a modified Weibull distribution, for progressively Type-II censored samples*, IEEE Trans. Reliab. 54 (2005), pp. 374–380.
[8] N. Balakrishnan and R. Aggarwala, *Progressive Censoring: Theory, Methods and Applications*, Birkhauser, Boston, MA, 2000.
[9] N. Balakrishnan, *Progressive censoring methodology: An appraisal*, Test 16(2) (2007), pp. 211–259.
[10] B. Pradhan, *Point and interval estimation for the lifetime distribution of a k-unit parallel system based on progressively Type-II censored data*, Econ. Qual. Control 22(2) (2007), pp. 175–186.
[11] C. Kim and K. Han, *Estimation of the scale parameter of the half-logistic distribution under progressively Type-II censored sample*, Statist. Papers 51(2010), pp. 375–387.
[12] G. Iliopoulous and N. Balakrishnan, *Exact likelihood inference for Laplace distribution based on Type-II censored samples*, J. Statist. Plann. Inference 141 (2011), pp. 1224–1239.
[13] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Statist. Soc. B 39 (1977), pp. 1–38.
[14] G.J. Mclachlan and T. Krishnan, *The EM Algorithm and Extensions*, John Wiley and Sons, New York, 1997.
[15] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, John Wiley and Sons, New York, 2002.
[16] B. Pradhan and D. Kundu, *On progressively censored generalized exponential distribution*, Test 18 (2009), pp. 497–515.
[17] H.K.T. Ng, P.S. Chan, and N. Balakrishnan, *Estimation of parameters from progressively censored data using EM algorithm*, Comp. Statist. Data Anal. 39 (2002), pp. 371–386.

[18] T.A. Louis, *Finding the observed information matrix using the EM algorithm*, J. Roy. Statist. Soc. B 44 (1982), pp. 226–233.
[19] R.R. Kumbhar and D.T. Shirke, *Tolerance limits for lifetime distribution of k-unit parallel system*, J. Statist. Comput. Simul. 74 (2004), pp. 201–213.
[20] W.Q. Meeker and L.A. Escober, *Statistical Methods for Reliability Data*, John Wiley & Sons, New York, 1998.
[21] C.L. Atwood, *Approximate tolerance intervals based on maximum likelihood estimator*, J. Amer. Statist. Assoc. 79 (1984), pp. 459–465.
[22] N. Balakrishnan and R.A. Sandhu, *A simple simulation algorithm for generating progressive Type-II censored samples*, Amer. Statist. 49 (1995), pp. 229–230.

Journal of Academia and Industrial Research (JAIR)
Volume 2, Issue 2 July 2013

ISSN: 2278-5213

RESEARCH MANUSCRIPT

$\overline{X}$ Charts with Variable Sampling interval and Warning limits

Shashibhushan B. Mahadik
Dept. of Statistics, Solapur University, Solapur-413255, India
sbmahadik@yahoo.com; +91 9921516765

Abstract

The idea of *variable sampling interval and warning limits* (VSIWL) is proposed for $\overline{X}$ charts. Expressions for the performance measures for the charts with VSIWL are developed. The methods presented are general and can be applied to other Shewhart control charts. The performances of VSIWL $\overline{X}$ charts are compared numerically with that of VSI $\overline{X}$ charts with and without runs rules for switching between sampling interval lengths. It is observed that in general the former charts perform significantly better than the later.

Keywords: Adaptive control chart, average number of samples to signal, statistical process control, Shewhart control charts.

Introduction

Nowadays it has been well recognized that adaptive control charts are significantly more efficient than the static ones. Reynolds *et al*. (1988) were the first to consider the intuitive notion of adapting sampling interval length of a control charts according to the status of a process indicated by the last plotted sample point. They proposed *variable sampling interval* (VSI) $\overline{X}$ charts. The principle of choosing the sampling interval length in a VSI chart is that as the location of the current sample point approaches the control limits, tighten the control by taking the next sample more quickly. The in-control area of the chart is partitioned into a central region and one or more warning regions. Each region determines length of the sampling interval for the next sample if the current sample point falls in it. Reynolds *et al.* (1988) showed that the idea of VSI substantially improves the statistical performance of $\overline{X}$ charts. Also, they showed that the statistical performance of a VSI $\overline{X}$ chart in detecting a shift of any magnitude that exists initially is optimized by using the dual sampling interval policy consisting of the shortest and longest possible sampling interval lengths. Afterwards, Prabhu*et al.* (1993) and Costa (1994) independently proposed variable sample size $\overline{X}$ charts. Prabhu *et al.* (1994) proposed variable sample size and sampling interval $\overline{X}$ charts. Costa (1999a) proposed the adaptive $\overline{X}$ charts in which all the three design parameters are variable. Tagaras (1998) reported an extensive survey of the research on adaptive control charts until 1997. Then also, various schemes of adaptive control charts have been proposed and extensively investigated with different perspectives. See for example, Amin and Widmaier (1999), Costa(1999b), Aparasi and Haro (2003), Costa and Rahim (2001), Epprecht *et al.* (2003), Zimmer *et al.* (1998), Reynolds and Stoumbos (2001), Wu *et al.* (2005), Yu and Hou (2006), Celano *et al.* (2006), Chen (2007), Wu *et al.* (2007), Yang and Su (2007), Mahadik and Shirke (2007a, b), Jiang *et al.* (2008), Jensen *et al.* (2008), Luo *et al.* (2009), Wu *et al.* (2009), Shi *et al.* (2009), De Magalhaes *et al.* (2009), Celano (2009), Faraz and Moghadam (2009), Mahadik and Shirke (2009, 2011), Li and Wang (2010), Epprecht *et al.* (2010), Shu *et al.* (2010), Mahadik (2012a, b, 2013), Chen *et al.* (2011), Dai *et al.* (2011),Faraz and Saniga (2011), Nenes (2011), Kooli and Limam (2011) and Lee (2011).

The weakness of any adaptive control chart is the inconvenience in its administration due to the frequent switches between the values of its adaptive design parameters. In order to reduce the frequency of switches between sampling interval lengths of VSI charts, Amin and Letsinger (1991) proposed the use runs rules for switching between these lengths. Amin and Hemasinha (1993) developed approximate expressions for the performance measures for VSI $\overline{X}$ charts with such runs rules while Mahadik (2011a) developed the exact expressions.

In the present study, the idea of variable warning limits is proposed for VSI $\overline{X}$ charts. This significantly improves statistical performances of the charts in detecting small to moderate shifts in the process mean and also dramatically reduces the frequency of switches between sampling interval lengths.

Materials and methods

A VSIWL $\overline{X}$ Chart: Let the quality characteristic *X* to be monitored follows a normal distribution with mean μ, and a known and constant standard deviation σ. Suppose μ_0 is the target value of μ.

An occurrence of an assignable cause results in a shift of size δ in μ, where δ is expressed in σ units. It is assumed that δ remains constant following the occurrence of a shift until it is detected. A VSIWL $\bar{X}$ chart to monitor μ is as described below.

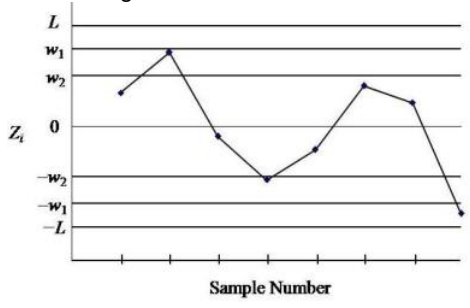
The chart statistic is the standardized sample mean $Z_i = \sqrt{n}\left(\bar{X}_i - \mu_0\right)/\sigma$, where $\bar{X}_i$, i = 1, 2, …, is the mean of i^{th} sample of size is n drawn on X. Note that when $\mu = \mu_0$, $Z_i \sim N(0, 1)$, and when $\mu = \mu_0 + \delta\sigma$, $Z_i \sim N(\sqrt{n}\delta, 1)$. Each control limit of the chart is at the distance of L units from its centerline. Let $t(i)$ be the length of sampling interval between $(i-1)^{st}$ and i^{th} trials and $w(i)$ be the distance of each warning limit from the centerline for the i^{th} trial, i = 1, 2, …. The values of $(t(i), w(i))$ can be either (t_1, w_1) or (t_2, w_2), where t_1, t_2, w_1, and w_2 are such that $t_{max} \geq t_1 \geq t_2 \geq t_{min}$ and $L > w_1 \geq w_2 > 0$, where t_{min} and t_{max} are the shortest and longest possible sampling intervals, respectively. When Z_{i-1} falls within the control limits, the pair of values of $(t(i), w(i))$, i = 2, 3, …,between (t_1, w_1) and (t_2, w_2) is chosen according to the following rule

$$(t(i), w(i)) = \begin{cases} (t_1, w_1), & \text{if } Z_{i-1} \in I_1 \\ (t_2, w_2), & \text{if } Z_{i-1} \in I_2, \end{cases}$$

where $I_1 = [-w(i), w(i)]$ and $I_2 = (-L, -w(i)) \cup (w(i), L)$ for the i^{th} trial, i = 1, 2, …

At start-up the values of $(t(1), w(1))$ can be chosen using an arbitrary probability distribution, as no prior sample is available. In practice, it is recommended to use the pair (t_2, w_2) for the first trial to provide additional protection against the problems that may exist initially. The trial following an out-of-control signal is again treated to be the first trial and the mechanism of choosing $(t(i), w(i))$is restarted from that. The chart signals an out-of-control state when a sample point falls beyond the control limits. Figure 1 shows a typical VSIWL $\bar{X}$ chart.

Fig. 1. A VSIWL $\bar{X}$ chart.



In practice, only one set of the warning limits may be shown anywhere on the chart within the control limits to represent the two sets in order to avoid the complexity in the administration. Suppose each warning limit of this set is at a distance of w units from the centerline. When $w(i) = w_j$, j = 1, 2, plot Z_i anywhere within $[-w, w]$,$(-L, -w)$, and (w, L), respectively, when it is within $[-w_j, w_j]$, $(-L, -w_j)$, and (w_j, L). Note that when $w_1 = w_2$, a VSIWL $\bar{X}$ chart is a VSI $\bar{X}$ chart. In the next section, expressions for performance measures for a VSIWL $\bar{X}$ chart are derived.

Performance measures: The appropriate measures of statistical performance of a VSIWL $\bar{X}$ chart are the steady-state average time to signal (SSATS) and the average number of samples to signal (ANSS). SSATS is the expected value of the time between a shift that occurs at some random time after the process starts and the time the chart signals while ANSS is the expected value of the number of samples taken from a shift to the time the chart signals. The administrative performance can be measured through average number of switches to signal (ANSW). ANSW is the expected value of the number of switches between two sampling interval lengths from a shift to the signal.

Let $SSATS_\delta$, $ANSS_\delta$, and $ANSW_\delta$ be the SSATS, ANSS, and ANSW, respectively of a control chart when the process mean has shifted from μ_0 to $\mu_1 = \mu_0 + \delta\sigma$. In the following, first the expressions for $SSATS_\delta$ and $ANSS_\delta$ are derived using a Markov chain approach. Brook and Evans (1972) were the first to use this approach to find the average run length of a control chart. Henceforth, the i^{th} trial refers to the i^{th} trial after a shift when $i > 0$ and the last trial before the $(i+1)^{st}$ trial when $i \leq 0$.Also, Z_i refers to the sample point corresponding to the i^{th} trial.

Define the three states 1, 2, and 3 of the Markov Chain corresponding to whether a sample point is plotted in I_1, I_2 and $I_3 = (-\infty, -L] \cup [L, \infty)$, respectively. State 3 is the absorbing state, as the process of taking samples is restarted when a sample point falls in region I_3. The transition probability matrix is given by

$$\mathbf{P}^{\boldsymbol{\delta}} = \begin{bmatrix} p_{11}^{\delta} & p_{12}^{\delta} & p_{13}^{\delta} \\ p_{21}^{\delta} & p_{22}^{\delta} & p_{23}^{\delta} \\ 0 & 0 & 1 \end{bmatrix},$$

Where p_{jk}^{δ} is the transition probability that j is the prior state and k is the current state, when the process mean has shifted by $\delta\sigma$.

For example,

$$p_{12}^{\delta} = \Pr_{\delta}[Z_i \in I_2 \mid Z_{i-1} \in I_1]$$
$$= \Pr_{\delta}[Z_i \in I_2 \mid w(i) = w_1]$$
$$= \Pr_{\delta}[-L < Z_i < -w_1] + P[w_1 < Z_i < L]$$
$$= \Phi(-w_1 - \sqrt{n}\delta) - \Phi(-L - \sqrt{n}\delta) + \Phi(L - \sqrt{n}\delta) - \Phi(w_1 - \sqrt{n}\delta),$$

Where $\Phi(\cdot)$ is the cumulative distribution function of standard normal variate.

Then, $SSATS_{\delta}$ and $ANSS_{\delta}$ are given by

$SSATS_{\delta} = \boldsymbol{b}'(\mathbf{I} - \mathbf{P}_1^{\delta})^{-1}\boldsymbol{t} - E(U)$ (1)

And $ANSS_{\delta} = \boldsymbol{b}'(\mathbf{I} - \mathbf{P}_1^{\delta})^{-1}\mathbf{1}$,

Where **I** is the identity matrix of order 2, $\mathbf{P}_1^{\delta}$ is the sub matrix of $\mathbf{P}^{\delta}$ that contains the probabilities associated with the transient states only, $\boldsymbol{t}' = (t_1, t_2), \mathbf{1}' = (1, 1)$, and $\boldsymbol{b}' = (b_1, b_2)$, b_j being the conditional probability that Z_0 falls in I_j given that it falls within the control limits, $j = 1, 2$. We note that $b_2 = 1 - b_1$. The Expression for b_1 is derived in appendix A.

E(*U*) in equation (1) is the expected value of the time *U* between the 0th trial and the shift. Assuming that an assignable cause of a process shift occurs according to a Poisson process, it can be shown that $E(U) = E[t(1)]/2$. Hence, $SSATS_{\delta} = \boldsymbol{b}'(\mathbf{I} - \mathbf{P}_1^{\delta})^{-1}\boldsymbol{t} - E[t(1)]/2$.

Now, to derive the expression for $ANSW_{\delta}$, let O_i be the number of switches between two sampling interval lengths following the i^{th} trial until the signal, $i = 1, 2, \ldots$ Further, let

$o_{j.i}^{\delta} = E_{\delta}(O_i \mid Z_{i-1} \in I_j)$, $i = 1, 2, \ldots, j = 1, 2$.

Then, the expression for $ANSW_{\delta}$ is given by

$ANSW_{\delta} = E_{\delta}[O_1] = b_1 o_{1.1}^{\delta} + b_2 o_{2.1}^{\delta} = \boldsymbol{b}'\mathbf{O}_1^{\delta}$,

Where, $\mathbf{O}_s^{\delta} = (o_{1.s}^{\delta}, o_{2.s}^{\delta})'$, $s = 1, 2, \ldots$

The expression for $\mathbf{O}_1^{\delta}$ is derived in appendix B.

Alternatively, the expression for $ANSW_{\delta}$ can also be obtained using the Markov Chain approach. For, let

$$Y_i = \begin{cases} 1, \text{ if } (Z_{i-1} \in I_1, Z_i \in I_2) \\ 2, \text{ if } (Z_{i-1} \in I_2, Z_i \in I_1) \\ 3, \text{ if } (Z_{i-1} \in I_1, Z_i \in I_1) \\ 4, \text{ if } (Z_{i-1} \in I_2, Z_i \in I_2) \\ 5, \text{ if } |Z_i| > L \end{cases}, i = 1, 2, \ldots$$

It is easy to see that $\{Y_i, i = 1, 2, \ldots\}$ is a Markov Chain with transition probability matrix

$$\mathbf{Q}^{\delta} = \begin{bmatrix} 0 & p_{21}^{\delta} & 0 & p_{22}^{\delta} & p_{23}^{\delta} \\ p_{12}^{\delta} & 0 & p_{11}^{\delta} & 0 & p_{13}^{\delta} \\ p_{12}^{\delta} & 0 & p_{11}^{\delta} & 0 & p_{13}^{\delta} \\ 0 & p_{21}^{\delta} & 0 & p_{22}^{\delta} & p_{23}^{\delta} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then, the expression for $ANSW_{\delta}$ is given by

$ANSW_{\delta} = \boldsymbol{a}'(\mathbf{I}_1 - \mathbf{Q}_1^{\delta})^{-1}\boldsymbol{e}$,

Where, $\mathbf{I}_1$ is the identity matrix of order 4, $\mathbf{Q}_1^{\delta}$ is the sub matrix of $\mathbf{Q}^{\delta}$ that contains the probabilities associated with the transient states only, $\boldsymbol{e} = (1, 1, 0, 0)'$, and $\boldsymbol{a} = (a_1, a_2, a_3, a_4)'$, a_j being the initial probability of state *j*, $j = 1, 2, 3, 4$, given by

$$a_j = \Pr_{\delta}[Y_1 = j] = \begin{cases} b_1 p_{12}^{\delta}, & j = 1 \\ b_2 p_{21}^{\delta}, & j = 2 \\ b_1 p_{11}^{\delta}, & j = 3 \\ b_2 p_{22}^{\delta}, & j = 4 \end{cases}.$$

Results and discussion

Performance evaluation of VSIWL $\bar{X}$ charts: The performances of VSIWL $\bar{X}$ charts are evaluated by comparing that with that of VSI, VSI (1, 3), and VSI (2, 3) $\bar{X}$ charts, where VSI (*k*, *m*) $\bar{X}$ charts refers to the VSI $\bar{X}$ charts with runs rule (*k*, *m*) for switching between sampling interval lengths. When the successive *m* sample points before the i^{th} trial fall within the control limits, rule (*k*, *m*) chooses sampling interval length t_1 for the i^{th} trial if among those *m* sample points, the number of sample points falling in each warning region is less than *k*, otherwise it chooses the sampling interval length t_2. See Mahadik (2011a) for the details of VSI (*k*, *m*) $\bar{X}$ charts.

Among various runs rules considered by Mahadik (2011a), runs rule (1, 3) reduces the ANSW values of VSI $\bar{X}$ charts the most without affecting their SSATS values for small to large shifts in the process mean. Further, runs rule (2, 3) significantly reduces both, the ANSW values for shifts of all sizes and SSATS values for small shifts without affecting that for large shifts. Hence, the VSI $\bar{X}$ charts with these runs rules are chosen for comparison.

Table 1. The SSATS values of the matched VSIWL, VSI, VSI (1, 3), and VSI (2, 3) $\bar{X}$ charts.

Chart	w_1	w_2	SSATS for a shift of size									
			0	0.25σ	0.5σ	0.75σ	1σ	1.5σ	2σ	2.5σ	3σ	4σ
Case 1: n= 1, t_1 = 2, t_2 = 0.2, L = 3												
VSIWL	1.40	0.16	369.90	267.80	129.49	55.91	24.35	5.79	2.10	1.11	0.75	0.54
VSIWL	2.20	0.03	369.90	259.80	118.03	49.16	22.06	6.42	2.72	1.45	0.92	0.57
VSI	0.59		369.90	274.69	142.14	66.98	31.51	7.67	2.40	1.11	0.73	0.54
VSI (1, 3)	1.18		369.90	268.63	130.38	55.85	23.65	5.23	1.93	1.06	0.74	0.54
VSI (2, 3)	0.39		369.89	267.35	128.13	54.04	22.70	5.30	2.23	1.37	0.98	0.63
Case 2: n = 2, t_1 = 1.8, t_2 = 0.4, L = 3												
VSIWL	1.30	0.18	369.90	209.38	71.76	24.76	9.78	2.49	1.07	0.68	0.55	0.50
VSIWL	2.00	0.04	369.90	201.66	65.60	22.89	9.66	2.74	1.18	0.71	0.56	0.50
VSI	0.56		369.90	216.33	79.68	29.08	11.50	2.62	1.06	0.67	0.55	0.50
VSI (1, 3)	1.16		369.90	209.20	70.82	23.72	9.17	2.40	1.07	0.68	0.55	0.50
VSI (2, 3)	0.38		369.89	207.90	69.51	23.23	9.18	2.64	1.27	0.79	0.59	0.50
Case 3: n = 3, t_1 = 1.2, t_2 = 0.6, L = 3												
VSIWL	1.60	0.27	369.90	175.09	50.91	16.48	6.50	1.73	0.80	0.56	0.51	0.50
VSIWL	2.10	0.08	369.90	172.12	49.12	16.25	6.67	1.83	0.82	0.57	0.51	0.50
VSI	0.96		369.90	179.09	54.71	18.20	7.03	1.73	0.79	0.56	0.51	0.50
VSI (1, 3)	1.52		369.90	175.73	51.36	16.53	6.46	1.72	0.80	0.56	0.51	0.50
VSI (2, 3)	0.64		369.90	173.64	49.67	15.97	6.42	1.86	0.88	0.59	0.51	0.50
Case 4: n = 4, t_1 = 1.5, t_2 = 0.5, L = 3												
VSIWL	1.40	0.20	369.90	140.23	32.02	9.14	3.51	1.03	0.60	0.51	0.50	0.50
VSIWL	1.80	0.09	369.90	136.30	30.59	9.05	3.60	1.07	0.60	0.51	0.50	0.50
VSI	0.67		369.90	147.14	36.19	10.31	3.71	1.02	0.59	0.51	0.50	0.50
VSI (1, 3)	1.26		369.90	140.48	31.60	8.83	3.41	1.03	0.60	0.51	0.50	0.50
VSI (2, 3)	0.45		369.92	138.65	30.74	8.80	3.58	1.18	0.65	0.52	0.50	0.50
Case 5: n = 5, t_1 = 1.4, t_2 = 0.3, L = 3												
VSIWL	1.50	0.29	369.90	115.18	20.52	4.94	1.86	0.69	0.52	0.50	0.50	0.50
VSIWL	1.90	0.12	369.90	110.48	19.12	4.98	2.00	0.73	0.53	0.50	0.50	0.50
VSI	0.91		369.90	122.74	24.70	5.98	2.00	0.68	0.52	0.50	0.50	0.50
VSI (1, 3)	1.47		369.90	115.52	20.38	4.78	1.83	0.69	0.52	0.50	0.50	0.50
VSI (2, 3)	0.60		369.91	111.77	18.84	4.71	2.05	0.85	0.56	0.50	0.50	0.50

Table 2. The ANSW values of the matched VSIWL, VSI, VSI (1, 3), and VSI (2, 3) $\bar{X}$ charts.

Chart	w_1	w_2	ANSW for a shift of size									
			0	0.25σ	0.5σ	0.75σ	1σ	1.5σ	2σ	2.5σ	3σ	4σ
Case 1: n = 1, t_1 = 2, t_2 = 0.2, L = 3												
VSIWL	1.40	0.16	52.29	40.26	22.60	11.49	5.65	1.42	0.52	0.31	0.21	0.07
VSIWL	2.20	0.03	8.27	6.57	3.91	2.12	1.15	0.47	0.30	0.23	0.16	0.06
VSI	0.59		182.42	137.53	73.98	36.49	17.74	4.18	1.10	0.43	0.24	0.07
VSI (1, 3)	1.18		77.77	59.09	32.00	15.38	6.95	1.37	0.49	0.32	0.22	0.07
VSI (2, 3)	0.39		108.88	80.05	39.82	17.18	7.00	1.31	0.55	0.36	0.22	0.05
Case 2: n = 2, t_1 = 1.8, t_2 = 0.4, L = 3												
VSIWL	1.30	0.18	60.61	37.18	14.66	5.32	1.95	0.45	0.24	0.12	0.05	0.00
VSIWL	2.00	0.04	13.59	8.72	3.66	1.46	0.67	0.31	0.20	0.11	0.04	0.00
VSI	0.56		180.93	107.51	40.55	14.51	5.17	0.80	0.27	0.13	0.05	0.00
VSI (1, 3)	1.16		77.91	46.75	17.25	5.51	1.70	0.41	0.25	0.13	0.05	0.00
VSI (2, 3)	0.38		108.28	61.04	19.55	5.44	1.62	0.47	0.26	0.11	0.03	0.00
Case 3: n = 3, t_1 = 1.2, t_2 = 0.6, L = 3												
VSIWL	1.60	0.27	52.79	29.67	10.84	3.54	1.25	0.42	0.20	0.06	0.01	0.00
VSIWL	2.10	0.08	16.31	9.82	3.82	1.40	0.67	0.34	0.17	0.05	0.01	0.00
VSI	0.96		164.18	85.94	29.82	10.12	3.42	0.60	0.22	0.06	0.01	0.00
VSI (1, 3)	1.52		62.27	34.70	12.58	3.97	1.29	0.43	0.21	0.06	0.01	0.00
VSI (2, 3)	0.64		95.60	48.90	14.50	3.70	1.16	0.42	0.15	0.03	0.00	0.00
Case 4: n = 4, t_1 = 1.5, t_2 = 0.5, L = 3												
VSIWL	1.40	0.20	58.82	25.95	6.72	1.70	0.60	0.24	0.08	0.01	0.00	0.00
VSIWL	1.80	0.09	25.62	11.85	3.21	0.94	0.45	0.22	0.07	0.01	0.00	0.00
VSI	0.67		184.70	76.37	19.08	4.68	1.26	0.27	0.08	0.01	0.00	0.00
VSI (1, 3)	1.26		76.21	32.51	7.60	1.59	0.55	0.25	0.08	0.01	0.00	0.00
VSI (2, 3)	0.45		109.31	41.28	7.69	1.48	0.60	0.24	0.06	0.01	0.00	0.00
Case 5: n = 5, t_1 = 1.4, t_2 = 0.3, L = 3												
VSIWL	1.50	0.29	61.72	25.44	6.18	1.47	0.58	0.22	0.04	0.00	0.00	0.00
VSIWL	1.90	0.12	25.80	11.40	2.91	0.86	0.46	0.20	0.04	0.00	0.00	0.00
VSI	0.91		170.96	64.84	15.67	3.72	1.03	0.24	0.04	0.00	0.00	0.00
VSI (1, 3)	1.47		65.75	26.94	6.34	1.37	0.55	0.22	0.04	0.00	0.00	0.00
VSI (2, 3)	0.60		99.72	35.45	6.31	1.23	0.57	0.17	0.02	0.00	0.00	0.00

The four charts mentioned above are designed such that their in-control statistical performances are matched. This is done by keeping the design parameters n, t_1, t_2, and L of all the charts the same and choosing the warning limits of each chart such that E[t(1)] = t_0 holds for each chart, where t_0 is some suitable constant.

As a VSIWL $\bar{X}$ chart has two sets of warning limits, by fixing one of them this condition uniquely determines the other. By fixing w_1, we get

$$w_2 = \Phi^{-1}\left\{\frac{(t_0 - t_2)[2\Phi(L) - 1 - 2\Phi(w_1)] + t_1 - t_2}{2(t_1 - t_0)}\right\},$$

or by fixing w_2, we get

$$w_1 = \Phi^{-1}\left\{\Phi(L) + \frac{[2\Phi(w_2) - 1](t_0 - t_1)}{2(t_0 - t_2)}\right\}.$$

In the same way the warning limits of VSI, VSI (1, 3), and VSI (2, 3) $\bar{X}$ charts are determined.

The SSATS and ANSW values of such statistically matched charts are then computed for shifts of various sizes. Tables 1 and 2, respectively, show these values for five different sets of the matched charts. Note that as all the charts in a set use the same values of L and n, their ANSS values will be the same. Hence, ANSS is not a relevant measure to compare the statistical performances of the charts. Computations of the SSATS and ANSW values indicate the following facts in general.

If the warning limits of a VSIWL $\bar{X}$ chart are chosen such that its SSATS values for the large shifts match that of a VSI $\bar{X}$ chart then for the small to moderate shifts, its SSATS values are slightly smaller than that of the VSI $\bar{X}$ chart and are similar to that of a VSI (1, 3) $\bar{X}$ chart. Further, the ANSW values of a VSIWL $\bar{X}$ chart are significantly smaller than that of the VSI and VSI (1, 3) $\bar{X}$ charts.

On the other hand, if the warning limits of a VSIWL $\bar{X}$ chart are chosen such that its SSATS values for the large shift are very slightly larger than that of a VSI $\bar{X}$ chart then for the small to moderate shifts, its SSATS values are significantly smaller than that of the VSI and VSI (1, 3) $\bar{X}$ charts and are similar to that of a VSI (2, 3) $\bar{X}$ chart. Besides, its ANSW values are dramatically smaller than that of the other charts and are about 5 to 15% of that of a VSI $\bar{X}$ chart.

Example

The statistically matched VSIWL, VSI (1, 3), VSI (2, 3), and VSI $\bar{X}$ charts with the design parameters, viz., n= 4, t_1 = 1.5 hours, t_2 = 0.5 hour, and L = 3 are implemented simultaneously and independently for a process. The process is initially in control when the implementation of the charts is started and a shift of size 0.75σ occurs at 5 hours after that. Table 3 shows the sample means taken for the two VSIWL $\bar{X}$ charts along with the corresponding times, sampling interval lengths, and the warning limits used. Table 4 shows the same for VSI (1, 3), VSI (2, 3), and VSI $\bar{X}$ charts. The pair (t_2, w_2) is used for the first trials for the VSIWL charts and the pairs for the subsequent trials are chosen according to the rule of the charts. Similarly, sampling interval length t_2 is used for the first trial for the VSI chart and for the first three trials for the VSI (1, 3) and VSI (2, 3) charts. Sampling interval lengths for the subsequent trials for these charts are chosen according to the respective rules of the charts. Table 5 shows the performances of the five charts which clearly demonstrate the superiority of the VSIWL charts.

Table 3. The details of the VSIWL $\bar{X}$ charts for the process in the example.

VSIWL with w_1 = 1.4, w_2 = 0.2			VSIWL with w_1 = 1.8, w_2 = 0.09		
Time in hours	($t(i)$, $w(i)$)	Zs	Time in hours	($t(i)$, $w(i)$)	Zs
0.5	(0.5, 0.2)	-0.42	0.5	(0.5, 0.09)	1.01
1	(0.5, 0.2)	-1.35	1	(0.5, 0.09)	0.14
1.5	(0.5, 0.2)	1.10	1.5	(0.5, 0.09)	0.50
2	(0.5, 0.2)	-1.43	2	(0.5, 0.09)	2.25
2.5	(0.5, 0.2)	1.33	2.5	(0.5, 0.09)	0.80
3	(0.5, 0.2)	0.46	3	(0.5, 0.09)	0.51
3.5	(0.5, 0.2)	-0.49	3.5	(0.5, 0.09)	-0.66
4	(0.5, 0.2)	0.36	4	(0.5, 0.09)	2.88
4.5	(0.5, 0.2)	0.97	4.5	(0.5, 0.09)	-0.83
5	(0.5, 0.2)	1.88	5	(0.5, 0.09)	2.28
5.5	(0.5, 0.2)	1.38	5.5	(0.5, 0.09)	2.82
6	(0.5, 0.2)	2.46	6	(0.5, 0.09)	3.26
6.5	(0.5, 0.2)	1.48			
7	(0.5, 0.2)	0.61			
7.5	(0.5, 0.2)	2.37			
8	(0.5, 0.2)	1.65			
8.5	(0.5, 0.2)	1.26			
9	(0.5, 0.2)	3.47			

Table 4.The details of the VSI (1, 3), VSI (2, 3), and VSI $\bar{X}$ charts for the process in the example.

VSI (1, 3) with w =1.26			VSI (2, 3) with w = 0.45			VSI with w = 0.67		
Time in hours	$t(i)$	Zs	Time in hours	$t(i)$	Zs	Time in hours	$t(i)$	Zs
0.5	0.5	0.16	0.5	0.5	-0.80	0.5	0.5	-1.08
1	0.5	-0.16	1	0.5	0.72	1	0.5	1.04
1.5	0.5	-0.51	1.5	0.5	-0.66	1.5	0.5	-0.64
3	1.5	-1.09	2	0.5	1.36	3	1.5	2.30
4.5	1.5	-0.22	2.5	0.5	-0.97	3.5	0.5	-0.11
			3	0.5	-0.20			
6	1.5	0.85	4.5	1.5	-0.15	4	1.5	1.33
7.5	1.5	0.91				4.5	0.5	1.42
9	1.5	1.53	6	1.5	2.08	5	0.5	1.18
9.5	0.5	1.70	7.5	1.5	3.13	5.5	0.5	-0.54
10	0.5	3.59				7	1.5	2.35
						7.5	0.5	0.80
						8	0.5	0.58
						9.5	1.5	2.37
						10	0.5	0.72
						10.5	0.5	1.23
						11	0.5	3.08

Table 5. Performances of the charts in the example.

Chart	Time from the shift to the signal	Number of switches during in-control period	Total number of switches until the signal
VSIWL with w_1 = 1.4, w_2 = 0.2	4 hours	0	0
VSIWL with w_1 = 1.8, w_2 = 0.09	1 hour	0	0
VSI (1, 3)	5 hours	1	2
VSI (2, 3)	2.5 hours	1	1
VSI	6 hours	2	8

Conclusion

The idea of variable warning limits is introduced for VSI $\bar{X}$ charts. Expressions for the performance measures, viz., SSATS, ANSS and ANSW for VSIWL $\bar{X}$ charts are developed. The methods presented are general and can be applied to other Shewhart control charts. The effects of variable warning limits on the performances of the charts are evaluated by comparing the performances of VSIWL $\bar{X}$ charts with that of VSI $\bar{X}$ charts with and without runs rules for switching between sampling interval lengths. It is observed that the variable warning limits dramatically reduce the ANSW values of the charts. The idea is even superior to that of runs rules for switching between sampling interval lengths for reducing the ANSW values. Also, it significantly reduces the SSATS values of the charts in detecting small to moderate shifts in the process mean without significantly affecting that in detecting large shifts. It would be interesting to study the application of variable warning limits to the other adaptive control charts.

References

1. Amin, R.W. and Hemasinha, R. 1993. The switching behavior of $\bar{X}$ charts with variable sampling intervals. *Commun. Stat. Theory Meth.*22: 2081-2102.
2. Amin, R.W. and Letsinger, W.C. 1991. Improved switching rules in control procedures using variable sampling intervals. *Commun. Stat. Simul.* 20: 205-230.
3. Amin, R.W. and Widmaier, O. 1999. Sign control charts with variable sampling intervals. *Commun. Stat. Theory Meth.* 28: 1961-1985.

4. Aparasi, F. and Haro, C. 2003. A comparison of T^2 control charts with variable sampling schemes as opposed to MEWMA chart. *Int. J. Prod. Res.* 41: 2169-2182.
5. Brook, D. and Evans, D.A. 1972. An approach to the probability distribution of CUSUM run length. *Biometrik.* 59: 539-549.
6. Celano, G. 2009. Robust design of adaptive control charts for manual manufacturing/inspection workstations. *J. Appl. Stat.* 36: 181-203.
7. Celano, G., Costa, A., and Fichera, S. 2006. Statistical design of variable sample size and sampling interval $\bar{X}$ control charts with run rules. *Int. J. Adv. Manufac. Tech.* 28: 966-977.
8. Chen, Y.K. 2007. Adaptive sampling enhancement for Hotelling's T^2 charts. *Eur. J. Oper. Res.* 178: 841-857.
9. Chen, Y.K., Liao, H.C. and Chang, H.H. 2011. Re-evaluation of adaptive $\bar{X}$ control charts: A cost-effectiveness perspective. *Int. J. Innov. Comp. Info. Cont.* 7: 1229-1242.
10. Costa, A.F.B. 1994. $\bar{X}$ charts with variable sample size. *J. Qual. Technol.* 26: 155-163.
11. Costa, A.F.B. 1999b. Joint $\bar{X}$ and *R* charts with variable sample sizes and sampling intervals. *J. Qual. Technol.* 31: 387-397.
12. Costa, A.F.B. and Rahim, M.A. 2001. Economic design of $\bar{X}$ charts with variable parameters: The Markov chain approach. *J. Appl. Stat.* 28: 875-885.
13. Costa, A.F.B.1999a. $\bar{X}$ charts with variable parameters. *J. Qual. Technol.* 31: 408-416.
14. Dai, Y., Luo, Y., Li, Z., and Wang, Z.2011. A new adaptive CUSUM control chart for detecting the multivariate process mean. *Qual. Reliab. Engg. Int.* 27: 877-884.
15. De Magalhaes, M.S., Costa, A.F.B. and MouraNeto, F.D. 2009. A hierarchy of adaptive $\bar{X}$ control charts. *Int. J. Prod. Econ.*119: 271-283.
16. Epprecht, E.K., Costa, A.F.B. and Mendes, F.C.T. 2003. Adaptive control charts for attributes. *IIE Trans.*35: 567-582.
17. Epprecht, E.K., Simoes, B.F.T. and Mendes, F.C.T. 2010. A variable sampling interval EWMA chart for attributes. *Int. J. Adv. Manufac. Tech.* 49: 281-292.
18. Faraz, A. and Moghadam, M. B. 2009. Hotelling's T^2 control chart with two adaptive sample sizes. *Qual. Quant.* 43: 903-912.
19. Faraz, A. and Saniga, E. 2011.A unification and some corrections to Markov chain approaches to develop variable ratio sampling scheme control charts. *Stat Pap.* 52: 799-811.
20. Jensen, W.A., Bryce, G.R. and Reynolds, M.R. 2008. Design issues for adaptive control charts. *Qual. Reliab. Engg. Int.* 24: 429-445.
21. Jiang, W., Shu, L., and Apley, D. 2008.Adaptive CUSUM procedures with EWMA-based shift estimators. *IIE Trans*. 40: 992-1003.
22. Kooli, I. and Limam, M. 2011. Economic design of an attribute *np* control chart using a variable sample size. *Sequential Anal.* 30: 145-159.
23. Lee, P.H. 2011. Adaptive R charts with variable parameters. *Comput. Stat. Data Anal.* 55: 2003-2010.
24. Li, Z. and Wang, Z. 2010. Adaptive CUSUM of Q chart. *Int. J. Prod. Res*. 48: 1287-1301.
25. Luo, Y., Li, Z. and Wang, Z. 2009. Adaptive CUSUM control chart with variable sampling intervals. *Comput. Stat. Data Anal*. 53: 2693-2701.
26. Mahadik, S.B. 2012a. Exact results for variable sampling interval Shewhart control charts with runs rules for switching between sampling interval lengths. *Commun. Stat. Theory Meth*. 41: 4453-4469.
27. Mahadik, S.B. 2012b. Variable sampling interval Hotelling's T^2 charts with runs rules for switching between sampling interval lengths. *Qual. Reliab. Engg. Int*. 28: 131-140.
28. Mahadik, S.B. 2013. Variable sample size and sampling interval $\bar{X}$ charts with runs rules for switching between sample sizes and sampling interval lengths. *Qual. Reliab. Engg. Int.* 29: 63-76.
29. Mahadik, S.B. and Shirke, D.T. 2007a.On superiority of a variable sampling interval control chart. *J. Appl. Stat.* 34: 443-458.
30. Mahadik, S.B. and Shirke, D.T. 2007b. Economic design of a modified variable sample size and sampling interval $\bar{X}$ chart. *Econ. Qual. Control*. 22: 273-293.
31. Mahadik, S.B. and Shirke, D.T. 2009. A special variable sample size and sampling interval $\bar{X}$ chart. *Commun. Stat. Theory.* 38: 1284-1299.
32. Mahadik, S.B. and Shirke, D.T. 2011. A special variable sample size and sampling interval Hotelling's T^2 chart. *Int. J. Adv. Manufac. Tech.* 53: 379-384.
33. Nenes, G. 2011. A new approach for the economic design of fully adaptive control charts. *Int. J. Prod. Econ.* 131: 631-642.
34. Prabhu, S.S., Montgomery, D.C. and Runger, G.C. 1994. A combined adaptive sample size and sampling interval $\bar{X}$ control scheme. *J. Qual. Technol.* 26: 164-176.
35. Prabhu, S.S., Runger, G.C. and Keats, J.B. 1993. An adaptive sample size $\bar{X}$ chart. *Int. J. Prod. Res.* 31: 2895-2909.
36. Reynolds, M.R., Jr. and Stoumbos, Z.G. 2001. Monitoring the process mean and variance using individual observations and variable sampling intervals. *J. Qual. Technol.* 33: 181-205.
37. Reynolds, M.R., Jr., Amin, R.W., Arnold, J.C. and Nachlas, J.A. 1988. $\bar{X}$ charts with variable sampling interval. *Technometric.* 30: 181-192.
38. Shi, L., Zou, C., Wang, Z., and Kapur, K. 2009. A new variable sampling control scheme at fixed times for monitoring the process dispersion. *Qual. Reliab. Engg. Int.* 25: 961-972.
39. Shu, L., Jiang, W. and Yeung, H.F. 2010.An adaptive CUSUM procedure for signaling process variance changes of unknown sizes. *J. Qual. Technol.* 42: 69-85.
40. Tagaras, G. 1998. A survey of recent developments in the design of adaptive control charts. *J. Qual. Technol.* 30: 212-231.
41. Wu, Z., Tian, Y. and Zhang, S. 2005. Adjusted-loss-function charts with variable sample sizes and sampling intervals. *J. Appl. Stat.* 32: 221-242.
42. Wu, Z., Wang, P. and Wang, Q. 2009. A loss function-based adaptive control chart for monitoring the process mean and variance. *Int. J. Adv. Manufac. Tech*. 40: 948-959.
43. Wu, Z., Zhang, S. and Wang, P.H. 2007. A CUSUM scheme with variable sample sizes and sampling intervals for monitoring the process mean and variance. *Qual. Reliab. Engg. Int.* 23: 157-170.
44. Yang, S.F. and Su, H.C. 2007. Adaptive sampling interval cause-selecting control charts. *Int. J. Adv. Manufac. Tech*. 31: 1169-1180.
45. Yu, F.J. and Hou, J.L. 2006. Optimization of design parameters for $\bar{X}$ control charts with multiple assignable causes. *J. Appl. Stat*. 33: 279-290.
46. Zimmer, L.S., Montgomery, D. C. and Runger, G.C. 1998. Evaluation of a three-state adaptive sample size $\bar{X}$ control chart. *Int. J. Prod. Res.* 36: 733-743.

Journal of Modern Applied Statistical Methods

Volume 13 | Issue 2 Article 10

11-2014

Comparison of Estimators in GLM with Binary Data

D. M. Sakate
Shivaji University, Kolhapur, India, dms.stats@gmail.com

D. N. Kashid
Shivaji University, Kolhapur, Maharashtra, India., dnk_stats@unishivaji.ac.in

Journal of Modern Applied Statistical Methods
November 2014, Vol. 13, No. 2, 185-200.

ISSN 1538 – 9472

Comparison of Estimators in GLM with Binary Data

D. M. Sakate
Shivaji University
Kolhapur, India

D. N. Kashid
Shivaji University
Kolhapur, India

Maximum likelihood estimates (MLE) of regression parameters in the generalized linear models (GLM) are biased and their bias is non negligible when sample size is small. This study focuses on the GLM with binary data with multiple observations on response for each predictor value when sample size is small. The performance of the estimation methods in Cordeiro and McCullagh (1991), Firth (1993) and Pardo et al. (2005) are compared for GLM with binary data using an extensive Monte Carlo simulation study. Performance of these methods for three real data sets is also compared.

Keywords: Binomial regression, modified score function, bias corrected MLE, Minimum ϕ-divergence estimation, Monte Carlo Simulation

Introduction

Generalized linear models (GLM) are frequently used to model small to medium size data. In case of binomial distributed response, logistic regression finds application to model the relationship between response and predictors. Maximum likelihood estimation (MLE) is usually used to fit a logistic regression model. It is well known that under certain regularity conditions, MLE of regression coefficients are consistent and asymptotically normal. However, for finite sample sizes, MLE tend to overestimate with an absolute bias that tends to increase with the magnitude of the parameter and with the ratio of the number of parameters to the number of observations. The bias in MLE decreases with the sample size and goes to zero as sample size tends to infinity. See Byth and McLachlan, (1978), Anderson and Richardson (1979), McLachlan (1980), Pike et al. (1980), Breslow (1981) and Hauck (1984) for the details. As a consequence, methods taking care of bias were explored. Jackknifed MLE and its versions and methods based on approximation

Dr. Sakate is an Assistant Professor in the Department of Statistics. Email him at dms.stats@gmail.com. Dr. Kashid is a Professor in the Department of Statistics. Email him at dnk_stats@unishivaji.ac.in.

Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Variable selection via penalized minimum φ-divergence estimation in logistic regression

D.M. Sakate[a] & D.N. Kashid[a]
[a] Department of Statistics, Shivaji University, Kolhapur, Maharashtra, India
Published online: 02 Dec 2013.

To cite this article: D.M. Sakate & D.N. Kashid (2014) Variable selection via penalized minimum φ-divergence estimation in logistic regression, Journal of Applied Statistics, 41:6, 1233-1246, DOI: 10.1080/02664763.2013.864262

To link to this article: http://dx.doi.org/10.1080/02664763.2013.864262

Journal of Applied Statistics, 2014
Vol. 41, No. 6, 1233–1246, http://dx.doi.org/10.1080/02664763.2013.864262

Taylor & Francis
Taylor & Francis Group

Variable selection via penalized minimum ϕ-divergence estimation in logistic regression

D.M. Sakate* and D.N. Kashid

Department of Statistics, Shivaji University, Kolhapur, Maharashtra, India

(*Received 9 November 2012; accepted 6 November 2013*)

We propose penalized minimum ϕ-divergence estimator for parameter estimation and variable selection in logistic regression. Using an appropriate penalty function, we show that penalized ϕ-divergence estimator has oracle property. With probability tending to 1, penalized ϕ-divergence estimator identifies the true model and estimates nonzero coefficients as efficiently as if the sparsity of the true model was known in advance. The advantage of penalized ϕ-divergence estimator is that it produces estimates of nonzero parameters efficiently than penalized maximum likelihood estimator when sample size is small and is equivalent to it for large one. Numerical simulations confirm our findings.

Keywords: ϕ-divergence; logistic regression; penalized MLE; SCAD; variable selection

1. Introduction

Logistic regression is one of the widely used generalized linear models (GLM) to describe the binary data. In logistic regression model, inference is done based on but not limited to likelihood. Minimum divergence estimators or minimum distance estimators are also used to model the discrete data [24]. Read and Cressie [25] and Pardo [22] outline the use and importance of the ϕ-divergence measures in Statistics. Minimum ϕ-divergence estimator [20] in logistic regression emerged as an attractive alternative to maximum likelihood estimator (MLE) when sample size is small. Based on this fact, Pardo and Pardo [21] introduced a method for variable selection using ϕ-divergence statistic. This method is a two-stage method which requires fitting and testing of several models to arrive at the best sub model.

It belongs to the broad class of sequential procedures for variable selection. It is well known that such procedures are time consuming and costly. Methods which perform estimation as well as variable selection simultaneously have become a good choice to overcome this difficulty. Penalized regression has evolved as a powerful tool to solve the problem of estimation and variable selection simultaneously. Anderson and Blair [4] introduced penalized logistic regression for the first time. Bridge regression [13] and least absolute shrinkage selection operator (LASSO) [28] are the members of class of penalized least-squares methods. l_1 type penalty of the LASSO has

*Corresponding author: Email: dms.stats@gmail.com

also found applications in logistic regression [14,26,27]. Fan and Li [11] extended the idea of penalized least squares to likelihood-based models in various statistical contexts. They introduced a penalty function called smoothly clipped absolute deviation penalty (SCAD).

In this article, we propose a penalized minimum ϕ-divergence estimator to obtain estimates of regression coefficients and simultaneous variable selection in logistic regression. We showed that this estimator is consistent, asymptotic normal and possesses oracle property. We used SCAD for the purpose of penalization. Our simulation study indicates that the proposed estimator performs better than SCAD penalized MLE.

The remaining article is organized as follows. Section 2 describes ϕ-divergence estimation in logistic regression. In Section 3, penalized minimum ϕ-divergence estimator is defined. Its sampling properties are also described in this section. Section 4 deals with simulation study to compare the performance of proposed method with existing ones. A real data application is also provided. This article ends with discussion in Section 5.

2. ϕ-divergence estimation in logistic regression

Let Z be a response binary random variable taking value 1 or 0, generally referred to as 'success' or 'failure', respectively. Let k explanatory variables $\boldsymbol{x} \in \mathbb{R}^k$ are observed along with the response variable. $\pi(\boldsymbol{x}) = P(Z = 1|\boldsymbol{x} \in \mathbb{R}^k)$ represents the conditional probability, of the value 1 given $\boldsymbol{x} \in \mathbb{R}^k$. Let X be the $N \times (k+1)$ matrix with rows $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ik})$, $i = 1, \ldots, N$ where, $x_{i0} = 1, \forall i$. The logistic regression model is defined by the conditional probability

$$\pi(\boldsymbol{x}_i) = \frac{\exp\left\{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right\}}{1 + \exp\left\{\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right\}}. \tag{1}$$

For more discussion on logistic regression see Hosmer and Lemeshow [17] and Agresti [1].

In laboratory or controlled setting, many individuals share same values for their explanatory variables. In other words, for each value of the explanatory variables there are several observed values of the random variable Z. Our focus is on this situation. The notations described earlier are required to be changed slightly. For this, we follow the notations used in [20]. Let there be I distinct values of $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ik})$, $i = 1, 2, \ldots, I$. We assume that for each $\boldsymbol{x}_i$ we have a binomial random variable $Y_i \equiv \sum_{i=1}^{n_i} Z_i$ with parameters n_i and $\pi(x_i)$. The values $n_{i1}, \ldots, n_{I1}$ are the observed values of the random variables $Y_1, \ldots, Y_I$, representing the number of successes in $n_1, \ldots, n_I$ trials respectively when the explanatory variables are fixed. This divides the entire sample of size N into I subgroups each of size n_i so that $N = \sum_{i=1}^{I} n_i$. Since, Z_i's are independent, Y_i's are also independent. Thus, the likelihood function for the logistic regression model is given by

$$L(\beta_0, \ldots, \beta_k) = \prod_{i=1}^{I} \binom{n_i}{n_{i1}} \pi(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^{n_{i1}} (1 - \pi(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}))^{n_i - n_{i1}}, \tag{2}$$

The MLE, $\hat{\boldsymbol{\beta}}$ is obtained by maximizing almost surely over

$$\Theta = \{\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k) : -\infty < \beta_j < \infty, j = 0, \ldots, k\},$$

the likelihood function given in Equation (2).

For simplicity, we denote by $\pi_{i1} = \pi(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ and $\pi_{i2} = 1 - \pi(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}), n_{i2} = n_i - n_{i1}$. To maximize (2) is equivalent to minimize the Kullback divergence measure between the probability vectors

$$\hat{\boldsymbol{p}} = (\hat{p}_{11}, \hat{p}_{12}, \ldots, \hat{p}_{I1}, \hat{p}_{I2})^{\mathrm{T}} = \left(\frac{n_{11}}{N}, \frac{n_{12}}{N}, \ldots, \frac{n_{I1}}{N}, \frac{n_{I2}}{N}\right)^{\mathrm{T}}$$

and

$$\boldsymbol{p}(\boldsymbol{\beta}) = (p_{11}(\boldsymbol{\beta}), p_{12}(\boldsymbol{\beta}), \ldots, p_{I1}(\boldsymbol{\beta}), p_{I2}(\boldsymbol{\beta}))^{\mathrm{T}} = \left(\pi_{11}\frac{n_1}{N}, \pi_{12}\frac{n_1}{N}, \ldots, \pi_{I1}\frac{n_I}{N}, \pi_{I2}\frac{n_I}{N}\right)^{\mathrm{T}}. \quad (3)$$

MLE for the GLM parameter $\boldsymbol{\beta}$ can be defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta\epsilon\Theta} D_{\text{Kullback}}(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})), \quad (4)$$

where the Kullback divergence measure is given by Kullback [19]

$$D_{\text{Kullback}}(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})) = \sum_{j=1}^{2}\sum_{i=1}^{I} \hat{p}_{ij} \log\left(\frac{\hat{p}_{ij}}{p_{ij}(\boldsymbol{\beta})}\right).$$

This measure is a particular case of the ϕ-divergence defined by Csiszár [8] and Ali and Silvey [2],

$$D_{\phi}(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})) = \sum_{j=1}^{2}\sum_{i=1}^{I} p_{ij}(\boldsymbol{\beta})\phi\left(\frac{\hat{p}_{ij}}{p_{ij}(\boldsymbol{\beta})}\right); \quad \phi\epsilon\Phi, \quad (5)$$

where Φ is the class of all convex functions $\phi(t)$, $t > 0$ and twice differentiable at $t = 1$, such that $\phi(1) = \phi'(1) = 0, \phi''(1) > 0$ and at $t = 0, 0\phi(0/0) = 0$ and $0\phi(p/0) = p\lim_{u\to\infty}\phi(u)/u$. For details, see Vajda [29] and Pardo [22].

Cressie and Read [7] introduced an important family of ϕ-divergences called as the power divergence family,

$$\begin{aligned}
\phi_{\lambda}(t) &= (\lambda(\lambda+1))^{-1}(t^{\lambda+1} - t); \quad \lambda \neq 0,\ \lambda \neq -1,\\
\phi_0(t) &= \lim_{\lambda\to 0}\phi_{\lambda}(t) = t\log(t) - t + 1,\\
\phi_{-1}(t) &= \lim_{\lambda\to -1}\phi_{\lambda}(t) = -\log(t) - t - 1.
\end{aligned} \quad (6)$$

It is interesting to note that

$$D_{\phi_0}(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})) = D_{\text{kullback}}(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})). \quad (7)$$

That is, for $\lambda = 0$, minimum power divergence estimator coincides to MLE. Use of power divergence family in the log linear models has produced good results [6,23].

The minimum ϕ-divergence estimator [20] is given by

$$\hat{\boldsymbol{\beta}}_{\phi} = \arg\min_{\beta\epsilon\Theta} D_{\phi}(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})). \quad (8)$$

To obtain a natural extension of the penalized MLE for a logistic regression model, in this article, we penalize the minimum ϕ-divergence estimator using appropriate penalty. The SCAD penalty proposed by Fan and Li [11] possesses attractive properties like asymptotic unbiasedness, sparsity and oracle property. Also, use of the SCAD penalty has yielded better performance with diverging number of parameters [12], penalized support vector machines [32], high dimensional linear regression models [18] and partially linear models [31]. Hence, we consider SCAD penalty for the purpose of penalization in the next section.

3. Penalized minimum ϕ-divergence estimator and variable selection

In GLM, likelihood-based inference is most common. Consider a data on response variable and covariates $\{(Y_i, \boldsymbol{x}_i)\}$ are collected independently. Let $f_i(g(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}), y_i)$, be the conditional density of Y_i given $\boldsymbol{x}_i$, where g is a known link function. Denote $l_i = \log f_i$, the conditional log likelihood of Y_i. Then the penalized likelihood is

$$\sum_{i=1}^{I} l_i(g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}), y_i) - N\sum_{j=1}^{k} J_\tau(|\beta_j|). \tag{9}$$

where J_τ is the penalty function and τ is the tuning parameter.

Maximizing (9) is equivalent to minimizing

$$-\sum_{i=1}^{I} l_i(g(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}), y_i) + N\sum_{j=1}^{k} J_\tau(|\beta_j|) \tag{10}$$

with respect to $\boldsymbol{\beta}$. Penalized MLE of $\boldsymbol{\beta}$ is obtained by minimizing (10) with respect to $\boldsymbol{\beta}$ for some thresholding parameter τ. Fan and Li [11] demonstrated that the good results can be obtained when SCAD penalty is used in Equation (10). SCAD penalty is continuous and differentiable and is defined by its derivative

$$J'_\tau(\theta) = \tau\left\{I(\theta \le \tau) + \frac{(a\tau - \theta)_+}{(a-1)\tau} I(\theta > \tau)\right\} \text{ for some } a > 2 \quad \text{and} \quad \theta > 0. \tag{11}$$

For simultaneous parameter estimation and variable selection in logistic regression, we define the penalized minimum ϕ-divergence estimator as follows.

DEFINITION 3.1 *Penalized minimum ϕ-divergence estimate of $\boldsymbol{\beta}$ is that value of $\boldsymbol{\beta}$ for which*

$$Q(\boldsymbol{\beta}) = D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})) + N\sum_{j=1}^{k} J_\tau(|\beta_j|) \tag{12}$$

is minimum. As penalization by SCAD results in an estimator with good properties, we use SCAD in Equation (12).

For brevity, we call the resulting estimator as ϕ SCAD estimator in the further discussion. In the following subsection we establish some asymptotic properties of the proposed estimator.

3.1 *Sampling properties and oracle properties*

Assume that X matrix is standardized. Let the parameter vector $\boldsymbol{\beta}$ be partitioned as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^{\mathrm{T}} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$. Similarly, true value of $\boldsymbol{\beta}$ that is $\boldsymbol{\beta}_0$ can be partitioned as $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{k0})^{\mathrm{T}} = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}$. Without loss of generality, assume that $\boldsymbol{\beta}_{20} = 0$. Let $I(\boldsymbol{\beta}_0)$ denotes the Fisher information matrix and $I_1(\boldsymbol{\beta}_{10}, 0)$ be the Fisher information knowing $\boldsymbol{\beta}_{20} = 0$. Let $Y_1, \ldots, Y_I$ be independent binomial variates with parameters n_i and π_{i1}. Since, minimizing (12) is equivalent to maximize

$$M(\boldsymbol{\beta}) = -D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})) - N\sum_{j=1}^{k} J_\tau(|\beta_j|), \tag{13}$$

we state our theorems based on maximization of $M(\boldsymbol{\beta})$.

THEOREM 1 *Let $\phi(t) \in \Phi$. If $\max\{J'_{\tau_N}(|\beta_{j0}|) : \beta_{j0} \neq 0\} \to 0$ then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $M(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(N^{-1/2} + a_N)$.*

Proof Let $\alpha_N = N^{-1/2} + a_N$. To prove Theorem 1, it is equivalent to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left(\operatorname*{Sup}_{\|\boldsymbol{u}\|=C} M(\boldsymbol{\beta}_0 + \alpha_N \boldsymbol{u}) < M(\boldsymbol{\beta}_0)\right) \geq 1 - \varepsilon. \tag{14}$$

That is, there exists a local maximum in the ball $\{\boldsymbol{\beta}_0 + \alpha_N \boldsymbol{u} : \|\boldsymbol{u}\| \leq C\}$ with probability at least $1 - \varepsilon$. Hence, a local maximizer exists such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\alpha_N)$.

Since, $J_{\tau_N}(0) = 0$, we have

$$\begin{aligned} W_N(\boldsymbol{u}) &= M(\boldsymbol{\beta}_0 + \alpha_N \boldsymbol{u}) - M(\boldsymbol{\beta}_0), \\ &\leq -D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta}_0 + \alpha_N \boldsymbol{u})) + D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta}_0)) - N \sum_{j=1}^{s} \{J_{\tau_N}(|\beta_{j0} + \alpha_N u_j|) - J_{\tau_N}(|\beta_{j0}|)\}, \end{aligned}$$

where s is the number of components of $\boldsymbol{\beta}_{10}$. Let $\nabla D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta}_0))$ be the gradient vector of $D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta}_0))$. Using the Taylor expansion of the phi divergence measure,

$$\begin{aligned} W_N(\boldsymbol{u}) \leq & -\alpha_N \nabla D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta}_0))^{\mathrm{T}} \boldsymbol{u} - \frac{1}{2} \boldsymbol{u}^{\mathrm{T}} I(\boldsymbol{\beta}_0) \boldsymbol{u} N \alpha_N^2 \{1 + O_P(1)\} \\ & - N \alpha_N \sum_{j=1}^{s} \{\nabla J_{\tau_N}(|\beta_{j0}|) \operatorname{sign}(\beta_{j0}) u_j + N \alpha_N^2 \nabla^2 J_{\tau_N}(|\beta_{j0}|) u_j^2\} \{1 + O_P(1)\}. \end{aligned} \tag{15}$$

Note that $N^{-1/2} \nabla D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta}_0)) = O_P(1)$. Thus, the first term on the right-hand side of Equation (15) is of the order $O_P(N^{1/2} \alpha_N)$. Second term dominates the first term uniformly in $\|\boldsymbol{u}\| = C$, for sufficiently large C. The third term in Equation (15) is bounded by Fan and Li [11]

$$\sqrt{s} N \alpha_N a_N \|\boldsymbol{u}\| + N \alpha_N^2 \max\{|\nabla^2 J_{\tau_N}(|\beta_{j0}|)| : \beta_{j0} \neq 0\} \|\boldsymbol{u}\|^2.$$

This is also dominated by the second term of Equation (15). Hence, by choosing a sufficiently large C, Equation (14) holds. Hence, the theorem is proved. ■

Thus, by choosing a proper τ_N, there exists a root-N consistent penalized minimum ϕ-divergence estimator. We now show that this estimator possess the sparsity property $\hat{\boldsymbol{\beta}}_2 = 0$ which is stated as follows.

THEOREM 2 *Let $\phi(t) \in \Phi$. Assume that*

$$\frac{\lim_{N \to \infty} \inf \lim_{\theta \to 0+} \inf J'_{\tau_N}(\theta)}{\tau_N} > 0. \tag{16}$$

If $\tau_N \to 0$ and $\sqrt{N} \tau_N \to \infty$ as $N \to \infty$, then with probability tending to 1, for any given $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(N^{-1/2})$ and any constant C,

$$M\left\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ 0 \end{pmatrix}\right\} = \max_{\|\boldsymbol{\beta}_2\| \leq C N^{-1/2}} M\left\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}\right\}.$$

Proof To prove Theorem 2, it is sufficient to show that for some small $\varepsilon_N = CN^{-1/2}$, with probability tending to 1 as $N \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(N^{-1/2})$ and

$$\begin{array}{ll} \dfrac{\partial M(\boldsymbol{\beta})}{\partial \beta_j} < 0 & \text{for } 0 < \beta_j < \varepsilon_N \\ > 0 & \text{for } -\varepsilon_N < \beta_j < 0 \end{array} \quad ; \quad j = s+1, \ldots, k. \tag{17}$$

Using Taylor's series expansion, we have

$$\begin{aligned} \frac{\partial M(\boldsymbol{\beta})}{\partial \beta_j} &= -\frac{\partial D_\phi(\boldsymbol{p}, \boldsymbol{p}(\boldsymbol{\beta}))}{\partial \beta_j} - NJ'_{\tau_N}(|\beta_j|)\text{sign}(\beta_j) \\ &= -\frac{\partial D_\phi(\boldsymbol{p}, \boldsymbol{p}(\boldsymbol{\beta}_0))}{\partial \beta_j} + \sum_{l=1}^{k} \frac{\partial^2 D_\phi(\boldsymbol{p}, \boldsymbol{p}(\boldsymbol{\beta}_0))}{\partial \beta_j \partial \beta_l}(\beta_l - \beta_{l0}) \\ &\quad + \sum_{l=1}^{k}\sum_{m=1}^{k} \frac{\partial^3 D_\phi(\boldsymbol{p}, \boldsymbol{p}(\boldsymbol{\beta}^*))}{\partial \beta_j \partial \beta_l \partial \beta_m}(\beta_l - \beta_{l0})(\beta_k - \beta_{k0}) - NJ'_{\tau_N}(|\beta_j|)\text{sign}(\beta_j), \end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Note that, $(1/N)(\partial D_\phi(\boldsymbol{p}, \boldsymbol{p}(\boldsymbol{\beta}_0))/\partial \beta_j) = O_P(N^{-1/2})$. By the assumption that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(N^{-1/2})$, we have

$$\frac{\partial M(\boldsymbol{\beta})}{\partial \beta_j} = N\tau_N \left\{ -\frac{1}{\tau_N} J'_{\tau_N}(|\beta_j|)\text{sign}(\beta_j) + O_P\left(\frac{N^{-1/2}}{\tau_N}\right) \right\}.$$

Whereas, $\lim_{N\to\infty} \inf \lim_{\theta \to 0+} \inf J'_{\tau_N}(\theta)/\tau_N > 0$ and $N^{-1/2}/\tau_N \to 0$, the sign of the derivative is completely determined by that of β_j. Hence, Equation (17) holds. ■

In the following theorem, we establish the oracle property of the proposed estimator. Denote $\Sigma = \text{diag}\{J''_{\tau_N}(|\boldsymbol{\beta}_{10}|), \ldots, J''_{\tau_N}(|\boldsymbol{\beta}_{s0}|)\}$ and $\boldsymbol{b} = (J'_{\tau_N}(|\boldsymbol{\beta}_{10}|)\text{sign}(\boldsymbol{\beta}_{10}), \ldots, J'_{\tau_N}(|\boldsymbol{\beta}_{s0}|)\text{sign}(\boldsymbol{\beta}_{s0}))^{\mathrm{T}}$, where, s is the number of components of $\boldsymbol{\beta}_{10}$.

THEOREM 3 (Oracle Property) *Let $\phi(t) \in \Phi$. Assume that the penalty function $J_{\tau_N}(|\beta_{j0}|)$ satisfies the condition in* Equation (16). *If $\tau_N \to 0$ and $\sqrt{N}\tau_N \to \infty$ as $N \to \infty$, then with probability tending to 1, the root-N consistent local maximizers $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$ in Theorem 1 must satisfy:*

(a) Sparsity: $\hat{\boldsymbol{\beta}}_2 = 0$
(b) Asymptotic Normality:

$$\sqrt{N}(I_1(\boldsymbol{\beta}_{10}) + \Sigma)\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (I_1(\boldsymbol{\beta}_{10}) + \Sigma)^{-1}\boldsymbol{b}\} \to N\{0, I_1(\boldsymbol{\beta}_{10})\}$$

in distribution, where $I_1(\beta_{10}) = I_1(\boldsymbol{\beta}_{10}, 0)$ is the Fisher information knowing $\boldsymbol{\beta}_2 = 0$.

Proof The proof of the part (a) follows from Theorem 2. Now we prove part (b). The minimum ϕ-divergence estimator satisfies the best asymptotic normal (BAN) decomposition and is the BAN estimator of $\boldsymbol{\beta}$ [20]. It means that the asymptotic behavior of minimum ϕ-divergence estimator is

same as that of MLE irrespective of the choice of the function ϕ. Following the same steps of the proof of Theorem 2 in Fan and Li [11], using Slutsky's theorem, it is easy to verify that,

$$\sqrt{N}(I_1(\boldsymbol{\beta}_{10}) + \Sigma)\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (I_1(\boldsymbol{\beta}_{10}) + \Sigma)^{-1}\boldsymbol{b}\} \to N\{0, I_1(\boldsymbol{\beta}_{10})\}$$

in distribution. ■

As a consequence, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_1$ is
$(1/N)(I_1(\boldsymbol{\beta}_{10}) + \Sigma)^{-1}I_1(\boldsymbol{\beta}_{10})(I_1(\boldsymbol{\beta}_{10}) + \Sigma)^{-1}$ which approximately equals $(1/N)I_1^{-1}(\boldsymbol{\beta}_{10})$ for the SCAD penalty if τ_N tends to 0.

3.2 *Algorithm*

Since, Equation (5) is continuous and twice differentiable with respect to $\boldsymbol{\beta}$, minimizing $D_\phi\left(\hat{\boldsymbol{p}}, \boldsymbol{p}\left(\boldsymbol{\beta}\right)\right)$ in respect to $\boldsymbol{\beta}$ is not a difficult task. This can be done using Newton–Raphson method and the $(t+1)^{\text{th}}$ step estimate, $\hat{\boldsymbol{\beta}}_\phi^{(t+1)}$, is obtained from $\hat{\boldsymbol{\beta}}_\phi^{(t)}$ as

$$\hat{\boldsymbol{\beta}}_\phi^{(t+1)} = \hat{\boldsymbol{\beta}}_\phi^{(t)} - G\left(\hat{\boldsymbol{\beta}}_\phi^{(t)}\right)^{-1} X^{\mathrm{T}}\mathrm{Diag}\left(\left(\frac{n_i}{N}\pi_{i1}^{(t)}\pi_{i2}^{(t)}\right)_{i=1,\ldots,I}\right) T\left(\hat{\boldsymbol{\beta}}_\phi^{(t)}\right),$$

where $G(\boldsymbol{\beta}) = X^{\mathrm{T}}\mathrm{Diag}(((n_i/N)\pi_{i1}\pi_{i2})_{i=1,\ldots,I})X$,

$$T(\boldsymbol{\beta}) = \left(\phi\left(\frac{\boldsymbol{n}_{(1)}}{\boldsymbol{m}(\boldsymbol{\beta})}\right) - \frac{\boldsymbol{n}_{(1)}}{\boldsymbol{m}(\boldsymbol{\beta})}\phi'\left(\frac{\boldsymbol{n}_{(1)}}{\boldsymbol{m}(\boldsymbol{\beta})}\right) - \phi\left(\frac{\boldsymbol{n}-\boldsymbol{n}_{(1)}}{\boldsymbol{n}-\boldsymbol{m}(\boldsymbol{\beta})}\right) + \frac{\boldsymbol{n}-\boldsymbol{n}_{(1)}}{\boldsymbol{n}-\boldsymbol{m}(\boldsymbol{\beta})}\phi'\left(\frac{\boldsymbol{n}-\boldsymbol{n}_{(1)}}{\boldsymbol{n}-\boldsymbol{m}(\boldsymbol{\beta})}\right)\right)$$

being $\boldsymbol{n} = (n_1, \ldots, n_I)^{\mathrm{T}}$, $\boldsymbol{n}_{(1)} = (n_{i1}, \ldots, n_{I1})^{\mathrm{T}}$ and $\boldsymbol{m}(\boldsymbol{\beta}) = (n_1\pi_{11}, \ldots, n_I\pi_{I1})^{\mathrm{T}}$. For details see Pardo *et al.* [20].

As the SCAD penalty is singular at the origin and does not have second-order derivative, it can be locally approximated by a quadratic function [11] as follows. Suppose that an initial value $\boldsymbol{\beta}^0$ is close to the minimizer of Equation (12). If β_j^0 is very close to 0, then set $\beta_j^0 = 0$. Otherwise, it can be locally approximated by a quadratic function as

$$J_\tau(|\beta_j|) \approx J_\tau(|\beta_j^0|) + \frac{1}{2}\left\{\frac{J_\tau'(|\beta_j^0|)}{|\beta_j^0|}\right\}((\beta_j)^2 - (\beta_j^0)^2) \quad \text{for } \beta_j \approx \beta_j^0. \tag{18}$$

Thus, Equation (12) can be locally approximated by

$$Q(\boldsymbol{\beta}) \approx D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta})) + \frac{1}{2}N\boldsymbol{\beta}^{\mathrm{T}}\Sigma_\tau(\boldsymbol{\beta}^0)\boldsymbol{\beta}, \tag{19}$$

where $\Sigma_\tau(\boldsymbol{\beta}^0) = \mathrm{Diag}((J_\tau'(|\beta_j^0|)/|\beta_j^0|)_{j=1,\ldots,k})$. To the minimization problem in Equation (12), Newton–Raphson procedure can be applied and the $(t+1)^{\text{th}}$ step estimate, $\hat{\boldsymbol{\beta}}^{(t+1)}$, is obtained from $\hat{\boldsymbol{\beta}}^{(t)}$ as

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - (\nabla^2 D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\hat{\boldsymbol{\beta}}^{(t)})) + N\Sigma_\tau(\hat{\boldsymbol{\beta}}^{(t)}))^{-1}(\nabla D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\hat{\boldsymbol{\beta}}^{(t)})) + N\mathbf{U}_\tau(\hat{\boldsymbol{\beta}}^{(t)})), \tag{20}$$

where $\nabla D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\hat{\boldsymbol{\beta}}^{(t)})) = X^{\mathrm{T}}\mathrm{Diag}(((n_i/N)\pi_{i1}^{(t)}\pi_{i2}^{(t)})_{i=1,\ldots,I})T(\hat{\boldsymbol{\beta}}^{(t)})$,
$\nabla^2 D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\hat{\boldsymbol{\beta}}^{(t)})) = G(\hat{\boldsymbol{\beta}}^{(t)})$ and $\boldsymbol{U}_\tau(\hat{\boldsymbol{\beta}}^{(t)}) = \Sigma_\tau(\hat{\boldsymbol{\beta}}^{(t)})\hat{\boldsymbol{\beta}}^{(t)}$. The iterations are terminated when the Euclidean distance between estimates of two successive iterations is smaller than some threshold.

We used 10^{-8} as the threshold value for termination of iterations. When the iterations are stopped, nonzero elements of $\hat{\boldsymbol{\beta}}$ approximately satisfy

$$\frac{\partial D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\hat{\boldsymbol{\beta}}))}{\partial \beta_j} + NJ'_\tau(|\hat{\beta}_j|)\text{sign}(\hat{\beta}_j) = 0.$$

3.3 *Selection of thresholding parameters*

The thresholding parameter τ plays a vital role in the performance of a procedure based on penalized loss function. If it is not appropriately chosen, the estimates become unstable. Use of cross validation to choose τ was the first choice to many researchers. Generalized cross validation (GCV) was used by Tibshirani [28] and Fan and Li [11] for this purpose. Fan and Li [11] also used V-fold cross validation but found results similar to that of GCV. Later, Wang *et al.* [30] showed that GCV is not consistent for selecting τ in SCAD. They proposed a consistent criterion based on information called Generalized information criterion of which Akaike information criterion and Bayes information criterion (BIC) are particular cases. Their simulation study established that BIC is a better selector. It is defined as

$$\text{BIC}(\tau) = \frac{1}{N} D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_\tau\right) + \frac{1}{N} \text{d}f_\tau \log(N), \tag{21}$$

where $\hat{\boldsymbol{\mu}}_\tau$ is penalized MLE of $\boldsymbol{\mu}$ when threshold parameter is τ and $D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_\tau\right)$ is the model deviance. $\text{d}f_\tau$ denotes number of nonzero components in $\hat{\boldsymbol{\beta}}_\tau$.

We used the BIC defined in Equation (21) but replaced $\hat{\boldsymbol{\mu}}_\tau$ by the penalized minimum ϕ-divergence estimate of $\boldsymbol{\mu}$ when threshold parameter is τ. Since, the form of this new selector is same as that of BIC, we call it as the BIC type selector. Our simulation study presented in the next section indicates that τ chosen using BIC type selector gives good results. This motivates us to use this criterion to choose tuning parameter τ in ϕ SCAD.

4. Simulation study

This section is divided into three subsections. In the first subsection, we present the results of simulation study performed to choose the value of λ. A real-life application of the proposed method for illustrative purpose is given in the second subsection. In the third subsection, we compare the performance of proposed method with SCAD using simulation.

4.1 *Selection of* λ

In the simulation study of Pardo *et al.* [20], $\lambda = \frac{2}{3}$ emerged as a good choice for λ in the minimum ϕ-divergence estimator. The performance of ϕ SCAD also depends on the choice of λ. Hence, it should be carefully chosen. To select the value of λ, we perform the simulation study similar to the one in [20].

Consider the Binomial regression model in which response Y_i has binomial distribution with parameters n_i and $\pi(\boldsymbol{x}_i)$ and $\pi(\boldsymbol{x}_i) = \exp\{\beta_0 + \sum_{j=1}^5 \beta_j x_{ij}\}/(1 + \exp\{\beta_0 + \sum_{j=1}^5 \beta_j x_{ij}\})$. The observations on the predictor variables are given in Table 1. The number of distinct $\boldsymbol{x}_i$'s in this example is $I = 20$. A correlation of 0.5 was introduced in first two predictors. The response Y_i follow binomial $(n_i, \pi(\boldsymbol{x}_i))$ and $\pi(\boldsymbol{x}_i)$ is as defined above with $\boldsymbol{\beta} = (3, 1.5, -2, 0, 0, 0)^{\text{T}}$ for Model I and $\boldsymbol{\beta} = (1.5, -1.5, 1.5, 2, -2, 1.5)^{\text{T}}$ for Model II. We simulated Models I and II, 500 times for

Table 1. The values of x_{ij} in Example 1.

x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}
0.8451	2.4896	0.4009	−0.136	−1.9752	−1.8927	−0.3368	0.979	−1.0472	−0.4425
−0.7435	0.6849	0.0697	−0.6224	0.4119	−0.699	−0.7438	0.351	−1.9229	−0.4719
0.1647	1.1647	−1.6608	−0.8612	−1.3012	1.3177	−1.3967	0.3339	−1.1499	−0.5843
−0.4278	−0.9471	0.0422	2.3386	−0.626	−0.9127	0.41	−0.5538	0.4145	−0.1557
0.3517	−0.4363	−0.5924	1.5747	0.5404	0.9559	0.2141	1.5588	0.4653	−0.1186
−0.4791	−0.2663	−0.7643	0.3917	1.1794	0.1867	−0.0012	0.2194	0.5182	0.8995
0.523	0.9216	−0.9703	−0.9815	−1.6545	1.3932	−1.1858	−0.4565	0.0075	1.8205
0.7523	−0.271	−0.2461	−0.6377	1.4705	−1.7296	0.9468	1.9457	0.283	0.4275
1.6461	0.5777	−1.9543	−0.0331	0.6814	−0.109	0.2732	0.113	−0.2961	−0.4615
−0.1057	−1.8081	1.4955	0.8164	−0.2752	−1.0374	−0.2893	−0.3106	0.8783	0.6355

Table 2. MSE of nonzero coefficients based on 500 repetitions.

λ	$-\frac{1}{2}$	0	$\frac{2}{3}$	1	2	3
Model I						
n^1	.1665	.1462	.1253	.1577	.1780	.1897
n^2	.1667	.1623	.1476	.1854	.2000	.2147
n^3	.0998	.0953	.0805	.0808	.1114	.1218
n^4	.1199	.1151	.0818	.1188	.1214	.1332
n^5	.1264	.1155	.0927	.1207	.1298	.1403
Average	.1358	.12688	.1055	.1326	.1481	.1599
Model II						
n^1	.1539	.1397	.1295	.1360	.1427	.1588
n^2	.1773	.1461	.1351	.1312	.1499	.1648
n^3	.1160	.0827	.0714	.0729	.0932	.1450
n^4	.1233	.1003	.0994	.1067	.1125	.1219
n^5	.1467	.1223	.1170	.1194	.1376	.1515
Average	.1434	.1182	.1104	.1132	.1271	.1484

different values of n_i's as given below.

$n^1 : 5, 5, 5, 5, 5, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10.$

$n^2 : 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10.$

$n^3 : 5, 5, 5, 5, 5, 15, 15, 15, 15, 15, 30, 30, 30, 30, 30, 40, 40, 40, 40, 40.$

$n^4 : 5, 5, 5, 5, 5, 10, 10, 10, 10, 10, 15, 15, 15, 15, 15, 20, 20, 20, 20, 20.$

$n^5 : 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10.$

For brevity, we shall denote SCAD penalized MLE as SCAD here onwards. To compute ϕ SCAD estimates, we considered the power divergence measure given in Equation (6). We used the value of $a = 3.7$ as suggested by Fan and Li [11] in SCAD penalty. The MSE of the $(k + 1)$ dimensional estimator will be a matrix. We call the trace of the MSE matrix as the total MSE (TMSE). Here, we divide the TMSE by the number of nonzero parameters and denote it by MSE. The MSE of ϕ SCAD for different values of λ is reported in Table 2.

The simulation results in Table 2 clearly indicate that the choice of $\lambda = \frac{2}{3}$ yields smaller MSE. This is not a surprise as this choice of λ is supported by simulation studies in [6,20,21,25].

Table 3. Frequency of zero estimates of zero coefficients for Model I based on 500 repetitions.

λ		$-\frac{1}{2}$	0	$\frac{2}{3}$	1	2	3
n^1	$\hat{\beta}_3$	485	487	490	491	493	495
	$\hat{\beta}_4$	493	492	495	496	497	499
	$\hat{\beta}_5$	490	494	498	498	498	499
	$\hat{\beta}_3$	480	481	484	487	490	492
n^2	$\hat{\beta}_4$	491	493	497	496	498	498
	$\hat{\beta}_5$	488	493	497	498	498	499
	$\hat{\beta}_3$	484	487	493	494	496	497
n^3	$\hat{\beta}_4$	493	495	499	499	499	499
	$\hat{\beta}_5$	490	496	499	499	499	499
	$\hat{\beta}_3$	490	492	497	496	499	499
n^4	$\hat{\beta}_4$	493	495	498	497	500	500
	$\hat{\beta}_5$	491	494	497	496	500	500
	$\hat{\beta}_3$	493	495	499	499	500	500
n^5	$\hat{\beta}_4$	494	495	499	499	500	500
	$\hat{\beta}_5$	490	495	497	499	500	500

Table 3 gives frequency of zero estimates of zero coefficients in the above example based on 500 repetitions. The frequency of zero estimates of zero coefficients for $\lambda = \frac{2}{3}$ is close to 500 based on 500 repetitions. This also supports our claim. For other choices of λ, MSE is large, however, frequency of zero estimates of zero coefficients is close to 500.

4.2 *Real data application*

We consider a real-life data [3, p. 171], used by Pardo and Pardo [21] to illustrate the variable selection method based on minimum ϕ-divergence estimator. The data consists of observations on six objective indicators $\{X_1, \ldots, X_6\}$ of the actual indoor climate (predictors) in 10 classrooms of a Danish Institute. The response variable is the number of yes-answers to the question whether they felt that the indoor climate at the moment was pleasant or not so pleasant and the number of students in each of the 10 classrooms is also reported.

We used MLE, minimum ϕ-divergence estimator ($\lambda = \frac{2}{3}$), ϕ SCAD ($\lambda = \frac{2}{3}$) and SCAD to estimate the regression coefficients. These are reported in Table 4. The method given in Pardo

Table 4. Parameter estimates for the real data.

Predictors	MLE	Mϕ DE[a]	Post variable selection Mϕ DE	Post variable selection MLE	ϕSCAD $\tau = 0.223$	SCAD $\tau = 0.19$
Intercept	4.6563	7.1557	−10.0163	−11.1599	−10.8423	5.1934
X_1	1.3204	1.2987	0.9516	1.0420	0.9516	1.3160
X_2	−1.1412	−1.1523	−0.6442	−0.7026	−0.6441	−1.3950
X_3	20.2955	19.2891	0.0000	0.0000	0.0000	19.1100
X_4	1.4486	1.4237	0.8521	0.9496	0.8521	1.4350
X_5	25.3047	25.2041	16.0221	17.5313	16.0227	24.9710
X_6	−0.0705	−0.077	0.0000	0.0000	0.0000	−0.0715

[a]Minimum ϕ-divergence estimator.

Table 5. Simulation results for performance comparison.

n	Method	Average number of correct zeroes	TMSE	MRME
n^1	ϕ SCAD	0.9819	0.8682 (1.2278)[a]	0.9072
	SCAD	0.9875	1.5884 (4.0638)	0.9164
	Oracle	1.0000	0.8292 (1.5944)	0.8324
n^2	ϕ SCAD	0.9900	1.4854 (1.2989)	0.8764
	SCAD	0.9996	1.8369 (4.4443)	0.9152
	Oracle	1.0000	0.9619 (1.7057)	0.8276
n^3	ϕ SCAD	0.9858	0.3564 (0.2901)	0.8673
	SCAD	0.9453	0.6402 (1.4743)	0.8910
	Oracle	1.0000	0.2757 (0.7355)	0.8399
n^4	ϕ SCAD	0.9745	0.7249 (0.9075)	0.8609
	SCAD	0.9815	1.2071 (5.5267)	0.8943
	Oracle	1.0000	0.6303 (0.8754)	0.8536
n^5	ϕ SCAD	0.9889	0.3381 (0.3658)	0.8503
	SCAD	0.9942	1.1983 (3.1253)	0.8891
	Oracle	1.0000	0.2999 (0.4208)	0.8460

[a]Figures in parenthesis indicate corresponding standard deviation.

and Pardo [21] selects set of predictors $\{X_1, X_2, X_4, X_5\}$ which coincides with that proposed by Andersen [3]. We present the estimates corresponding to this set of predictors using minimum ϕ-divergence estimator ($\lambda = \frac{2}{3}$) and MLE in the columns 4 and 5, respectively of Table 4.

For these data, the SCAD fails to identify the correct set of predictors. ϕ SCAD selected the same set of predictors proposed by Pardo and Pardo [21] and Andersen [3]. Also, estimates of nonzero coefficients are very close to the one obtained using minimum ϕ-divergence estimator ($\lambda = \frac{2}{3}$) assuming that the predictors X_3 and X_6 are absent in the model.

4.3 *Performance comparison*

In this subsection, we compare the performance of ϕ SCAD with $\lambda = \frac{2}{3}$ and SCAD for different combinations of n_i's. Sample size was fixed to 20. The predictors X_1, X_2 and X_3 were generated from standard normal distribution such that the correlation between X_1 and X_2 is 0.5. The response Y_i follow binomial $(n_i, \pi(\boldsymbol{x}_i))$ and $\pi(\boldsymbol{x}_i)$ is as defined in Equation (1) with $\boldsymbol{\beta} = (3, 1.5, 0, 2)^{\mathrm{T}}$. We simulated five different models 1000 times characterized by the values of n_i's mentioned in Section 4.3.

We report the average number of correct zeroes, TMSE and median of relative model error (MRME) in Table 5. Box plots of TMSE are presented in Figure 1.

MRME is computed relative to the model error of full model based on the unpenalized MLE. Oracle estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}}$ is obtained by maximizing the likelihood assuming that $\beta_2 = 0$.

The average TMSE of ϕ SCAD and SCAD, averaged over all the five models are 0.7546 and 1.2942, respectively. TMSE and MRME of ϕ SCAD are close to that of oracle estimator. Moreover, both the methods give more or less same average number of zeroes.

5. Discussion

We proposed computation of ϕ SCAD using the Newton–Raphson method based on local quadratic approximation of penalty function [11]. Even though this method is fast and efficient, it is very

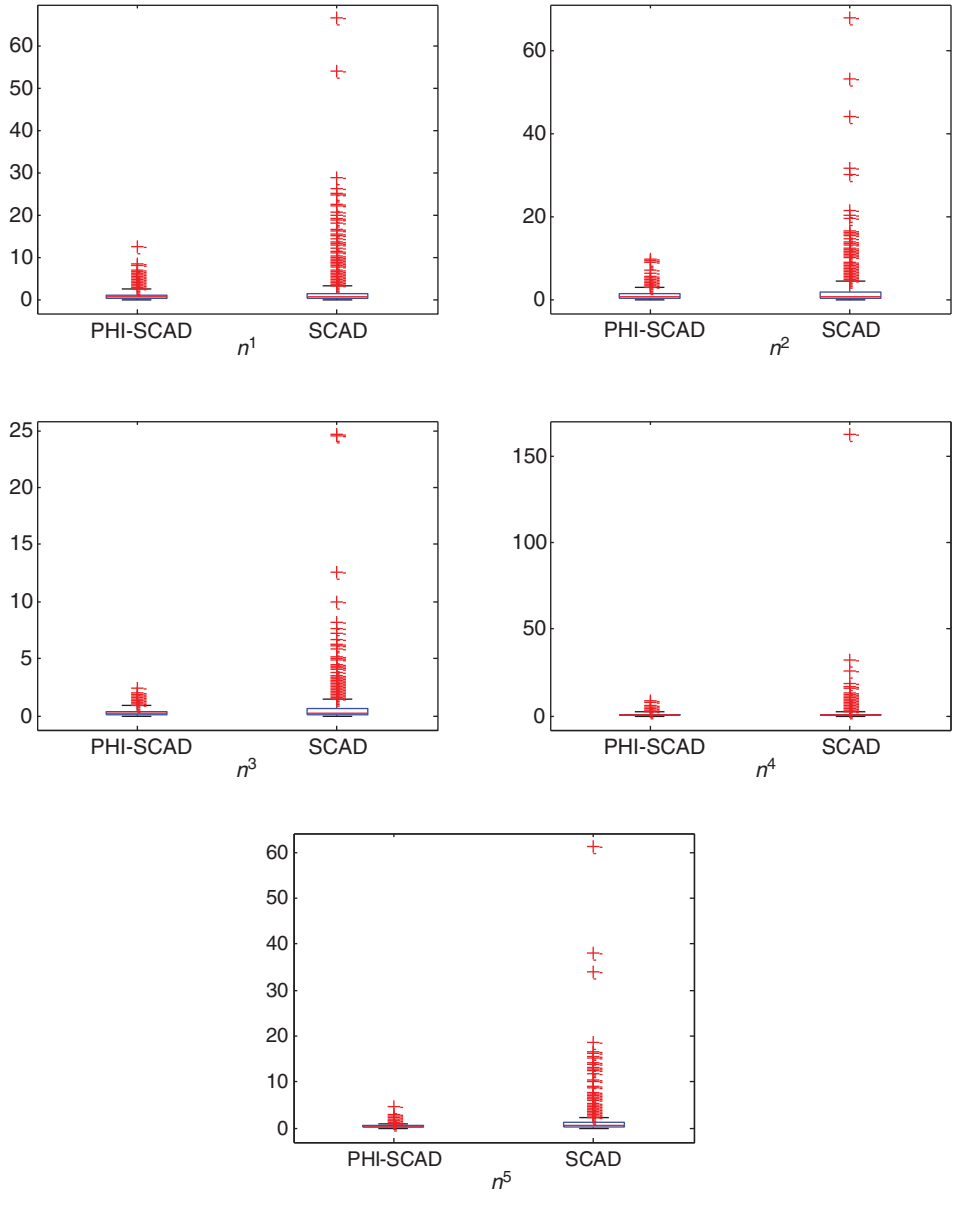


Figure 1. Box plot of TMSE (Example 3).

sensitive to starting points. Particularly, if $D_\phi(\hat{\boldsymbol{p}}, \boldsymbol{p}(\boldsymbol{\beta}))$ is very flat near its minimizer, Newton–Raphson algorithm may not converge if starting values are not chosen properly. Such a case is very rare in practice. We suggest the use of Expectation–Maximization (EM) algorithm to avoid this potential issue. The efficacy and usefulness of EM algorithm for penalized likelihood estimation is proved by Green [15] and De Pierro [9].

We proposed the penalized ϕ-divergence estimation using SCAD for simultaneous estimation and variable selection in logistic regression. It is interesting to note that SCAD and ϕ SCAD identify number of zeroes efficiently. Our simulation study indicates that MRME of ϕ SCAD is less than that of SCAD and is close to that of oracle estimator. It is evident that ϕ SCAD performs

as well as if $\boldsymbol{\beta}_{20} = 0$ were known. In the language of Donoho and Johnstone [10], the resulting estimator performs as well as the oracle estimator, which knows in advance that $\boldsymbol{\beta}_{20} = 0$. ϕ SCAD estimates the nonzero parameters more efficiently than SCAD penalized MLE in MSE sense. This makes ϕ SCAD an attractive alternative to penalized MLE when sample size is small in logistic regression. Moreover, we theoretically showed that ϕ SCAD is equivalent to penalized MLE asymptotically.

The minimum ϕ-divergence estimation has also emerged as a good estimation procedure in more complex models like log linear models with multinomial sampling scheme [5,6,23] and polytomous logistic regression [16]. As per the suggestion of one of the referees, the ϕ SCAD can also be extended to such models. The detail study of the properties and performance of ϕ SCAD for more complex models like polytomous logistic regression or multinomial probit models can constitute the material for a new research paper.

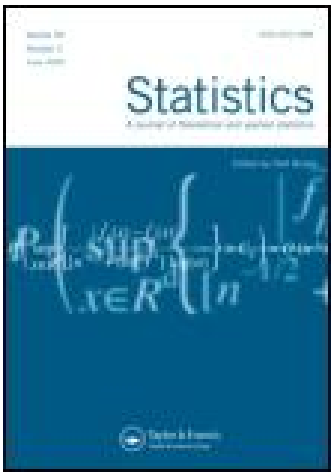
Acknowledgements

The authors thank the Editor and anonymous referees whose suggestions led to the significant improvement in this paper.

References

[1] A. Agresti, *Categorical Data Analysis*, John Wiley and Sons, New York, 1990.
[2] S.M. Ali and S.D. Silvey, *A general class of coefficients of divergence of one distribution from another*, J. R. Stat. Soc. Ser. B 26 (1966), pp. 131–142.
[3] E.B. Andersen, *Introduction to the Statistical Analysis of Categorical Data I*, Springer-Verlag, Heidelberg, 1997.
[4] J. Anderson and V. Blair, *Penalized maximum likelihood estimation in logistic regression and discrimination*, Biometrika 69 (1982), pp. 123–136.
[5] N. Cressie and L. Pardo, *Minimum ϕ -divergence estimator and hierarchical testing in loglinear models*, Stat. Sin. 10 (2000), pp. 867–884.
[6] N. Cressie, L. Pardo, and M.C. Pardo, *Size and power considerations for testing Loglinear models using ϕ -divergence test statistics*, Stat. Sin. 13 (2003), pp. 555–570.
[7] N. Cressie and T.R.C. Read, *Multinomial goodness of fit tests*, J. R. Stat. Soc. Ser. B 46 (1984), pp. 440–464.
[8] I. Csiszár, *Eine Informationtheorestiche Ungleichung und ihre Anwendung anf den Beweis der Ergodizität Markoffshen Ketten*, Publ. Math. Inst. Hung. Acad. Sci. Ser. A 8 (1963), pp. 84–108.
[9] A.R. De Pierro, *A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography*, IEEE Trans. Med. Imaging 14(1) (1995), pp. 132–137.
[10] D.L. Donoho and I.M. Johnstone, *Ideal spatial adaptation by wavelet shrinkage*, Biometrika 81 (1994), pp. 425–455.
[11] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Am. Stat. Assoc. 96 (2001), 1348–1360.
[12] J. Fan and H. Peng, *Nonconcave penalized likelihood with a diverging number of parameters*, Ann. Stat. 32 (2004), 928–961.
[13] I.E. Frank and J.H. Friedman, *A statistical view of some chemometrics regression tools*, Technometrics 35 (1993), 109–148.
[14] A. Genkin, D.D. Lewis, and D. Madigan, *Large-scale Bayesian logistic regression for text categorization*, Technometrics 49 (2007), pp. 291–304.
[15] P.J. Green, *On use of the EM for penalized likelihood estimation*, J. R. Stat. Soc. Ser. B (Methodol.) 52(3) (1990), pp. 443–452.
[16] A.K. Gupta, D. Kasturiratna, T. Nguyen, and L. Pardo, *A new family of BAN estimators for polytomous logistic regression models based on φ-divergence measures*, Stat. Methods Appl. 15 (2006), 159–176.
[17] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, John Wiley, New York, 1989.
[18] Y. Kim, H. Choi, and H. Oh, *Smoothly clipped absolute deviation on high dimensions*, J. Am. Stat. Assoc. 103 (2008), pp. 1665–1673.
[19] S. Kullback, *Kullback information*, in *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz and N.L. Johnson, eds., John Wiley, New York, 1985, pp. 421–425.
[20] J.A. Pardo, L. Pardo, and M.C. Pardo, *Minimum ϕ -divergence estimator in logistic regression models*, Stat. Pap. 47 (2005), pp. 91–108.
[21] J.A. Pardo and M.C. Pardo, *Minimum ϕ -divergence estimator and ϕ -divergence statistics in generalized linear models with binary data*, Methodol. Comput. Appl. Probab. 10 (2008), pp. 357–379.

[22] L. Pardo, *Statistical Inference Based on Divergence Measures*, Chapman and Hall/CRC, Taylor and Francis Group, Boca Raton, FL, 2006.
[23] L. Pardo and M.C. Pardo, *Minimum power-divergence in three-way contingency tables*, J. Stat. Comput. Simul. 73(11) (2003), pp. 819–831.
[24] W.C. Parr, *Minimum distance estimation: a bibliography*, Commun. Stat. Theory Methods 10 (1981), pp. 1205–1224.
[25] T.R.C. Read and N. Cressie, *Goodness of Fit Statistics for Discrete Multivariate Data*, Springer, New York, 1988.
[26] V. Roth, *The generalized LASSO*, IEEE Trans. Neural Networks 15 (2004), pp. 16–28.
[27] S. Shevade and S. Keerthi, *A simple and efficient algorithm for gene selection using sparse logistic regression*, Bioinformatics 19 (2003), pp. 2246–2253.
[28] R.J. Tibshirani, *Regression shrinkage and selection via LASSO*, J. R. Stat. Soc. B 58 (1996), pp. 267–288.
[29] I. Vajda, *Theory of Statistical Inference and Information*, Kluwer, Boston, MA, 1989.
[30] H. Wang, R. Li, and C.L. Tsai, *On the consistency of SCAD tuning parameter selector*, Biometrika 94 (2007), pp. 553–568.
[31] H. Xie and J. Huang, *SCAD-penalized regression in high-dimensional partially linear models*, Ann. Stat. 37 (2009), pp. 673–696.
[32] H. Zang, J. Ahn, X. Lin, and C. Park, *Gene selection using support vector machines with non-convex penalty*, Bioinformatics 22 (2006), 88–95.



Statistics: A Journal of Theoretical and Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gsta20

A deviance-based criterion for model selection in GLM

D. M. Sakate[a] & D. N. Kashid[a]
[a] Department of Statistics, Shivaji University, Kolhapur (MS), India
Published online: 30 Jul 2012.

To cite this article: D. M. Sakate & D. N. Kashid (2014) A deviance-based criterion for model selection in GLM, Statistics: A Journal of Theoretical and Applied Statistics, 48:1, 34-48, DOI: 10.1080/02331888.2012.708035

To link to this article: http://dx.doi.org/10.1080/02331888.2012.708035

Statistics, 2014
Vol. 48, No. 1, 34–48, http://dx.doi.org/10.1080/02331888.2012.708035

Taylor & Francis
Taylor & Francis Group

A deviance-based criterion for model selection in GLM

D.M. Sakate* and D.N. Kashid

Department of Statistics, Shivaji University, Kolhapur (MS), India

(*Received 31 December 2011; final version received 27 June 2012*)

Model selection is the most persuasive problem in generalized linear models. A model selection criterion based on deviance called the deviance-based criterion (DBC) is proposed. The DBC is obtained by penalizing the difference between the deviance of the fitted model and the full model. Under certain weak conditions, DBC is shown to be a consistent model selection criterion in the sense that with probability approaching to one, the selected model asymptotically equals the optimal model relating response and predictors. Further, the use of DBC in link function selection is also discussed. We compare the proposed model selection criterion with existing methods. The small sample efficiency of proposed model selection criterion is evaluated by the simulation study.

Keywords: GLM; deviance; model selection; consistency; observed L_2 efficiency

AMS Subject Classification: 62J12

1. Introduction

Regression is the most widely used technique to model the relationship between a response variable and predictors. Some of these predictors may be redundant in nature and are required to be eliminated based on the observed data. Model selection plays an important role to identify the necessary predictors which are related to the response variable. In classical regression, Mallows' [1] C_p is one of the most widely used model selection criteria. AIC [2], AICc [3], BIC [4], cAIC [5] and others [6,7] are some of the model selection criteria to be noted.

Nelder and Wedderburn [8] introduced the generalized linear model (GLM) as a unification of linear and nonlinear regression models that incorporated a rich family of normal and non-normal distributions for the response variable. The GLM is a powerful tool to model the relationship between predictors and the function of the mean for continuous and discrete response variables. In practice, the GLM is used to model the various kinds of data like clinical trials data, ecological data, meteorological data, etc. Lawless and Singhal [9,10], Nordberg [11,12] and Hosmer *et al*. [13] provided methods for model selection in the GLM. If the likelihood is specified, AIC is still applicable. Qian *et al*. [14] proposed a model selection criterion in the class of the GLM based on the predictive minimum description length principle and the theory of quasi likelihood known as a predictive least quasi-deviance (PLQD) criterion. PLQD requires fitting of sequence of models and is computationally complex. Recently, Hu and Shao [15] proposed a model selection criterion based on adjusted R^2 which is consistent under weak conditions.

*Corresponding author. Email: dms.stats@gmail.com

In this article, we propose a deviance-based criterion (DBC) which penalizes the difference in deviance of the fitted and the full model by complexity via the number of predictors in the model. Deviance is familiar to investigators using the GLM as a modelling tool. Moreover, the DBC is computationally simpler as compared to PLQD. Under certain weak conditions, minimizing the DBC results in a consistent model selection in the sense that with the probability approaching to one model selected is asymptotically equal to the optimal model which contains no redundant variables.

The remaining article is organized as follows. Section 2 discusses the set up of the GLM and describes models. We propose a DBC and establish its consistency for model selection in Section 3. Use of the DBC for the link function selection is discussed in Section 4. The performance of the DBC is evaluated by simulation study in Section 5. Also, it is compared with existing model selection criteria. Section 6 presents some concluding remarks.

2. Generalized linear models

The GLM is defined via a link function

$$g(\mu_i) = X_i'\beta, \quad i = 1, 2, \ldots, n, \tag{1}$$

where, $\beta \in \mathcal{R}^k$ is a vector of regression parameters, $X_i = (1, X_1, \ldots, X_{k-1}) \in \mathcal{R}^k$ and $k < n$. The maximum-likelihood estimator (MLE) of β after using iteratively reweighted least squares at convergence is

$$\hat{\beta}^f = (X'V^{-1}X)^{-1}X'V^{-1}z,$$

where, the superscript f denotes the estimate corresponding to the fitted model, X is an $n \times k$ real matrix and V is an $n \times n$ diagonal matrix whose diagonal elements are $v_i = (\mathrm{d}\theta_i/\mathrm{d}\mu_i)a(\phi)$ and $z_i = g(\hat{\mu}_i) + (y_i - \hat{\mu}_i)(\mathrm{d}g(\mu_i)/\mathrm{d}\mu_i)$. Following McCullagh and Nelder [16], the discrepancy of the fitted GLM is twice the difference between the maximum log likelihood achievable in a saturated model with n parameters $L(y, \phi; y)$ and that achieved by the model under investigation $L(\hat{\mu}, \phi; y)$. A saturated model has n parameters, one per observation, and μ_i's derived from it match the data exactly. The saturated model consigns all the variation in y_i's to the systematic component leaving none for the random component. Denote $\hat{\theta}^f = \theta(\hat{\mu})$ and $\tilde{\theta} = \theta(y)$, the estimates of the location (canonical) parameters under the fitted model and the saturated model, respectively, and we assume $a_i(\phi) = \phi/w_i$. The discrepancy between the model under investigation and saturated model is given by

$$\frac{\sum 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i^f) - b(\tilde{\theta}_i) + b(\hat{\theta}_i^f)\}}{\phi} = \frac{D(y, \hat{\beta}^f)}{\phi},$$

where, $D(y, \hat{\beta}^f)$ is commonly known as the deviance of the fitted model.

In the GLM, the model selection involves identifying relevant predictors and link function. For a GLM, we denote a model by M_α, where $\alpha = \alpha_0 \cup \alpha_l, \alpha_0 = \{0\}$ denotes intercept and α_l denotes a non-empty subset of $\{1, 2, \ldots, k-1\}$. The model M_α, is defined as

$$g(\mu_{i,\alpha}) = X_{i,\alpha}'\beta_\alpha, \tag{2}$$

where, $X_{i,\alpha}$ denotes the sub-vector of X_i containing components indexed by α, β_α is a p_α-vector and p_α denotes cardinality of α.

Suppose, α_N denote all necessary predictors. Following Shao [17], each candidate model can be associated with one of the following two categories.

(1) Class of wrong models $\mathcal{M}_w = \{M_\alpha :$ at least one necessary predictor is missing$\}$, i.e. $\mathcal{M}_w = \{M_\alpha : \alpha_N \nsubseteq \alpha\}$.
(2) Class of correct models $\mathcal{M}_c = \{M_\alpha :$ all necessary predictors are present$\}$, i.e. $\mathcal{M}_c = \{M_\alpha : \alpha_N \subseteq \alpha\}$.

The models in $\mathcal{M}_c$ are correct models and those in $\mathcal{M}_w$ are the wrong models. There are more than one correct models unless $\alpha_N = \alpha_0 \cup \{1, 2, \ldots, k-1\}$. The optimal model is M_{α_N}.

3. Model selection using DBC

The deviance is a function of the data only and is used to define a statistic for model selection. Let $D(y, \hat{\beta})$ denote the deviance of the full model. If the difference in deviance of model M_α and full model $D(y, \hat{\beta}_\alpha) - D(y, \hat{\beta})$ is small, then the model M_α can be regarded as good as the full model for prediction. This cannot serve the purpose of the model selection criterion because for the model M_{α_*} such that $\alpha_* \supset \alpha$, the difference is smaller than that for the model M_α and is zero when α corresponds to the full model. Hence, it becomes difficult to identify the optimal model. A good model selection criterion should take into account goodness of fit as well as the complexity of the model [18]. A natural measure of the complexity of the model is the number of parameters p_α involved in it. Therefore, we define a model selection criterion in the GLM based on the penalized difference between deviance of model M_α and the full model. The DBC can be expressed as

$$\mathrm{DBC}(M_\alpha) = \frac{D(y, \hat{\beta}_\alpha) - D(y, \hat{\beta})}{\phi} - (k - p_\alpha) + C(n, p_\alpha), \tag{3}$$

where 'ϕ' is the dispersion parameter and is usually known. If it is unknown, it is replaced by its MLE.

Under normality of the response and $C(n, p_\alpha) = p_\alpha$, the criterion in Equation (3) is equivalent to Mallows' C_p (see Lemma 3.1 for the details). Minimum Mallows' C_p is not a consistent model selection criterion [19]. Its inconsistency is due to the constant penalty p_α which does not increase when the sample size is increased. It is necessary to consider a complexity measure $C(n, p_\alpha)$ which will make the criterion consistent.

Under criterion (3), those candidate models having better goodness of fit and smaller complexity will be preferred than the others; and the best model will be the one achieving the smallest DBC value.

In order to establish the consistency of DBC, we require the following condition. This condition ensures that the wrong model is asymptotically worse than any correct model.

Condition 3.1 For $M_\alpha \in \mathcal{M}_w$ and $M_{\alpha_*} \in \mathcal{M}_C$,

$$\lim_{n \longrightarrow \infty} \inf (I + (p_\alpha - p_{\alpha_*}) + C(n, p_\alpha) - C(n, p_{\alpha_*})) > 0$$

where, $I = \sum_{i=1}^{n} 2w_i\{y_i(\hat{\theta}_{i,\alpha_*} - \hat{\theta}_{i,\alpha}) + b(\hat{\theta}_{i,\alpha_*})\}/\phi$.

If M_α is a wrong model, then the deviance of M_α is larger than that of a correct model M_{α_*}. Hence, quantity I is positive and large. Thus, the assumption in Condition 3.1 is reasonable.

The following theorem indicates that, if we choose a model by minimizing DBC over all possible models, then asymptotically, the model selected by using DBC falls in the class of correct models.

THEOREM 3.1 *Under Condition 3.1, for any correct model $M_{\alpha_*} \in \mathcal{M}_C$ and any wrong model M_α we have,*

$$\lim_{n\longrightarrow\infty} \inf \Pr(\mathrm{DBC}(M_\alpha) > \mathrm{DBC}(M_{\alpha_*})) = 1.$$

Proof The deviance of a model M_α can be expressed as

$$\begin{aligned} D(y, \hat{\beta}_\alpha) &= \sum_{i=1}^{n} 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_{i,\alpha_*}) - b(\tilde{\theta}_i) + b(\hat{\theta}_{i,\alpha_*})\} \\ &\quad + \sum_{i=1}^{n} 2w_i\{y_i(\hat{\theta}_{i,\alpha_*} - \hat{\theta}_{i,\alpha}) + b(\hat{\theta}_{i,\alpha}) - b(\hat{\theta}_{i,\alpha_*})\} \\ &= D(y, \hat{\beta}_{\alpha_*}) + \phi I. \end{aligned}$$

Therefore,

$$\Pr(\mathrm{DBC}(M_\alpha) - \mathrm{DBC}(M_{\alpha_*}) > 0) = \Pr(I + (p_\alpha - p_{\alpha_*}) + C(n, p_\alpha) - C(n, p_{\alpha_*}) > 0).$$

Hence,

$$\begin{aligned} &\lim_{n\longrightarrow\infty} \inf \Pr(\mathrm{DBC}(M_\alpha) > \mathrm{DBC}(M_{\alpha_*})) \\ &= \lim_{n\longrightarrow\infty} \inf \Pr(I + (p_\alpha - p_{\alpha_*}) + C(n, p_\alpha) - C(n, p_{\alpha_*}) > 0) \\ &\geq \Pr(\lim_{n\longrightarrow\infty} \inf (I + (p_\alpha - p_{\alpha_*}) + C(n, p_\alpha) - C(n, p_{\alpha_*}) > 0)). \end{aligned}$$

Using Condition 3.1, we have

$$\lim_{n\longrightarrow\infty} \inf \Pr(\mathrm{DBC}(M_\alpha) > \mathrm{DBC}(M_{\alpha_*})) = 1. \qquad \blacksquare$$

It follows from the above theorem that with the probability approaching to one, value of the DBC for a wrong model is larger than that for any correct model. Further, we state some lemmas.

LEMMA 3.1 *If M_{α_*} is a correct model and n is large, $(D(y, \hat{\beta}_{\alpha_*}) - D(y, \hat{\beta}))/\phi$ has an approximately chi-square distribution with $k - p_{\alpha_*}$ degrees of freedom (d.f.) [20].*

Condition 3.2 $C(n, p_\alpha) = o(n)$ and $C(n, p_\alpha) \to \infty$ as $n \to \infty$.

Let M_n denote the model selected by using DBC when the sample size is n. Moreover, if $C(n, p_\alpha)$ satisfies Condition 3.2, the following theorem indicates that our model selection criterion is consistent.

THEOREM 3.2 *Under Condition 3.2, with probability approaching to one, as n tends to infinity, DBC selects the optimal model in the class of all correct models, i.e.*

$$\lim_{n\longrightarrow\infty} \Pr(M_n = M_{\alpha_{\mathrm{N}}}) = 1.$$

Proof In the light of Theorem 3.1, for large n model selected by the DBC falls in the class of correct models. Therefore, we shall confine ourselves to the class of correct models only. For a

correct model $M_{\alpha_*} \in \mathcal{M}_C$,

$$\begin{aligned}
&\Pr(\mathrm{DBC}(M_{\alpha_*}) > \mathrm{DBC}(M_{\alpha_N})) \\
&= \Pr\left(\frac{D(y,\hat{\beta}_{\alpha_*}) - D(y,\hat{\beta}_{\alpha_N})}{\phi} > (p_{\alpha_N} - p_{\alpha_*}) + C(n,p_{\alpha_N}) - C(n,p_{\alpha_*})\right) \\
&= \Pr\left(\frac{D(y,\hat{\beta}_{\alpha_N}) - D(y,\hat{\beta}_{\alpha_*})}{\phi} < (p_{\alpha_*} - p_{\alpha_N}) + C(n,p_{\alpha_*}) - C(n,p_{\alpha_N})\right)
\end{aligned}$$

Since, $C(n,p_\alpha) \to \infty$ as $n \to \infty$ and $p_{\alpha_*} > p_{\alpha_N}$ for any correct model $M_{\alpha_*} \in \mathcal{M}_C$, and by Lemma 3.1 we have,

$$\lim_{n\to\infty} \Pr(\mathrm{DBC}(M_{\alpha_*}) > \mathrm{DBC}(M_{\alpha_N})) = \Pr(\chi^2_{(p_{\alpha_*} - p_{\alpha_N})} < \infty) = 1.$$

This indicates that, with probability approaching to one, asymptotically value of DBC for the optimal model is the smallest in the class of all correct models. Moreover, the DBC selects that model for which its value is minimum among all possible models.

Therefore,

$$\lim_{n\to\infty} \Pr(M_n = M_{\alpha_N}) = 1. \qquad \blacksquare$$

This proves that, DBC is a consistent model selection criterion.

LEMMA 3.2 *If M_α is a correct model then $E(\mathrm{DBC}) \cong C(n,p_\alpha)$.*

Proof According to Lemma 3.1, distribution of the first term in Equation (3) is approximately chi-square with $k - p_\alpha$ d.f., we have

$$\begin{aligned}
E(\mathrm{DBC}) &\cong (k - p_\alpha) - (k - p_\alpha) + C(n,p_\alpha) \\
&= C(n,p_\alpha). \qquad \blacksquare
\end{aligned}$$

LEMMA 3.3 *If $C(n,p_\alpha) = p_\alpha$ and distribution of response is normal then DBC and Mallows' C_p are equivalent.*

Proof Let $Y_1, Y_2, \ldots, Y_n$ be independent $N(\mu_i, \sigma^2)$.

Then, $D(y,\hat{\beta}) = \mathrm{RSS}_k$, $D(y,\hat{\beta}_\alpha) = \mathrm{RSS}_{p_\alpha}$ and $\phi = \sigma^2$.

Therefore,

$$\mathrm{DBC}(M_\alpha) = \frac{\mathrm{RSS}_{p_\alpha} - \mathrm{RSS}_k}{\sigma^2} - (k - 2p_\alpha).$$

Since, σ^2 is unkown, replace it by its OLS estimator $\hat{\sigma}^2 = \mathrm{RSS}_k/n - k$.

Thus, we have

$$\mathrm{DBC}(M_\alpha) = \frac{\mathrm{RSS}_{p_\alpha}}{\hat{\sigma}^2} - (n - 2p_\alpha) = C_p. \qquad \blacksquare$$

Remark 3.1 The term $k - p_\alpha$ is included in Equation (3) because for the optimal and full model, the expectation $E((D(y,\hat{\beta}_{\alpha_N}) - D(y,\hat{\beta}))/\phi - (k - p_\alpha)) \cong 0$ with equality for the full model and $E((D(y,\hat{\beta}_{\alpha_N}) - D(y,\hat{\beta}))/\phi - (k - p_\alpha)) > 0$, otherwise. Also, complexity of the full model is larger than that of the optimal model. This indicates that the value of the DBC is small for the optimal model. This helps to identify the optimal model easily.

It is evident that the DBC belongs to the class of likelihood-based model selection criteria of which AIC and BIC are widely used. They are defined as follows:

$$\mathrm{AIC}(M_\alpha) = -2L(\hat{\mu}_\alpha, \hat{\phi}; y) + 2p_\alpha, \tag{4}$$

$$\mathrm{BIC}(M_\alpha) = -2L(\hat{\mu}_\alpha, \hat{\phi}; y) + p_\alpha \log(n). \tag{5}$$

AIC is an efficient model selection criterion [21]. It has been shown that AIC is not consistent [15]. BIC is a consistent model selection criterion [19]. A recently proposed consistent model selection criterion based on modified adjusted R^2 is $\bar{R}^2$ [15]. It is defined as

$$\bar{R}^2(M_\alpha) = 1 - \frac{n-1}{n - \lambda_n p_\alpha} \frac{\sum_{i=1}^n (y_i - \hat{\mu}_{i,\alpha})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \tag{6}$$

where, $\hat{\mu}_{i,\alpha}$ is obtained by maximizing the quasi-likelihood of the model M_α and λ_n is a penalty term satisfying Condition 3.2. The simulation study in [15] reveals that $\bar{R}^2$ performs better than PLQD, AIC, AICc and is compatible with BIC. They used $\lambda_n = \log n$ and $\lambda_n = \sqrt{n}$ for the simulation purpose.

Model selection in the GLM involves identification of relevant predictors and a selection of the link function. We discussed the identification of relevant predictors, using the DBC when the link function is known in this section. In the next section, the use of the DBC for model selection when the true link function is unknown but belongs to a parametric family is discussed.

4. Model selection with parametric link function

When the true link function is known, the DBC defined in Equation (3) can be used as it is for model selection. If the link function is unknown and is to be selected from a finite set $\mathcal{G}$ of continuous monotone link functions, the DBC in Equation (3) cannot be used as it is because of the presence of deviance of the full model in it. When there are more than one candidate link functions, if the deviance of the full model is based on a respective link to compute the DBC for all possible models corresponding to each link function, then it may happen that the minimum DBC corresponds to a wrong model (in the sense of incorrect link as well as predictors). This can be overcome by initial screening of all the candidate link functions using deviance. Let $g \in \mathcal{G}$, be one of the finitely many link functions in $\mathcal{G}$ and M_α^g be the GLM defined in Equation (2) when the link function is g. Denote the mean of the response of model M_α^g by μ_α^g and the regression parameter by β_α^g. Then the DBC for model selection when the link function is unknown is defined as

$$\mathrm{DBC}(M_\alpha^g) = \frac{D(y, \hat{\beta}_\alpha^g) - D(y, \hat{\beta}^{g^*})}{\phi} - (k - p_\alpha) + C(n, p_\alpha), \tag{7}$$

where $D(y, \hat{\beta}^g)$ is the deviance of the full model corresponding to the link function $g^* \in \mathcal{G}$ such that

$$\min_{g \in \mathcal{G}} D(y, \hat{\beta}^g) = D(y, \hat{\beta}^{g^*})$$

The problem of parametric link selection is addressed by Pregibon [22] and later by Czado [23], Czado and Munk [24] and Hu and Shao [15]. Pregibon [22] proposed a test for checking whether a modification to the hypothetical link function is necessary or not. This test is based on the reduction in deviance of the model when the modified link is used and if it is significant, then it is necessary to modify the hypothetical link. Moreover, he emphasized that this is the first logical step towards optimal link identification. The notion behind using the term $D(y, \hat{\beta}^{g^*})$ in Equation (7) follows from this.

Table 1. Penalty functions.

Sr. No.	Penalty function $C(n, p_\alpha)$
1	$P_1 = p_\alpha$
2	$P_2 = 2p_\alpha$
3	$P_3 = 2p_\alpha + \dfrac{2(p_\alpha+1)(p_\alpha+2)}{n-p_\alpha-2}$
4	$P_4 = p_\alpha \log(n)$
5	$P_5 = p_\alpha(\log(n)+1)$

Further, the first term in Equation (7) is always positive. Theorems 3.1 and 3.2 proved for DBC in Equation (3) can also be proved for DBC defined in Equation (7) on the similar lines. Hence, the consistency property of DBC for optimal model selection can be established when the link function is to be selected from $\mathcal{G}$.

The performance of DBC for the link function selection in the GLM is evaluated and compared with existing methods using simulation and the results are presented in Part C of the next section.

5. Simulation results

In this section, we present the results of the simulation study. The study is divided into three parts. In part (A), two examples are discussed. Example 5.1 uses the data given in [14] to compare the performance of DBC with some existing model selection criteria. In Example 5.2, simulated data are used to examine and compare the performance of DBC. Part (B) presents the small sample observed L_2 efficiency of DBC, $\bar{R}^2$, AIC and BIC. In Part (C), we report the findings of the performance study of various model selection criteria for the link function selection. At the end, we discuss the choice of the penalty function in DBC. In the entire simulation study, we use five different penalty functions which appear in the literature and are presented in Table 1.

(A) *Performance and comparison study*

Example 5.1 We consider a Poisson regression model where response Y follows the Poisson distribution with mean μ and $\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$. The same design matrix ($n = 36$)

Table 2. Probabilities of selecting each model ($n = 36$).

		DBC					$\bar{R}^2$,	$\bar{R}^2$,	
β	Model	P_1	P_2	P_3	P_4	P_5	$\lambda_n = \log(n)$	$\lambda_n = \sqrt{n}$	BIC
{2,1,0,0}[a]	{0,1}$\{M_{\alpha_N}\}$	0.703	0.827	0.872	0.922	0.972	0.764	0.881	0.868
	{0,1,2}[b]	0.141	0.088	0.068	0.042	0.018	0.104	0.055	0.072
	{0,1,3}	0.130	0.078	0.058	0.036	0.010	0.103	0.56	0.059
	{0,1,2,3}	0.026	0.007	0.002	0	0	0.029	0.008	0.001
{2,1,0.5,0.35}	{0,1}	0	0.001	0.003	0.003	0.009	0	0	0
	{0,1,2}	0.050	0.081	0.112	0.164	0.276	0.055	0.092	0.093
	{0,1,3}	0	0.004	0.003	0.008	0.028	0.004	0.008	0.007
	{0,1,2,3}$\{M_{\alpha_N}\}$	0.950	0.914	0.882	0.825	0.687	0.941	0.900	0.900
{2,3,0,0.1}	{0,1}	0.381	0.510	0.580	0.675	0.808	0.312	0.399	0.596
	{0,1,2}	0.059	0.039	0.032	0.022	0.012	0.092	0.085	0.040
	{0,1,3} $\{M_{\alpha_N}\}$	0.480	0.418	0.370	0.292	0.176	0.440	0.420	0.342
	{0,1,2,3}	0.080	0.033	0.018	0.011	0.004	0.156	0.096	0.022

[a] $\{i, j, \ldots\}$ denote $\{\beta_0, \beta_1, \beta_2, \beta_3\}$.
[b] $\{i, j, \ldots\}$ denote suffixes of $\{X_0, X_1, X_2, X_3\}$.

and parameter structure used for the purpose of simulation in [14] is used. Model selection using DBC is carried out. Table 2 shows the probabilities of selecting models for three different β's based on 1000 runs (values for BIC and $\bar{R}^2$ are obtained from [15]).

In the first case, where predictors X_2 and X_3 are redundant, performance of DBC with penalty P_4 and P_5 is better than BIC and $\bar{R}^2$. In the second case, where no predictor is redundant, performance of DBC with penalty P_1, is better than BIC and compatible with $\bar{R}^2$. This is obvious because a smaller penalty will perform better when the full model itself is the optimal model. In the third case, the parameter structure is very interesting. Predictor X_2 is redundant while X_3 is near to being redundant. DBC with P_1 and P_2, BIC and $\bar{R}^2$ select evenly the optimal model and the near to optimal model (containing intercept and X_1 only). In this case, DBC (with P_2 to P_5) tend to select the near to optimal model with higher probability than the optimal model, and the probability being highest for P_5. Hu and Shao [15] used the same data to compare the performance of $\bar{R}^2$ with existing model selection criteria and found that it performs better than PLQD, AIC and AICc and has a performance similar to that of BIC. The above simulation study indicates that DBC performs better than $\bar{R}^2$.

Example 5.2 In this example, we assessed the performance of DBC and compared it with some existing methods by counting the frequency of selecting the optimal model based on the simulated data. Five response distributions namely normal, Poisson, Bernoulli, gamma and negative binomial are considered in this study. Four sample sizes 100, 200, 400 and 500 are used. For each response distribution, two different parameter structures are considered. These structures consist of discrete and continuous predictors. Table 3 gives different parameter structures and link functions corresponding to response distributions.

The design matrix X is of the order $n \times 6$ with the first column as ones. The columns 2–6 consist of random numbers from Normal(0, 1), Uniform(0, 1), Poisson(1), Binomial(1, 0.4) and Normal(0, 1) distributions, respectively. The GLM involves dispersion parameter ϕ which is to be estimated from the data. In the GLM, ϕ plays a very important role. As the expression of DBC involves ϕ, its performance may differ for different values of ϕ. The theoretical value of ϕ for Poisson and Bernoulli is one. Value of ϕ other than one cannot be used for the simulation in case of these distributions. Therefore, we considered $\phi = 1$ for the simulation from Poisson and Bernoulli response models. For normal, gamma and negative binomial distribution, we considered four different values of ϕ for the simulation. We used MLE of ϕ in DBC, AIC and BIC. Tables 4–6 present the frequency of selecting optimal model by DBC, $\bar{R}^2$, AIC and BIC over 1000 realizations for model I and II when the distribution of response is normal, gamma and negative binomial, respectively. The frequency count of the optimal model selection by various criteria when distribution of response is Poisson and Bernoulli are reported in Table 7.

Table 3. Parameter structure and link functions.

Distribution of response	Model no.	Parameter structure	Link function
Normal	I	$g(\mu) = 1 + 5X_1 + 5X_2$	Identity
	II	$g(\mu) = 1 + 5X_1 + 5X_2 + 5X_3$	
Poisson	I	$g(\mu) = 1 + X_2 + X_3$	Log
	II	$g(\mu) = 1 + X_1 + 2X_4$	
Bernoulli	I	$g(\mu) = 2 + X_1 + X_2 + X_3$	Logit
	II	$g(\mu) = 1 + X_1 - 2X_3$	
Gamma	I	$g(\mu) = 1 + 15X_1 + 15X_2$	Log
	II	$g(\mu) = 1 + 5X_1 - 5X_2 + 5X_3$	
Negative binomial	I	$g(\mu) = 1.5 + 1.5X_1 - 1.5X_2$	Log
	II	$g(\mu) = 0.5 + 1.5X_1 - 2X_3$	

Table 4. Frequency count of optimal model selection when response distribution is normal.

		DBC					$\bar{R}^2$			
ϕ	n	P_1	P_2	P_3	P_4	P_5	$\lambda_n = \log n$	$\lambda_n = \sqrt{n}$	AIC	BIC
Model I										
0.5	100	794	859	868	964	976	848	976	580	887
	200	796	863	869	975	986	891	998	556	915
	400	813	885	888	983	989	948	998	599	954
	500	785	870	873	989	991	958	1000	568	969
1	100	810	875	886	970	984	861	979	594	908
	200	795	879	883	983	990	915	995	581	934
	400	809	894	903	987	992	958	1000	609	962
	500	802	880	882	984	992	957	1000	590	965
2	100	812	881	888	967	980	853	973	579	897
	200	813	880	882	984	989	921	995	576	940
	400	779	862	865	991	994	945	1000	590	953
	500	797	871	874	988	994	953	1000	601	960
4	100	791	864	874	958	977	850	974	585	886
	200	799	878	883	980	984	900	998	596	923
	400	794	876	883	982	993	947	1000	590	957
	500	802	879	880	988	994	950	1000	600	959
Model II										
0.5	100	831	886	901	970	983	885	974	709	922
	200	830	894	903	980	987	934	997	684	949
	400	845	903	906	991	997	972	1000	727	976
	500	821	895	898	991	994	971	1000	704	973
1	100	840	910	919	981	987	899	976	712	937
	200	815	904	906	982	988	934	997	690	951
	400	813	888	890	984	992	967	1000	679	974
	500	838	908	910	992	997	973	1000	720	977
2	100	855	914	927	973	982	897	976	701	937
	200	810	898	905	982	988	931	995	682	952
	400	853	920	925	994	998	968	1000	726	973
	500	830	903	907	995	999	976	1000	696	983
4	100	829	903	911	975	986	890	983	701	927
	200	825	906	914	982	991	942	997	696	953
	400	831	911	918	992	997	973	1000	714	978
	500	826	904	907	985	991	955	999	696	963

Note: Number of simulations for each combination is 1000.

It seems that, irrespective of values of ϕ and n considered, DBC with P_4 and P_5 and $\bar{R}^2$ perform equally and are better than AIC and BIC for the normal response distribution. Moreover, frequency of the optimal model selection of each criterion for various values of ϕ is same for the respective sample sizes in this case.

When the distribution of the response is gamma, for $\phi = 0.5$, 1 and 2, DBC with P_4 and P_5 perform better than $\bar{R}^2$, AIC and BIC, irrespective of the sample size. For $\phi = 4$, DBC performs better than $\bar{R}^2$ and AIC for large sample sizes but its performance is lower than BIC. It means that the value of ϕ plays a significant role in the performance of DBC for a gamma response.

Table 6 gives some interesting findings regarding the performance of various model selection criteria in the GLM form of the negative binomial regression. DBC with P_4 and P_5 perform better than AIC, BIC and $\bar{R}^2$ irrespective of sample size and the values of ϕ. However, for $\phi = 4$, performance of all model selection criteria is lower but that of DBC with P_4 and P_5 is moderate.

Table 7 presents the results of the performance study in the case of the Poisson and the Bernoulli response distributions for $\phi = 1$. For Poisson response distribution, all sample sizes considered, the frequency of selecting the optimal model by DBC with P_4 and P_5 is larger than that for AIC, BIC and $\bar{R}^2$. When the response distribution is Bernoulli, the DBC with P_4 and P_5 perform better

Table 5. Frequency count of optimal model selection when response distribution is gamma.

		DBC					$\bar{R}^2$			
ϕ	n	P_1	P_2	P_3	P_4	P_5	$\lambda_n = \log n$	$\lambda_n = \sqrt{n}$	AIC	BIC
Model I										
0.5	100	625	796	818	960	981	206	275	596	890
	200	651	824	838	975	988	235	325	627	943
	400	619	796	799	986	992	213	359	592	961
	500	640	799	803	986	994	203	360	620	966
1	100	653	812	839	964	980	224	298	610	907
	200	656	806	817	980	992	215	326	627	939
	400	684	840	844	991	996	222	377	649	967
	500	644	822	828	984	989	193	351	604	957
2	100	692	817	838	931	945	184	253	638	896
	200	723	856	864	983	990	205	302	658	932
	400	707	856	859	989	990	189	349	651	963
	500	714	864	873	995	999	183	333	653	966
4	100	235	278	282	307	312	217	259	555	695
	200	338	407	408	454	455	205	265	581	788
	400	527	629	631	704	706	231	332	637	894
	500	576	663	665	740	744	204	301	649	900
Model II										
0.5	100	739	860	879	974	987	265	324	718	934
	200	743	872	885	984	992	294	383	722	965
	400	751	864	870	985	991	286	418	739	971
	500	760	871	876	992	996	284	404	745	982
1	100	757	877	903	971	984	272	335	725	938
	200	766	890	898	990	997	275	362	738	966
	400	762	883	888	990	995	246	343	731	970
	500	774	881	884	994	999	246	376	750	974
2	100	766	886	903	957	969	243	279	727	927
	200	792	902	910	985	993	279	361	753	963
	400	777	899	904	992	998	265	352	731	974
	500	789	901	908	994	998	261	358	742	982
4	100	238	275	277	298	300	198	212	447	535
	200	434	479	483	514	514	253	303	623	738
	400	630	703	705	748	748	215	302	723	877
	500	675	758	762	813	814	232	332	730	909

Note: Number of simulations for each combination is 1000.

than AIC, BIC and $\bar{R}^2$ for large n. In general, the DBC with P_4 and P_5 performs better than $\bar{R}^2$, AIC and BIC.

(B) *Observed L_2 efficiency*

McQuarrie *et al*. [6] adopted the approach of Shibata [21] to define the observed L_2 efficiency and used it to compare AIC, AICc, and AICu in classical regression. Following the same approach, we define the observed L_2 efficiency of a model selection criterion in the GLM. Let $\mathcal{M} = \mathcal{M}_C \cup \mathcal{M}_w$ be the class of all possible models and

$$L_2(M_m) = \min_{M \in \mathcal{M}} L_2(M),$$

where, $L_2(M) = \|\mu_{\alpha_N} - \hat{\mu}_\alpha\|^2/n$. In addition, let M_S denote the model selected by the specific model selection criterion. The observed L_2 efficiency of a model selection criterion in the GLM is then defined as $L_2(M_m)/L_2(M_S)$. The efficiency of a model selection criterion will be high if it selects the model which best approximates the optimal model. Hence, a good model selection criterion will select a model which yields high efficiency. As the correct model is asymptotically

Table 6. Frequency count of optimal model selection when response distribution is negative binomial.

		DBC					$\bar{R}^2$			
ϕ	n	P_1	P_2	P_3	P_4	P_5	$\lambda_n = \log n$	$\lambda_n = \sqrt{n}$	AIC	BIC
Model I										
0.5	100	681	785	803	868	861	112	187	320	418
	200	688	801	810	958	966	99	228	202	359
	400	688	809	816	966	974	84	242	86	216
	500	736	824	829	973	978	92	270	56	132
1	100	600	667	676	661	624	96	161	239	325
	200	654	760	771	903	891	108	188	147	259
	400	705	806	808	948	956	90	241	52	131
	500	668	778	779	931	940	86	228	44	100
2	100	454	498	498	425	371	70	99	174	186
	200	568	666	674	688	652	82	157	121	195
	400	659	752	759	872	871	96	182	42	80
	500	613	731	734	911	921	87	211	34	67
4	100	279	303	306	242	224	83	103	115	84
	200	406	457	460	446	416	80	135	82	69
	400	469	566	571	658	644	86	151	26	31
	500	504	619	620	758	759	83	187	13	21
Model II										
0.5	100	489	624	656	849	889	107	194	658	869
	200	536	675	689	888	922	111	225	536	796
	400	569	682	689	911	940	73	220	388	689
	500	593	709	713	925	948	90	248	368	642
1	100	491	627	647	833	879	93	164	528	793
	200	539	670	681	865	892	100	208	444	720
	400	573	694	698	904	924	82	226	318	607
	500	553	688	692	897	920	78	260	278	569
2	100	476	611	631	809	853	91	168	500	750
	200	512	618	632	832	860	93	204	371	676
	400	484	595	603	834	856	84	259	243	532
	500	537	632	635	867	898	83	250	202	474
4	100	392	504	536	720	761	96	125	450	643
	200	433	544	554	769	803	103	200	342	635
	400	476	568	571	814	847	99	248	223	492
	500	462	563	566	815	847	91	262	204	430

Note: Number of simulations for each combination is 1000.

closest to the optimal model, according to McQuarrie *et al*. [6] observed L_2 efficiency can also be used as a measure of consistency. Because the small sample efficiency of the proposed model selection criterion is vital to be noted, we present the average L_2 efficiency and its standard deviation (SD) for the sample size 50 and $\phi = 1$ in Table 8.

In case of normal, binomial and gamma response distributions, the average L_2 efficiency of DBC, $\bar{R}^2$, AIC and BIC are the same. For a negative binomial response distribution, average L_2 efficiency of DBC is larger than that of $\bar{R}^2$. Moreover, DBC with P_4 and P_5 is compatible with AIC and BIC in the sense of the average L_2 efficiency. In case of the Poisson response distribution, the average L_2 efficiency of DBC is larger than that of $\bar{R}^2$. DBC with P_4 and P_5 have larger average L_2 efficiency than that of AIC and BIC.

(C) *Link function selection*

We considered the same problem in Part A but added the selection of the link function from the set $\mathcal{G} = \{\log(\mu), \sqrt{\mu}\}$. This set-up was used by Hu and Shao [15] to demonstrate the use of $\bar{R}^2$

Table 7. Frequency count of optimal model selection when response distribution is Poisson and Bernoulli.

Distribution of response	Sample size^n	DBC					$\bar{R}^2$		AIC	BIC
		P_1	P_2	P_3	P_4	P_5	$\lambda_n = \log n$	$\lambda_n = \sqrt{n}$		
Poisson						Model I				
	100	613	788	816	947	981	356	567	623	911
	200	634	808	822	967	995	370	622	625	942
	400	620	775	785	977	989	313	671	607	960
	500	596	784	788	970	988	333	681	592	951
						Model II				
	100	652	806	821	948	983	369	557	624	908
	200	608	773	783	970	989	390	669	580	934
	400	624	789	795	976	994	388	775	603	958
	500	623	764	771	982	992	385	782	592	959
Bernoulli						Model I				
	100	494	511	523	400	346	483	331	545	423
	200	581	677	689	712	665	683	506	642	731
	400	623	766	771	925	922	867	768	675	929
	500	634	788	800	969	969	910	813	705	971
						Model II				
	100	504	660	686	818	818	675	777	566	847
	200	529	695	705	933	961	791	963	588	939
	400	538	714	721	969	981	867	999	590	966
	500	518	702	706	951	972	868	996	574	953

Note: Number of simulations for each combination is 1000.

for the link the function selection in the GLM. We performed the model selection when the link function was unknown but a member of $\mathcal{G}$. The results are given in Table 9.

As indicated in Hu and Shao [15], $\bar{R}^2$, AIC and BIC were not able to distinguish between the two link functions when $\beta = \{2, 1, 0, 0\}$ and $n = 36$. However, to some extent, the DBC with $P4$ and P_5 were able to do so. When the sample size is increased to 288 by generating eight independent Poisson responses for each covariate value, all the criteria were able to identify the true link function with a larger probability. We used one more parameter structure $\beta = \{2, 3, 0, 0\}$ and for the sample size 288, all the criteria were able to identify the true link function with probability one. Moreover, the DBC with P_4 and P_5 selected the optimal model with probability larger than that of $\bar{R}^2$, AIC and BIC. Performance of all the criteria for $\beta = \{2, 1, 0, 1.2\}$ and sample size 36 is identical in the sense of optimal model selection to $\beta = \{2, 3, 0, 0\}$ and sample size 288. It is interesting to note that DBC, $\bar{R}^2$, AIC and BIC have more or less the same performance for the link function selection for different parameter structures and sample sizes considered.

5.1. *Choice of penalty function $C(n, p_\alpha)$*

DBC involves a penalty term $C(n, p_\alpha)$ which plays a vital role in its performance. The choice of $C(n, p_\alpha)$ is crucial and should be done carefully. The penalty functions given in Table 3 appear in the existing model selection criteria. P_1, P_2 and P_3 are the part of those model selection criteria which are not consistent. Consistency is a desirable property for any model selection criterion. Therefore, $C(n, p_\alpha)$ should be chosen in such a way that DBC becomes consistent. It can be done by opting for a $C(n, p_\alpha)$ which satisfies Condition 3.2. The penalty functions P_4 and P_5 satisfy this condition and are a good choice for $C(n, p_\alpha)$ as revealed from the results of our simulation study. Of course, P_4 and P_5 are not the only options for $C(n, p_\alpha)$ and performance of DBC can be enhanced when the optimal model is strongly identifiable by using $C(n, p_\alpha)$ satisfying Condition 3.2 which

Table 8. Average L_2 efficiency and its SD for DBC and $\bar{R}^2(n = 50)$.

Model selection criteria	Penalty functions	Distribution of response				
		Normal	Poisson	Binomial	Gamma	Negative binomial
Model I						
DBC	P_1	0.972849	0.774015	0.352249	0.193805	0.741968
		(0.019354)[a]	(0.290175)	(0.01253)	(0.168298)	(0.285569)
	P_2	0.961475	0.851873	0.344588	0.194187	0.764588
		(0.026567)	(0.250818)	(0.140035)	(0.168397)	(0.281524)
	P_3	0.957265	0.877451	0.378643	0.194085	0.778043
		(0.02984)	(0.228507)	(0.105468)	(0.168223)	(0.275755)
	P_4	0.945188	0.91378	0.375384	0.194657	0.759611
		(0.038855)	(0.19038)	(0.112473)	(0.168456)	(0.287012)
	P_5	0.939696	0.949129	0.378865	0.194603	0.736897
		(0.043574)	(0.129454)	(0.100986)	(0.169181)	(0.297624)
$\bar{R}^2$	$\lambda_n = \log n$	0.963576	0.711327	0.361127	0.233095	0.554319
		(0.026336)	(0.294101)	(0.097766)	(0.203662)	(0.290255)
	$\lambda_n = \sqrt{n}$	0.947381	0.75532	0.371327	0.237309	0.560648
		(0.038602)	(0.28619)	(0.103942)	(0.208665)	(0.296184)
AIC	–	0.974995	0.7718	0.357816	0.194086	0.712637
		(0.018244)	(0.293903)	(0.095089)	(0.168411)	(0.306414)
BIC	–	0.955156	0.890916	0.371167	0.194581	0.646936
		(0.032568)	(0.217742)	(0.102679)	(0.168639)	(0.326928)
Model II						
DBC	P_1	0.972803	0.754145	0.344879	0.966541	0.699728
		(0.019495)	(0.324749)	(0.013547)	(0.046558)	(0.307471)
	P_2	0.961191	0.846499	0.344734	0.966732	0.753929
		(0.026775)	(0.28119)	(0.131435)	(0.045103)	(0.300773)
	P_3	0.956416	0.877996	0.379378	0.966731	0.772348
		(0.029703)	(0.253003)	(0.105532)	(0.045104)	(0.293737)
	P_4	0.945815	0.927001	0.376473	0.966747	0.821579
		(0.036612)	(0.198384)	(0.114783)	(0.045167)	(0.269755)
	P_5	0.940505	0.958527	0.378934	0.96676	0.844496
		(0.039997)	(0.133717)	(0.101473)	(0.045199)	(0.253927)
$\bar{R}^2$	$\lambda_n = \log n$	0.962823	0.627062	0.360543	0.967987	0.569977
		(0.025697)	(0.325587)	(0.10423)	(0.047582)	(0.304165)
	$\lambda_n = \sqrt{n}$	0.948814	0.666999	0.368366	0.968089	0.585653
		(0.03528)	(0.329841)	(0.109567)	(0.047605)	(0.309547)
AIC	–	0.975469	0.747043	0.357199	0.966538	0.850013
		(0.017958)	(0.327994)	(0.100016)	(0.046554)	(0.258175)
BIC	–	0.954654	0.89822	0.36731	0.966723	0.888216
		(0.031259)	(0.235647)	(0.107404)	(0.045204)	(0.215051)

[a]Figure in parenthesis indicates corresponding SD.

penalizes the difference in deviance of model M_α and the full model more than these can. See McQuarrie *et al*. [6] for details on identifiability of a model.

6. Concluding remarks

We proposed a new model selection criterion based on deviance in the GLM which takes into account goodness of fit as well as complexity of the model. The complexity of a model is quantified by the penalty term $C(n, p_\alpha)$. From the practical implementation point of view, $C(n, p_\alpha)$ has to be appropriately chosen. We studied the penalty functions given in Table 1 as typical choices of penalty terms in the simulation study. It is quite evident that a larger penalty performs better when the optimal model involves large number if redundant predictors. The simulation study reveals that the DBC is an attractive alternative to the existing likelihood-based model selection criteria.

Table 9. Probabilities of selecting each model.

			DBC					$\bar{R}^2$,	$\bar{R}^2$,		
β and n	Link	Model	P_1	P_2	P_3	P_4	P_5	$\lambda_n = \log(n)$	$\lambda_n = \sqrt{n}$	AIC	BIC
$\{2, 1, 0, 0\}$	$\log(\mu)$	$\{0, 1\}\{M_{\alpha_N}\}$	0.451	0.547	0.57	0.605	0.625	0.489	0.56	0.467	0.584
$n = 36$		$\{0, 1, 2\}$	0.087	0.041	0.036	0.02	0.013	0.066	0.027	0.081	0.032
		$\{0, 1, 3\}$	0.081	0.049	0.037	0.021	0.013	0.075	0.048	0.082	0.029
		$\{0, 1, 2, 3\}$	0.015	0.004	0.001	0	0	0.014	0.005	0.009	0.001
	$\sqrt{\mu}$	$\{0, 1\}$	0.254	0.285	0.302	0.317	0.326	0.287	0.322	0.252	0.306
		$\{0, 1, 2\}$	0.047	0.031	0.023	0.014	0.01	0.03	0.017	0.045	0.021
		$\{0, 1, 3\}$	0.047	0.037	0.026	0.021	0.012	0.032	0.017	0.05	0.024
		$\{0, 1, 2, 3\}$	0.018	0.006	0.005	0.002	0.001	0.007	0.004	0.014	0.003
$\{2, 1, 0, 0\}$	$\log(\mu)$	$\{0, 1\}\{M_{\alpha_N}\}$	0.591	0.711	0.715	0.83	0.838	0.797	0.842	0.592	0.814
$n = 288$		$\{0, 1, 2\}$	0.122	0.06	0.058	0.006	0.003	0.017	0	0.119	0.015
		$\{0, 1, 3\}$	0.109	0.068	0.068	0.013	0.008	0.028	0.001	0.113	0.02
		$\{0, 1, 2, 3\}$	0.022	0.006	0.005	0	0	0	0	0.021	0.001
	$\sqrt{\mu}$	$\{0, 1\}$	0.108	0.123	0.123	0.144	0.148	0.144	0.157	0.106	0.14
		$\{0, 1, 2\}$	0.024	0.019	0.018	0.005	0.002	0.008	0	0.025	0.006
		$\{0, 1, 3\}$	0.018	0.012	0.012	0.002	0.001	0.006	0	0.018	0.004
		$\{0, 1, 2, 3\}$	0.006	0.001	0.001	0	0	0	0	0.006	0
$\{2, 3, 0, 0\}$	$\log(\mu)$	$\{0, 1\}\{M_{\alpha_N}\}$	0.722	0.846	0.848	0.981	0.988	0.811	0.979	0.718	0.967
$n = 288$		$\{0, 1, 2\}$	0.12	0.066	0.066	0.01	0.008	0.096	0.011	0.125	0.014
		$\{0, 1, 3\}$	0.134	0.081	0.079	0.009	0.004	0.087	0.01	0.131	0.019
		$\{0, 1, 2, 3\}$	0.024	0.007	0.007	0	0	0.006	0	0.026	0
	$\sqrt{\mu}$	$\{0, 1\}$	0	0	0	0	0	0	0	0	0
		$\{0, 1, 2\}$	0	0	0	0	0	0	0	0	0
		$\{0, 1, 3\}$	0	0	0	0	0	0	0	0	0
		$\{0, 1, 2, 3\}$	0	0	0	0	0	0	0	0	0
$\{2, 1, 0, 1.2\}$	$\log(\mu)$	$\{0, 1, 2\}$	0	0	0	0	0	0	0	0	0
$n = 36$		$\{0, 1, 3\}\{M_{\alpha_N}\}$	0.728	0.793	0.816	0.833	0.843	0.724	0.779	0.734	0.814
		$\{0, 1, 2, 3\}$	0.141	0.075	0.048	0.03	0.019	0.145	0.089	0.135	0.051
	$\sqrt{\mu}$	$\{0, 1, 2\}$	0	0	0	0	0	0	0	0	0
		$\{0, 1, 3\}$	0.108	0.12	0.131	0.136	0.137	0.118	0.124	0.109	0.13
		$\{0, 1, 2, 3\}$	0.023	0.012	0.005	0.001	0.001	0.013	0.008	0.022	0.005

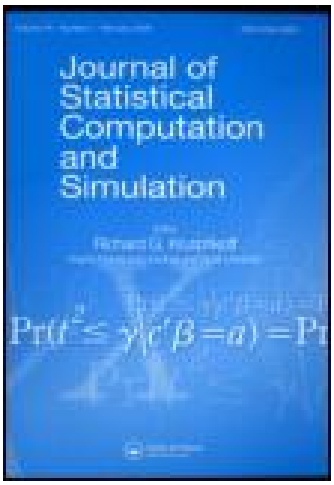
Acknowledgements

We thank the Editor and the anonymous referee for their valuable suggestions which led to the improvement of this paper.

References

[1] C.L. Mallows, *Some comments on C_p*, Technometrics 15 (1973), pp. 661–675.
[2] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automat. Control 19 (1974), pp. 716–723.
[3] C.M. Hurvich and C.L. Tsai, *Regression and time series model selection in small samples*, Biometrika 76 (1989), pp. 297–307.
[4] H. Akaike, *A Bayesian analysis of the minimum AIC procedure*, Ann. Inst. Statist. Math. A 30 (1978), pp. 9–14.
[5] D.R. Anderson and K.P. Burnham, *AIC model selection in overdispersed capture-recapture data*, Ecology 75 (1994), pp. 1780–1793.
[6] A. McQuarrie, R. Shumway, and C.L. Tsai, *The model selection criterion AICu*, Statist. Probab. Lett. 34 (1997), pp. 285–292.
[7] Z. Bai, C.R. Rao, and Y. Wu, *Model selection with data-oriented penalty*, J. Statist. Plann. Inference 77 (1999), pp. 103–117.
[8] J. Nelder and R. Wedderburn, *Generalized linear models*, J. R. Stat. Soc. A 135 (1972), pp. 370–384.
[9] J.F. Lawless and K. Singhal, *ISMOD: An all-subsets regression program for generalized linear models I statistical and computational background*, Comput. Methods Programs Biomed. 24 (1987), pp. 117–124.
[10] J.F. Lawless and K. Singhal, *ISMOD: An all-subsets regression program for generalized linear models I statistical and computational background*, Comput. Methods Programs Biomed. 24 (1987), pp. 125–134.
[11] L. Nordberg, *Stepwise selection of explanatory variables in the binary logit model*, Scand. J. Statist. 8 (1981), pp. 17–26.

[12] L. Nordberg, *On variable selection in generalized and related linear models*, Comm. Statist. Theory Methods 11 (1982), pp. 2427–2449.
[13] D.W. Hosmer, B. Jovanovic, and S. Lemeshow, *Best subsets logistic regression*, Biometrics 45 (1989), pp. 1265–1270.
[14] G. Qian, G. Gabor, and R.P. Gupta, *Generalized linear model selection by the predictive least quasi-deviance criterion*, Biometrika 83 (1996), pp. 41–54.
[15] B. Hu and J. Shao, *Generalized linear model selection using* R^2, J. Statist. Plann. Inference 138 (2008), pp. 3705–3712.
[16] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, 2nd ed., Chapman and Hall, London, 1989.
[17] J. Shao, *Linear model selection by cross validation*, J. Amer. Statist. Assoc. 422 (1993), pp. 484–494.
[18] H. Bozdogan and D.M.A. Haughton, *Informational complexity criteria for regression models*, Comput. Statist. Data Anal. 28 (1998), pp. 51–76.
[19] R. Shibata, *Consistency of model selection and parameter estimation*, J. Appl. Probab. 23 (1986), pp. 127–141.
[20] D.C. Montgomery, E.A. Peck, and G.G. Vining, *Introduction to Linear Regression Analysis*, Wiley, New York, 2006.
[21] R. Shibata, *An optimal selection of regression variables*, Biometrika 68 (1981), pp. 45–54.
[22] D. Pregibon, *Goodness of link test for generalized models*, Appl. Statist. 29 (1980), pp. 15–24.
[23] C. Czado, *On selecting parametric link transformation families in generalized linear models*, J. Statist. Plann. Inference 61 (1997), pp. 125–139.
[24] C. Czado and A. Munk, *Noncanonical links in generalized linear models-when is the effort justified*? J. Statist. Plann. Inference 87 (2000), pp. 317–345.



Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

A modified one-sample test for goodness-of-fit

B.R. Dhumal[a] & D.T. Shirke[b]

[a] Krantisinh Nana Patil College, Walwe, Sangli, Maharashtra, India

[b] Department of Statistics, Shivaji University, Kolhapur, Maharashtra, India

Published online: 12 Aug 2013.

To cite this article: B.R. Dhumal & D.T. Shirke (2015) A modified one-sample test for goodness-of-fit, Journal of Statistical Computation and Simulation, 85:2, 422-429, DOI: 10.1080/00949655.2013.825720

To link to this article: http://dx.doi.org/10.1080/00949655.2013.825720

Journal of Statistical Computation and Simulation, 2015
Vol. 85, No. 2, 422–429, http://dx.doi.org/10.1080/00949655.2013.825720

Taylor & Francis
Taylor & Francis Group

A modified one-sample test for goodness-of-fit

B.R. Dhumal[a]* and D.T. Shirke[b]

[a]*Krantisinh Nana Patil College, Walwe, Sangli, Maharashtra, India;* [b]*Department of Statistics, Shivaji University, Kolhapur, Maharashtra, India*

(*Received 19 October 2012; final version received 12 July 2013*)

This paper introduces a modified one-sample test of goodness-of-fit based on the cumulative distribution function. Damico [A new one-sample test for goodness-of-fit. Commun Stat – Theory Methods. 2004;33:181–193] proposed a test for testing goodness-of-fit of univariate distribution that uses the concept of partitioning the probability range into n intervals of equal probability mass $1/n$ and verifies that the hypothesized distribution evaluated at the observed data would place one case into each interval. The present paper extends this notion by allowing for m intervals of probability mass r/n, where $r \geq 1$ and $n = m \times r$. A simulation study for small and moderate sample sizes demonstrates that the proposed test for two observations per interval under various alternatives is more powerful than the test proposed by Damico (2004).

Keywords: distribution-free; goodness-of-fit; greatest integer function; non-parametric test; one-sample test

1. Introduction

Goodness-of-fit techniques are methods of examining how well a sample of data agrees with a specified distribution as its population. In the formal framework of hypothesis testing, the null hypothesis H_0 is that a given random variable X follows a stated probability law $F(x)$; the random variable may come from a process which is under investigation. The goodness-of-fit techniques applied to test H_0 are based on measuring in some way the conformity of the sample data (a set of x-values) to the hypothesized distribution, or equivalently, its discrepancy from it.

Some of the popular techniques discussed in literature for goodness-of-fit problem are: tests of chi-squared type, test based on empirical distribution function; characteristic function; moment-generating function, test based on regression, correlation, moments, test based on transformation methods, etc. In the course of his Mathematical Contributions to the Theory of Evolution, Karl Pearson abandoned the assumption that biological populations are normally distributed, introducing the Pearson system of distributions to provide other models. The need to test fit arose naturally in this context, and in 1900 Pearson invented his chi-squared test. This test and others related to it remain among the most used statistical procedures. Modern developments have increased the flexibility of chi-squared test, especially when unknown parameters are to be estimated in the hypothesized family. Log-likelihood ratio, Neymann modified chi-squared and Freeman–Tukey test play classical role in chi-squared type test. The most well-known empirical distribution

*Corresponding author. Email: brd_stats@yahoo.in

function test is introduced by Kolmogorov–Smirnov. For testing many distributional families, Stephens [1] has given modifications for empirical distribution function statistics. A comprehensive review of the theory of empirical distribution function tests is given in Durbin.[2] An extensive review of literature on goodness-of-fit techniques is given in D'Agostino and Stephens.[3] The rest of the article is organized as follows.

In Section 2, we discuss the test due to Damico [4] and in Section 3 we propose a modified test for goodness-of-fit. In Section 4, a Monte Carlo study is done to estimate power of the modified test for various alternatives. Section 5 gives concluding remarks.

2. Test based on *A*-statistic

For testing goodness-of-fit of a completely specified univariate distribution, Damico [4] has proposed a one-sample test for goodness-of-fit. The test is easy to describe and compute and so is a useful teaching tool. Damico [4] uses a simple technique where one divides the probability range into n intervals of equal probability mass $1/n$, and verifies whether the hypothesized distribution evaluated at the observed data would place one observation into each interval. Consider the problem of testing the following null hypothesis,

H_0: A random sample of n X-values comes from a completely specified distribution $F(\bullet)$.

The test statistic proposed by Damico [4] for testing H_0 is

$$A = \sum_{i=1}^{n} |\text{Gif}(n \times F1) - i|,$$

where Gif$(\bullet)$ is the greatest integer function and F is the cumulative distribution function.

Goodness-of-fit test based on A-statistic has been studied and simulated powers are given by Damico.[4] In the following section, we extend Damico's idea and obtain a modified test statistic.

3. Test based on *T*-statistic

While defining the A-statistic, Damico [4] assumes one observation from the sample to occur in each of the n intervals under the null hypothesis. In the following, we have modified the A-statistic by allowing for m intervals of probability mass r/n, where $r \geq 1$ and $n = m \times r$. Further, we verify whether the hypothesized distribution evaluated at the observed data would place r observations in each interval. To test H_0, we suggest the following modified test statistic:

$$T = \sum_{k=1}^{m} |S_k - r \times k|,$$

where $|\cdot|$ is an absolute function. Also,

$$S_k = \sum_{i=(k-1)r+1}^{kr} \text{Gif}(m * F(X_{(i)}) + 1), \quad k = 1, 2, 3, \ldots, m.$$

It is clear that for r equal to one, the statistics T and A are identical. The procedure of understanding the modified test statistic is as follows:

(a) Arrange the given values in ascending order $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$.

(b) Compute $F(X_{(i)})$ and Gif$(m \times F(X_{(i)}) + 1)$, $i = 1, 2, \ldots, n$.
(c) Compute S_k, $k = 1, 2, \ldots, m$.
(d) Compute the T-statistic.

Large values of T indicate that the sample is not from the hypothesized distribution. Therefore, we reject the null hypothesis at the significant level α, if $T \geq C_\alpha$. The critical point C_α is determined by the αth quantile of the distribution of the T-statistic by means of Monte Carlo simulations.

In Tables 1 and 2, we present the results of Monte Carlo study conducted at a α-nominal level with 10,000 replications to assess the empirical critical values of T-statistic for r equal to 2 and 3, respectively. In each case, the four α levels were 0.20, 0.10, 0.05 and 0.01. A code in *R* was written to compute the empirical critical values.

The following example illustrates the procedure of finding the T-statistic for r equal to two. Suppose we have a random sample comprising the following 10 values: 0.018, 0.026, 0.277, 0.306, 0.426, 0.479, 0.502, 0.551, 0.720 and 0.892. We wish to test the hypothesis that these 10 values were drawn from a uniform distribution over (0, 1). We begin by defining five equal and non-over-lapping intervals and finding the number of observations in each. Further, we find the number of moves required to produce the ground state (i.e. two observations per interval).

Interval	Frequency	First move	Second move	Third move
(0.0, 0.2)	2	2	2	2
(0.2, 0.4)	2	2	2	2
(0.4, 0.6)	4	3	2	2
(0.6, 0.8)	1	2	3	2
(0.8, 1.0)	1	1	1	2

Table 1. Critical values for T-statistic ($r = 2$).

n	Cr. value T^*	P $[T \geq T^*]$	n	Cr. value T^*	P $[T \geq T^*]$	n	Cr. value T^*	P $[T \geq T^*]$	n	Cr. value T^*	P $[T \geq T^*]$
4	1	0.6288		19	0.0477		56	0.0105	50	73	0.1873
	2	0.1227		24	0.0089	30	34	0.1872		88	0.1013
	3	0.0000	18	16	0.1784		41	0.1015		99	0.0594
6	3	0.2250		19	0.1000		47	0.0496		131	0.0096
	4	0.0769		22	0.0521		61	0.0109	60	95	0.2032
	5	0.0179		29	0.0092	32	37	0.1921		115	0.0978
	7	0.0000	20	18	0.2061		44	0.1012		134	0.0508
8	5	0.1636		22	0.1070		52	0.0502		174	0.0094
	6	0.0830		26	0.0484		67	0.0092	70	119	0.1997
	7	0.0367		33	0.0105	34	41	0.2071		147	0.0921
	8	0.0137	22	21	0.1951		49	0.1009		169	0.0499
10	6	0.2547		26	0.0929		58	0.0480		218	0.0103
	8	0.1006		29	0.0560		75	0.0099	80	147	0.2008
	9	0.0559		38	0.0118	36	44	0.1953		179	0.0986
	12	0.0095	24	24	0.1971		55	0.0900		209	0.0497
12	8	0.2353		29	0.1029		63	0.0492		268	0.0104
	11	0.0794		34	0.0544		81	0.0099	90	175	0.1985
	12	0.0514		44	0.0112	38	48	0.1939		211	0.1014
	15	0.0130	26	27	0.2004		58	0.0998		246	0.0507
14	11	0.1853		33	0.1024		67	0.0490		289	0.0167
	13	0.1067		38	0.0541		87	0.0095	100	202	0.2060
	15	0.0511		49	0.0098	40	52	0.1877		249	0.1002
	19	0.0130	28	30	0.1985		62	0.0910		292	0.0499
16	13	0.1978		37	0.0981		73	0.0522		381	0.0100
	16	0.0998		43	0.0525		95	0.0100			

Hence, the number of moves required to get two observations per interval is 3. The mathematical method of understanding T-statistic is simple. First find $\text{Gif}(m \times F(X_{(i)}) + 1)$, $i = 1, 2, \ldots, 10$ and then S_k, $k = 1, 2, \ldots, 5$. So, for our example:

i	$X_{(i)}$	$F(X_{(i)})$	$\text{Gif}(m \times F(X_{(i)}) + 1)$	k	S_k	$\lvert S_k - r \times k \rvert$
1	0.018	0.018	$\text{Gif}(5 \times 0.018 + 1) = 1$	1	$\text{Gif}(5 \times 0.018 + 1) +$	$\lvert 2 - 2 \times 1 \rvert$
2	0.026	0.026	$\text{Gif}(5 \times 0.026 + 1) = 1$		$\text{Gif}(5 \times 0.026 + 1) = 2$	$= 0$
3	0.277	0.277	$\text{Gif}(5 \times 0.277 + 1) = 2$	2	$\text{Gif}(5 \times 0.277 + 1) +$	$\lvert 4 - 2 \times 2 \rvert$
4	0.306	0.306	$\text{Gif}(5 \times 0.306 + 1) = 2$		$\text{Gif}(5 \times 0.306 + 1) = 4$	$= 0$
5	0.426	0.426	$\text{Gif}(5 \times 0.426 + 1) = 3$	3	$\text{Gif}(5 \times 0.426 + 1) +$	$\lvert 6 - 2 \times 3 \rvert$
6	0.479	0.479	$\text{Gif}(5 \times 0.479 + 1) = 3$		$\text{Gif}(5 \times 0.479 + 1) = 6$	$= 0$
7	0.502	0.502	$\text{Gif}(5 \times 0.502 + 1) = 3$	4	$\text{Gif}(5 \times 0.502 + 1) +$	$\lvert 6 - 2 \times 4 \rvert$
8	0.551	0.551	$\text{Gif}(5 \times 0.551 + 1) = 3$		$\text{Gif}(5 \times 0.551 + 1) = 6$	$= 2$
9	0.720	0.720	$\text{Gif}(5 \times 0.720 + 1) = 4$	5	$\text{Gif}(5 \times 0.720 + 1) +$	$\lvert 9 - 2 \times 5 \rvert$
10	0.892	0.892	$\text{Gif}(5 \times 0.892 + 1) = 5$		$\text{Gif}(5 \times 0.892 + 1) = 9$	$= 1$

The computed value of the T-statistic is $0 + 0 + 0 + 2 + 1 = 3$. The probability under the null hypothesis that the T-statistic assumes a value ≥ 3 is 0.7275. This α-level would generally not be considered significant, and so the null hypothesis would not be rejected.

4. Performance study of the test based on T-statistic

While studying the performance of A-statistic, Damico [4] has used several statistical tests that first appeared in Stephens.[1] These statistical tests are Kolmogorov–Smirnov (D), Cramér–von Mises (W^2), Kuiper (V), Watson (U^2), Anderson–Darling (A^2), $Q(= \sum_i \ln Z_i)$ and chi-square. We have studied the performance of test based on T-statistic for r equal to 1, 2, 3, 4 and 5. The null hypothesis is that we have a uniform random number on the interval (0, 1). The seven alternative distributions which have been considered by Damico [4] for studying power of the test statistic are as follows:

$$F : F(x) = 1 - (1 - x)^k, \quad 0 \leq x \leq 1$$

for k equal to 1.5 and 2,

$$G : F(x) = \begin{cases} 2^{(k-1)}x^k, & 0 \leq x \leq 0.5 \\ 1 - 2^{(k-1)}(1 - x)^k, & 0.5 \leq x \leq 1 \end{cases}$$

for k equal to 1.5, 2 and 3,

$$H : F(x) = \begin{cases} (0.5 - x)^k, & 0 \leq x \leq 0.5, \\ 0.5 + 2^{(k-1)}(x - 0.5)^k, & 0.5 \leq x \leq 1 \end{cases}$$

for k equal to 1.5 and 2.

According to Stephens,[1] alternative F gives points closer to zero than expected under the hypothesis of uniformity, whereas G gives points near to 0.5 and H gives two clusters (close to 0 and 1). The same set of alternatives is used to study the performance of the test based on T-statistic. An

Table 2. Critical values for T-statistic ($r = 3$).

n	Cr. value T^*	P $[T \geq T^*]$	n	Cr. value T^*	P $[T \geq T^*]$	n	Cr. value T^*	P $[T \geq T^*]$
6	1	0.6914	39	33	0.1980	72	82	0.2009
	2	0.2255		40	0.1030		100	0.1000
	3	0.0336		47	0.0520		118	0.0504
	4	0.0000		61	0.0110		152	0.0100
9	4	0.1560	42	37	0.1907	75	88	0.1991
	5	0.0609		45	0.1018		106	0.0998
	7	0.0061		53	0.0515		124	0.0492
	8	0.0000		68	0.0099		160	0.0102
12	6	0.1719	45	41	0.2004	78	94	0.2012
	7	0.0990		51	0.0960		115	0.0988
	8	0.0527		58	0.0523		133	0.0493
	10	0.0123		77	0.0097		172	0.0102
15	8	0.1936	48	45	0.2020	81	99	0.2028
	10	0.0876		55	0.1018		121	0.0993
	11	0.0576		65	0.0478		142	0.0489
	14	0.0118		83	0.0098		181	0.0101
18	10	0.2205	51	49	0.1992	84	103	0.2050
	13	0.0910		60	0.1018		126	0.1011
	15	0.0460		71	0.0496		146	0.0492
	19	0.0091		92	0.0102		194	0.0099
21	13	0.1979	54	54	0.2002	87	110	0.2000
	16	0.0984		66	0.1008		134	0.1011
	19	0.0428		78	0.0490		155	0.0501
	24	0.0095		99	0.0104		204	0.0098
24	16	0.1890	57	58	0.2080	90	117	0.2000
	19	0.1060		71	0.1017		142	0.1006
	23	0.0450		82	0.0524		167	0.0494
	29	0.0100		107	0.0098		214	0.0101
27	19	0.1990	60	63	0.1979	93	120	0.2015
	24	0.0910		76	0.1007		148	0.0989
	27	0.0540		89	0.0492		173	0.051
	35	0.0100		117	0.0098		220	0.0099
30	22	0.2070	63	68	0.1971	96	128	0.2018
	28	0.0920		81	0.1090		156	0.0987
	32	0.0510		97	0.0498		181	0.0502
	42	0.0100		126	0.0101		232	0.0101
33	26	0.1940	66	73	0.1965	99	134	0.2009
	32	0.0930		88	0.1005		163	0.1015
	37	0.0500		103	0.0487		190	0.0504
	48	0.0090		132	0.0105		244	0.0100
36	29	0.1970	69	78	0.1977			
	36	0.0950		94	0.1023			
	41	0.0500		111	0.0497			
	52	0.0110		145	0.0101			

empirical study was conducted for the power estimates of the test for different values of r and sample sizes. Along with the power estimates, the mean and standard deviation of the T-statistic were also recorded. Table 3 shows the power estimates of the test based on T-statistic for different values of r (including $r = 1$) for F, G and H alternatives, respectively, for the nominal level 10%. The mean and standard deviation of the T-statistic for different values of r (including $r = 1$), for F, G and H alternatives, respectively, are given in Table 4. The entries in Tables 3 and 4 are proportion of 10,000 Monte Carlo samples that resulted in rejection of H_0. The sample sizes are selected so as to cover the cases of r equal to 2, 3, 4 and 5. The performance of Kolmogorov–Smirnov (D), Cramér–von Mises (W^2), Kuiper's (V), Watson (U^2), Anderson–Darling (A^2), $Q(= \sum_i \ln Z_i)$ and chi-square tests are not included in the tables as we are interested in comparing the performance

Table 3. Power comparisons for different values of r (α-level 0.10).

r	1	2	3	4	5	r	1	2	3	4	5
n	Alternative $F_{k=1.5}$					n	Alternative $F_{k=2}$				
12	0.317	0.353	0.332	0.327	–	12	0.664	0.703	0.649	0.656	–
18	0.411	0.429	0.404	–	–	18	0.825	0.837	0.812	–	–
20	0.447	0.465	–	0.448	0.419	20	0.867	0.882	–	0.861	0.842
24	0.522	0.536	0.500	0.493	–	24	0.925	0.928	0.912	0.908	–
30	0.609	0.608	0.616	–	0.604	30	0.963	0.963	0.969	–	0.961
36	0.681	0.680	0.678	0.675	–	36	0.987	0.987	0.986	0.985	–
40	0.726	0.742	–	0.738	0.727	40	0.993	0.994	–	0.993	0.992
42	0.744	0.744	0.752	–	–	42	0.994	0.994	0.995	–	–
48	0.804	0.808	0.802	0.803	–	48	0.997	0.997	0.997	0.996	–
n	Alternative $G_{k=1.5}$					n	Alternative $G_{k=2}$				
12	0.078	0.117	0.090	0.101	–	12	0.134	0.197	0.148	0.190	–
18	0.092	0.114	0.102	–	–	18	0.224	0.268	0.225	–	–
20	0.109	0.130	–	0.130	0.111	20	0.295	0.333	–	0.316	0.245
24	0.128	0.146	0.120	0.118	–	24	0.389	0.421	0.358	0.343	–
30	0.165	0.164	0.178	–	0.130	30	0.548	0.547	0.561	–	0.435
36	0.203	0.203	0.200	0.198	–	36	0.672	0.670	0.667	0.648	–
40	0.226	0.251	–	0.248	0.230	40	0.745	0.770	–	0.753	0.730
42	0.241	0.254	0.250	–	–	42	0.778	0.777	0.781	–	–
48	0.291	0.308	0.296	0.292	–	48	0.857	0.866	0.857	0.855	–
n	Alternative $G_{k=3}$					n	Alternative $H_{k=1.5}$				
12	0.424	0.510	0.341	0.501	–	12	0.159	0.162	0.138	0.149	–
18	0.736	0.776	0.689	–	–	18	0.162	0.165	0.141	–	–
20	0.837	0.860	–	0.816	0.655	20	0.174	0.176	–	0.149	0.126
24	0.932	0.940	0.900	0.872	–	24	0.188	0.188	0.150	0.135	–
30	0.988	0.988	0.988	–	0.955	30	0.221	0.208	0.199	–	0.140
36	0.997	0.997	0.996	0.996	–	36	0.243	0.240	0.214	0.198	–
40	0.999	0.999	0.999	0.999	–	40	0.267	0.269	–	0.250	0.226
42	0.999	0.999	0.999	0.999	–	42	0.274	0.271	0.265	–	–
48	0.999	0.999	0.999	0.999	–	48	0.315	0.319	0.297	0.288	–
n	Alternative $H_{k=2}$										
12	0.237	0.229	0.159	0.157	–						
18	0.311	0.289	0.211	–	–						
20	0.358	0.334	–	0.216	0.155						
24	0.437	0.424	0.319	0.258	–						
30	0.581	0.535	0.501	–	0.300						
36	0.682	0.648	0.609	0.557	–						
40	0.751	0.756	–	0.677	0.612						
42	0.773	0.761	0.741	–	–						
48	0.848	0.848	0.826	0.805	–						

of the proposed test for different values of r with the test due to Damico.[4] The power of the T-statistic for r less than three compares very favourably with both the Kolmogorov–Smirnov (D) statistic and the Cramer–von Mises (W^2) statistic for almost all alternatives.

5. Concluding remarks

Although the technique of partitioning the range of the probability distribution is same as that of the chi-square test, the test due to Damico [4] is superior for small samples. The test proposed here is modified version of the test due to Damico [4] for more than one observation per interval.

Table 4. Mean and standard deviation of T-statistic for different values of r.

r	1	2	3	4	5	r	1	2	3	4	5
n	Alternative $F_{k=1.5}$					n	Alternative $F_{k=2}$				
12	017.26	008.26	005.23	003.67	–	12	024.68	012.00	007.69	005.44	–
	(08.07)	(04.15)	(02.81)	(02.11)	–		(08.78)	(04.44)	(02.95)	(02.22)	–
18	036.10	017.65	011.47	–	–	18	054.68	026.99	017.67	–	–
	(15.84)	(08.08)	(05.42)	–	–		(16.78)	(08.43)	(05.63)	–	–
20	043.61	021.71	–	010.47	008.04	20	067.57	033.32	–	016.16	012.61
	(18.78)	(09.77)	–	(04.88)	(03.94)		(20.08)	(09.96)	–	(05.02)	(04.03)
24	062.19	030.70	020.14	014.85	–	24	097.24	048.27	031.86	023.56	–
	(25.93)	(13.12)	(08.81)	(06.64)	–		(26.88)	(13.48)	(08.99)	(06.72)	–
30	094.64	047.09	031.04	–	018.03	30	150.32	075.27	049.75	–	029.14
	(37.40)	(18.88)	(12.60)	–	(07.65)		(37.87)	(19.05)	(12.47)	–	(07.67)
36	134.74	066.97	044.32	032.98	–	36	217.19	108.23	071.88	053.57	–
	(50.59)	(25.45)	(17.02)	(12.83)	–		(49.69)	(24.88)	(16.59)	(12.46)	–
40	165.01	082.11	–	040.28	032.25	40	267.23	133.77	–	066.10	052.61
	(59.92)	(30.54)	–	(14.92)	(12.01)		(58.63)	(29.12)	–	(14.56)	(11.67)
42	182.29	090.73	060.15	–	–	42	294.66	146.99	097.69	–	–
	(65.45)	(32.89)	(21.98)	–	–		(63.36)	(31.69)	(21.15)	–	–
48	236.35	117.75	078.17	058.35	–	48	385.45	192.40	127.95	095.61	–
	(79.17)	(39.75)	(26.57)	(19.99)	–		(78.63)	(39.33)	(26.22)	(19.64)	–
n	Alternative $G_{k=1.5}$					n	Alternative $G_{k=2}$				
12	013.09	006.23	003.91	002.72	–	12	015.32	007.36	004.62	003.24	–
	(04.96)	(02.68)	(01.88)	(01.48)	–		(04.50)	(02.42)	(01.77)	(01.45)	–
18	025.87	012.56	008.13	–	–	18	032.14	015.77	010.18	–	–
	(09.07)	(04.75)	(03.31)	–	–		(08.35)	(04.35)	(03.03)	–	–
20	031.10	015.15	–	007.12	005.46	20	039.17	019.28	–	009.18	006.96
	(10.68)	(05.57)	–	(02.97)	(02.43)		(09.88)	(05.13)	–	(02.78)	(02.31)
24	042.31	020.75	013.55	009.92	–	24	054.93	027.19	017.83	013.07	–
	(14.03)	(07.23)	(04.96)	(03.84)	–		(13.00)	(06.65)	(04.57)	(03.53)	
30	062.05	030.74	020.22	–	011.54	30	083.44	041.49	027.37	–	015.97
	(19.52)	(10.10)	(06.86)	–	(04.23)		(18.37)	(09.23)	(06.31)	–	(4.01)
36	085.77	042.44	027.97	020.71	–	36	118.40	058.92	039.00	028.97	–
	(26.03)	(13.24)	(08.96)	(06.83)	–		(23.88)	(12.07)	(08.16)	(06.22)	–
40	103.41	051.26	–	025.12	019.80	40	144.93	072.15	–	035.55	028.12
	(30.60)	(15.56)	–	(08.00)	(06.45)		(28.42)	(14.30)	–	(07.31)	(06.00)
42	113.40	056.25	037.17	–	–	42	158.90	079.14	052.46	–	–
	(33.54)	(17.02)	(11.48)	–	–		(30.62)	(15.41)	(10.36)	–	–
48	143.60	071.28	047.23	035.14	–	48	206.00	102.70	068.18	050.86	–
	(40.60)	(20.55)	(13.85)	(10.46)	–		(37.96)	(19.12)	(12.82)	(09.70)	–
n	Alternative $G_{k=3}$					n	Alternative $H_{k=1.5}$				
12	019.63	009.50	005.90	004.58	–	12	016.80	007.46	004.25	002.59	–
	(03.88)	(02.12)	(01.57)	(01.47)	–		(06.45)	(03.39)	(02.35)	(01. 98)	–
18	043.01	021.29	013.76	–	–	18	027.84	013.09	008.13	–	–
	(07.11)	(03.73)	(02.61)	–	–		(11.56)	(05.94)	(04.06)	–	–
20	052.81	026.13	–	012.80	009.31	20	033.05	015.65	–	006.87	004.98
	(08.33)	(04.28)	–	(02.54)	(02.02)		(13.43)	(06.90)	–	(03.62)	(02.99)
24	075.39	037.42	024.60	017.97	–	24	045.11	021.61	013.78	009.63	–
	(10.96)	(05.62)	(03.86)	(03.02)	–		(17.32)	(08.98)	(06.03)	(04.63)	–
30	116.74	058.02	038.39	–	022.14	30	065.76	032.05	020.62	–	011.26
	(15.41)	(07.90)	(05.39)	–	(03.38)		(24.14)	(12.36)	(08.25)	–	(05.12)
36	166.60	083.04	055.02	041.15	–	36	089.93	043.48	028.20	020.48	–
	(20.39)	(10.29)	(06.96)	(05.42)	–		(31.18)	(15.84)	(10.84)	(08.04)	–
40	205.42	102.44	–	050.63	040.40	40	108.82	053.23	–	025.31	019.63
	(24.03)	(12.11)	–	(06.21)	(05.05)		(36.37)	(18.35)	–	(09.33)	(07.54)
42	226.40	112.90	074.94	–	–	42	117.00	057.58	037.67	–	–
	(25.69)	(12.94)	(08.73)	–	–		(38.65)	(19.52)	(13.34)	–	–
48	293.80	146.70	097.50	072.75	–	48	148.70	073.02	047.78	034.99	–
	(31.43)	(15.82)	(10.63)	(08.05)	–		(47.43)	(23.85)	(15.97)	(12.04)	–

(*Continued*)

Table 4. Continued.

r	1	2	3	4	5
n			Alternative $H_{k=2}$		
12	016.83	007.49	004.24	002.59	–
	(06.41)	(03.40)	(02.33)	(01.97)	–
18	034.32	016.12	009.86	–	–
	(11.18)	(05.86)	(03.93)	–	–
20	041.45	019.54	–	008.29	005.94
	(12.71)	(06.51)	–	(03.43)	(02.79)
24	058.24	027.76	017.52	012.25	–
	(16.72)	(08.33)	(05.66)	(04.37)	–
30	087.12	042.15	027.11	–	014.64
	(22.34)	(11.33)	(07.70)	–	(04.73)
36	122.22	059.63	038.56	028.00	–
	(28.93)	(14.55)	(09.76)	(07.44)	–
40	148.49	072.68	–	034.45	026.61
	(32.74)	(16.42)	–	(08.26)	(06.67)
42	162.19	079.82	052.08	–	–
	(35.92)	(17.79)	(12.01)	–	–
48	209.03	103.43	067.73	049.42	–
	(43.76)	(21.89)	(14.51)	(10.90)	–

Note: Value in the bracket is the standard deviation of T-statistic.

The modification reduces computational work as compared with test proposed by Damico,[4] as the number of observation per interval increases without further loss of power. While considering the power performance of the test for different values of r, we observe that the test performs better for two observations per interval as compared with one observation per interval for all F_k, G_k and H_k alternatives except for an alternative H_k with $k = 2$ and sample sizes considered for the study. The estimates of power decreases for r equal to three and above for almost all alternatives.

The test statistic is designed as a general technique for testing the goodness-of-fit of completely specified distribution. One can study the performance of the test even if parameters are to be estimated for a particular probability distribution. If the sample size is a prime number then further modification of the proposed test could be a topic of future research.

Acknowledgements

The authors are grateful to the editor and the expert referee for making constructive and valuable comments that have significantly improved the contents of this article.

References

[1] Stephens MA. EDF statistics for goodness-of-fit and some comparisons. J Am Stat Assoc. 1974;69(347): 730–737.
[2] Durbin J. Distribution theory for tests based on the sample distribution function. Philadelphia: SIAM; 1973.
[3] D'Agostino RB, Stephens MA. Goodness-of-fit techniques. New York: Marcel Dekker inc; 1986.
[4] Damico J. A new one-sample test for goodness-of-fit. Commun Stat – Theory Methods. 2004;33:181–193.

Int J Adv Manuf Technol (2015) 78:1305–1314
DOI 10.1007/s00170-014-6735-1

ORIGINAL ARTICLE

Fraction nonconforming control charts with m-of-m runs rules

S. K. Khilare · D. T. Shirke

Received: 16 December 2011 / Accepted: 18 December 2014 / Published online: 7 January 2015

Abstract In this article, we proposed m-of-m control chart for fraction nonconforming to increase sensitivity of the standard p-chart. The m-of-m control chart for fraction nonconforming is described, and the 3-of-3 control chart is discussed in detail as an illustration of the m-of-m control chart. The performance of proposed control charts is measured in terms of average run length, standard deviation of run length and quartiles under zero-state and steady-state modes. The Markov chain approach is used to compute average run length, standard deviation of run length and quartiles. Comparison study revealed that the performance proposed m-of-m control chart with $m=2, 3$ is significantly better than the standard p-chart. Average run length values of the proposed control charts are at least 15 % less than that of the standard p-chart values for small to moderate shifts. Standard deviation of run length values of the proposed control charts is also at least 18 % less than the standard p-chart. An example is given to illustrate an application of procedures discussed.

Keywords Fraction nonconforming · Zero state · Steady state · Markov chain · Runs rules · Warning limits

1 Introduction

The Shewhart's p chart is widely used in the industry to monitor the fraction nonconforming units in a process. Nonconforming unit is a product which fails to meet at least one specified requirements. The control limits and performance study of the fraction nonconforming control chart are typically based on the binomial distribution. The sample fraction nonconforming is defined as the ratio of the number of nonconforming units (X) in the sample to the sample size (n). That is,

$$\hat{p} = \frac{X}{n}. \tag{1}$$

It is clear that X follows binomial distribution with parameters n and p, where p is the probability that a unit is nonconforming. The objective of a control chart is to control the quality of the characteristic or to detect quickly an increase in a process fraction nonconforming (p). When the process is in the state of in-control, the mean and variance of $\hat{p}$ are $\mu_0 = p_0$ and $\sigma_0^2 = \frac{p_0(1-p_0)}{n}$, respectively, where p_0 is the fraction nonconforming in the production process when the process is in the in-control state. If p_0 is unknown, it will be estimated from the observed data. The 3σ control limits for p using normal approximation are given by

$$\begin{aligned} \mathrm{UCL} &= p_0 + 3\sqrt{p_0(1-p_0)/n}, \\ \mathrm{CL} &= p_0, \\ \mathrm{LCL} &= p_0 - 3\sqrt{p_0(1-p_0)/n}. \end{aligned}$$

Alternatively, the chart could be based on standardized statistic Z, where Z is defined as follows:

$$Z_j = \frac{\hat{p}_j - p_0}{\sqrt{p_0(1-p_0)/n}}, \quad j = 1, 2, 3, \ldots \tag{2}$$

Here, Z_j is approximately distributed as a standard normal variate.

The Shewhart $\overline{X}$ and p control charts are most popular control charts, respectively, for monitoring mean and fraction nonconforming of a process distribution. The Shewhart standard control charts are based on the three sigma control limits and give out-of-control signal if a single point plots outside the control limits. To detect large shifts in a process, the Shewhart

S. K. Khilare (✉)
Department of Statistics, R.B.N.B. College, Shrirampur 413709, MS, India
e-mail: shashi.khilare@gmail.com

D. T. Shirke
Department of Statistics, Shivaji University, Kolhapur 416004, MS, India
e-mail: dtshirke@gmail.com

Production, v. 25, n. 4, p. 739-749, out./dez. 2015
http://dx.doi.org/10.1590/0103-6513.113112

Steady-state behavior of nonparametric control charts using sign statistic

Shashikant Kuber Khilare[a]*, Digambar Tukaram Shirke[b]

[a]*Raobahadur Narayanrao Borawake College, Shrirampur, India, shashi.khilare@gmail.com
[b]Shivaji University, Kolhapur, India

Abstract

If process is running for a long period in an in-control condition, it will reach in a steady-state condition. In order to study the long term properties of a control chart, it is appropriate to investigate the steady-state average time to signal. In this article, we discussed runs rules representation of a nonparametric synthetic control chart using sign statistic for detecting shifts in location parameter. We compared zero-state average time to signal with steady-state average time to signal of the synthetic control chart for symmetric and asymmetric distributions. We also present the m-of-m control chart using sign statistic. For comparison study, we computed average time to signal of the m-of-m control chart, the sign chart (1-of-1 chart) and the synthetic control chart for normal, Cauchy, double exponential and gamma distributions. Steady-state and zero-state performance of the m-of-m control chart with m = 2, 3 compared with the sign chart (1-of-1 chart) and synthetic control chart. The zero-state and steady-state average time to signal of the synthetic and the m-of-m control charts computed using Markov chain approach.

Keywords

Steady-state. Markov chain. Synthetic. Nonparametric. Average time to signal.

1. Introduction

In a process control environment with variables data, it is assumed that the process output follow the normal distribution. The statistical properties of commonly employed control charts such as the Shewhart $\overline{X}$ chart, the cumulative sum control chart and the exponentially weighted moving average control chart are the exact only if assumption of normality is satisfied. If the underlying process distribution is non-normal, performance of these charts are not up to the mark. Such considerations provide reasons for the development and applications of control charts that are not specifically designed under the assumption of normality or any other parametric distribution. When the distribution of process output is non-normal, distribution-free or nonparametric control charts can be useful.

Nonparametric control charts are used for detecting the changes in the process median (or mean) or changes in the process variability. Most of the control charts are based on the sample means when observations are taken sequentially under the normality condition. If the distribution of observations is non-normal then the central limit theorem is usually used to justify the assumption that the distribution of sample mean is approximately normal. The nonparametric control charts used for monitoring the process median (or mean) based on the signs computed within samples and used in place of sample means in the Shewhart chart. The chart is labelled to be the nonparametric chart if in-control average time to signal (ATS) does not depend on the underlying process distribution. In case of charts based on signs, ATS will be same for all distributions for which median equals to the target value. In nonparametric control charts the assumption of normality is not necessary for calculating the control limits. Another advantage is that the nonparametric control charts are usually more efficient than the charts based on $\overline{X}$ when the distribution of the observations is heavy tailed, that is when observations in the tails of the distribution have

Received: Dec. 4, 2012; Accepted: Apr. 21, 2014

a higher probability than for normal distribution. In nonparametric control charts variance of the process need not to be known or estimated in order to apply the control chart. In fact, these control charts for controlling median are not affected by changes in the variance as long as location parameter is constant. The nonparametric control charts may be particularly useful when a process is just start up. It is desirable to apply control charts before there is an enough data to get a reasonable estimate of variance and/or assess the normality of the process.

In quality control applications McGilchrist & Woodyer (1975) proposed a distribution-free cumulative sum technique for monitoring rainfall amounts. Bakir (2006) developed distribution-free quality control charts based on signed-rank-like statistic. Bakir (2004) proposed a distribution-free Shewhart quality control chart based on signed-ranks. Bakir & Reynolds Junior (1979) studied a nonparametric procedure for process control based on within-group ranking. Amin & Searcy (1991) studied the behavior of the EWMA control chart using the Wilcoxon signed-rank statistic. Amin et al. (1995) developed the nonparametric quality control charts based on the sign statistic. Chakraborti & Eryilmaz (2007) proposed control charts based on signed-rank statistic. Chakraborti & Van de Wiel (2008) proposed Mann-Whiteny statistic based control chart. Human et al. (2010) studied nonparametric Shewhart-type sign control charts based on runs. Ho & Costa (2011) proposed monitoring a wandering mean with an np chart and this chart is also work with sign statistics. Crosier (1986) suggested a technique for obtaining steady-state ARL of CUSUM chart using the Markov chain approach. Saccucci & Lucas (1990) given a FORTRAN computer program for the computation of ARL of EWMA and combined Shewhart-EWMA control schemes. The program calculates zero-state and steady-state ARL using the Markov chain approach. Champ (1992) computed steady-state ARL of Shewhart control chart with supplementary runs rules. Davis & Woodall (2002) studied the steady-state properties of synthetic control chart to monitor shifts in process mean. Lim & Cho (2009) developed a control charts with m-of-m runs rules to study the economical-statistical properties of control chart using steady-state ARL.

The rest of article is organized as follows:

Section 2 gives the Shewhart charts using sign statistic. Section 3 gives conforming run length control chart. In Section 4, operations and design procedure of synthetic control chart using sign statistic are given and also in this we explained the Markov chain model and steady-state ATS of synthetic control chart. In Section 5, we present m-of-m runs rules schemes using sign statistic. In this Section, we also study steady-state and zero-state ATS performance of the m-of-m chart for process median. Section 6 gives conclusions.

The Shewhart control chart using sign statistic is explained in brief in following section.

2. Shewhart chart using sign statistic

Let X be a continuous random variable with cumulative distribution function (c.d.f.) F(.). Let μ and μ_0 be the median and target value of median respectively. A sample of n observations is taken at regular time interval from the process. Let $X_i = (X_{i1}, X_{i2}, ..., X_{in})$ be the sample taken at the i^{th} time point. At any time t, each observation from the sample is compared with target value μ_0 and the number of observations above and below μ_0 is recorded.

Define,

$$sign(X_{ij} - \mu_0) = \begin{cases} 1 & if\ X_{ij} > \mu_0 \\ 0 & if\ X_{ij} = \mu_0 \\ -1 & if\ X_{ij} < \mu_0 \end{cases} \quad (1)$$

where X_{ij} is the j^{th} observation in the i^{th} sample. Since the distribution of observations is assumed to be continuous, $pr(X_{ij} - \mu_0 = 0) = 0$. In practice occasional zero may occur which can be signed alternatively +1 and -1.

Let

$$SN_i = \left\{ \sum_{j=1}^{n} sign(X_{ij} - \mu_0) \qquad i = 1,2,3,... \right. \quad (2)$$

where SN_i is the difference between number of observations above μ_0 and number of observations below μ_0 in the i^{th} sample. A random variable $T_i = SN_i + n/2$ gives the number of positive signs in the sample of size n and has binomial distribution with parameters n and p, where $p = P(X_{ij} > \mu_0)$. As long as median remains at μ_0, we have $p = p_0 = 1/2$. That is, $P(X_{ij} > \mu_0) = P(X_{ij} < \mu_0) = 1/2$ and $E[SN_i] = 0$. The chart signals that shift has occurred if $|SN_i| \geq c$, where $c > 0$ is a specified constant (upper control limit = c and lower control limit = -c). The chart signals that shift has occurred in the positive direction if $SN_i \geq c$ and chart signals that the shift has occurred in the negative direction if $SN_i < -c$.

The largest possible in-control average run length (ARL) values of symmetric one-sided and two-sided control chart are 2^n and 2^{n-1} respectively, when $p = 1/2$ and $SN_i = n$. Unless 'n' is of a moderate size, it may be difficult to achieve even approximately a specified in-control ARL (0).

In following section we discuss conforming run length control chart in detail.

3. The conforming run length control chart

The conforming run length (CRL) chart is proposed by Bourke (1991). The Conforming run length is the number of inspected units between two consecutive nonconforming units including ending nonconforming unit. In Figure 1 below, the white and black circles denote the conforming and nonconforming units respectively. Suppose process start at t=0, then the three samples of CRL are displayed. CRL1=4, CRL2=5, CRL3=3. The idea behind the CRL chart is that the conforming run length will change when the fraction nonconforming in a process p changes. Namely, the CRL is shortened as p increases and is lengthened as p decreases (Figure 1).

The random variable CRL follows a geometric distribution. The probability mass function of CRL is

$$P(CRL) = p(1-p)^{CRL}, \quad CRL = 1,2,3,... \tag{3}$$

The cumulative probability function and mean value of CRL are respectively

$$F(CRL) = 1-(1-p)^{CRL}$$
$$\mu_{CRL} = \frac{1}{p} \tag{4}$$

If CRL is less than lower control limit (L) of CRL chart, then an upward process shift is signaled. Therefore, for detection of an upward process shift (increase in p), a single lower control limit L of CRL chart is sufficient and L can be derived from Equation 4, we have,

$$\alpha_{CRL} = F(L) = 1-(1-p)^{L}$$
$$L = \frac{\ln(1-\alpha_{CRL})}{\ln(1-p_0)} \tag{5}$$

where α_{CRL} is the type-I error probability of the CRL chart and p_0 is the in-control fraction nonconforming. L must be rounded to an integer. If a sample CRL is a less than or equal to the L, then the fraction nonconforming p has increased and out-of-control status will be signaled.

For the CRL chart, ARL_{CRL}, is the average number of CRL samples required to detect out-of-control fraction nonconforming p is given by

$$ARL_{CRL} = \frac{1}{\alpha_{CRL}}$$
$$ARL_{CRL} = \frac{1}{1-(1-p)^{L}} \tag{6}$$

Finally, let ANI_{CRL} be the average number of the inspected units required to signal a fraction nonconforming shift and be equal to the product of μ_{CRL} and ARL_{CRL}.

$$ANI_{CRL} = \mu_{CRL} \times ARL_{CRL}$$
$$ANI_{CRL} = \frac{1}{p} \times \frac{1}{1-(1-p)^{L}} \tag{7}$$

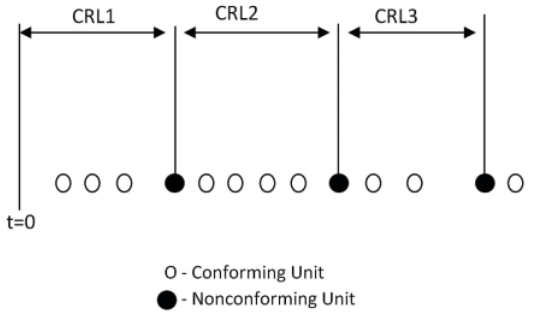


Figure 1. Conforming Run Length.

For CRL chart, if a CRL value falls between lower and upper control limits of the CRL chart, then the process is considered to be under control. However, if CRL value is less than the lower control limit of CRL chart, then upward process shift is signaled and if CRL value greater than upper control limit of CRL chart, then downward process shift is signaled. The presentation of CRL chart usually based on the 100% inspection, because every unit has to be accounted for and classified as either conforming unit or nonconforming one.

In following section we explain synthetic control chart using sign statistic.

4. Synthetic control chart using sign statistic

In the literature, Wu & Spedding (2000) studied the synthetic control chart for detecting small shifts in the process mean. Wu et al. (2001) proposed the synthetic control chart for fraction nonconforming and reported that the synthetic control chart has higher power of detecting out-of-control signal. Wu & Spedding (2001) developed the synthetic control charts for attributes. Khilare & Shirke (2010) proposed a nonparametric synthetic control chart using sign statistic and it performs significantly better than the Shewhart type $\overline{X}$ and sign control charts. The proposed nonparametric synthetic control chart is a combination of the nonparametric sign chart and the *CRL* chart. Basically, the operations of the nonparametric synthetic control chart are similar to that of the synthetic control chart for process mean proposed by Wu & Spedding (2000), except that the

subgroup mean is replaced by the sign statistic SN_i. However, we do not follow the same design procedure due to Wu & Spedding (2000) in order to ensure that the synthetic control chart is nonparametric.

The operations of the synthetic chart using sign statistic are outlined below.

1 Determine sign chart based upper control limit 'c' (> 0), sample size n and CRL based lower control limit (L).

2 Take a sample of 'n' units for inspection and calculate SN_i.

3 If SN_i < c, a sample is a conforming one and control flow goes back to step (2). Otherwise, a sample is a nonconforming one and control flow continues to the next step.

4 Check number of samples between the current and previous nonconforming samples. This number is taken as CRL value for synthetic chart.

5 If CRL > L, then the process is said to be under control and control flow goes back to the step (2). Otherwise the process is taken as out-of-control and control flow continues to the next step.

6 Take action to locate and remove the assignable causes. Then go back to step (2).

4.1. *Design of synthetic control chart*

The synthetic chart has two parameters namely, L and c. For given in-control *ARL* and subgroup sample size n, the parameters L and c are obtained as follows:

Let $ARL_s(\mu)$ be the out-of-control ARL of the synthetic control chart and it is given by

$$ARL_S(\mu) = \frac{1}{P(\delta)[1-(1-P(\delta))^L]} \quad (8)$$

Let $ARL_s(\mu_0)$ be in-control ARL of the synthetic control chart. If $\mu_0 = 0$, then in-control ARL is

$$ARL_S(0) = \frac{1}{P(0)[1-(1-P(0))^L]} \quad (9)$$

and $P(\delta) = Pr(SN_i > c/\mu = \mu_0 + \delta)$.

Here, $P(\delta)$ is the probability that the sample is nonconforming when the permanent upward step shift of δ units occurs. When there is no shift, δ is equal to zero. We note that in Equation 7, "p" is the probability that a unit is nonconforming.

Suppose the desired in-control *ARL* is ARL(0) and the subgroup sample size is n. We compute the $ARL_s(0)$ values using Equation 9 for c = 1, 2, ...,n and L = 1, 2, ... Now choose that pair of (L, c) for which the $ARL_s(0)$ is close to ARL(0). We may note that for a fixed value of c, $ARL_s(0)$ is a decreasing function of L, while for a fixed value of L, $ARL_s(0)$ is a non-decreasing function of c.

Table 1 gives the values of $ARL_s(0)$ for n = 10. As an example, suppose we wish to set ARL(0) = 1024. Then, from Table 1, we see that L = 9 and 8 = 10 is the required pair as the $ARL_s(0)$ corresponding to these values is 1005. Due to the discrete nature of the charting statistic SN_i, for a fixed value of L, we get the same value of $ARL_s(0)$ for two successive values of c (except for c = 1).

The complete design procedure for the synthetic chart can be outlined as below:

1 Specify subsample size n and ARL(0).

2 Initialize L as 1 and $1 \leq c \leq n$.

3 Calculate $ARL_s(0)$ from the current values of L and c using Equation 9.

4 If $ARL_s(0)$ is not close to the specified in-control ARL, increase L by one and go to step 3.

5 If $ARL_s(0)$ is close to the specified in-control ARL, take current values of L and c as final values in the synthetic control chart.

In following section we discuss runs rule representation of the synthetic control chart.

Table 1. In control ARL values for upward sided synthetic control chart for various values of c and L when n = 10.

c↓	L									
	1	2	3	4	5	6	7	8	9	10
1	2.58	1.87	1.70	1.64	1.62	1.61	1.61	1.61	1.61	1.61
2	7.04	4.34	3.50	3.12	2.93	2.82	2.75	2.71	2.69	2.68
3	7.04	4.34	3.50	3.12	2.93	2.82	2.75	2.71	2.69	2.68
4	33.85	18.52	13.47	10.98	9.53	8.59	7.94	7.47	7.12	6.86
5	33.85	18.52	13.47	10.98	9.53	8.59	7.94	7.47	7.12	6.86
6	334.37	171.88	117.78	90.77	74.60	63.85	56.19	50.47	46.04	42.51
7	334.37	171.88	117.78	90.77	74.60	63.85	56.19	50.47	46.04	42.51
8	8665.92	4356.36	2919.89	2201.70	1770.82	1483.60	1278.46	1124.63	1005.00	909.31
9	8665.92	4356.36	2919.89	2201.70	1770.82	1483.60	1278.46	1124.63	1005.00	909.31
10	1048576	524544	349866	262528	210125	175189	150236	131520	116964	105319

4.2. *Runs rule representation of the synthetic control chart*

Davis & Woodall (2002) discussed the runs rule representation of synthetic control chart to detect shifts in the process mean. Here, we discuss the runs rule representation of a nonparametric synthetic control chart for process median using sign statistic. Suppose that each observed sign statistic SN_i is classified as either '0' (conforming) or 1 (nonconforming). If value of sign statistic falls within control limit/limits, the sample is conforming and if it falls out-side the control limit/limits then sample is nonconforming. A sequence of SN_i can be represented by a string of zeros and ones. For example 10001000 would indicate that in a sequence of eight samples, the first and fifth samples are nonconforming samples.

For simplicity, suppose that L = 3. This means that any sequence of SN_i with pattern 1001, 101 or 11 will generate an out-of-control signal for synthetic chart. Note that this sequence also generate signal under the following runs rule:

If two successive sign statistics (SN_i values) fall out-side of the control limits out of L + 1 sign statistics then the two-of-L+1 chart signals an out-of-control status.

On initial pattern of 001, the synthetic control chart will signal using L=3, while two of L + 1 chart would not. The performance of control charts can be made identical over all the samples using head start feature in the runs rule representation; that is , it is assumed that the there is an observation at time zero and that falls out-side of the control limits. With this head start, both charts will signal on initial patterns 1, 01, and 001 but not on the initial pattern 0001. Thus, performance of the charts is now identical for all possible sequences of SN_i. If CRL value is less than or equal to L, then declare that the process is out-of-control. Thus, the synthetic control chart using sign statistic is identical to the above runs rule with the head start a sign statistic at time zero is observed and is nonconforming.

In the following subsection, we present the Markov chain model and ARL results of synthetic control chart.

4.3. *The Markov chain model and steady-state ATS of synthetic control chart*

The formula for ARL can be obtained by using the transition probability matrix (t. p. m.) of an absorbing Markov chain based on the states depending on a lower control limit of the CRL chart.

Consider the case where L = 4. This chart is an identical to a chart which signals if two of the five consecutive sign statistics fall out-sides of the control limits, assuming that a sign statistic at time zero is out-side of control limits.

Let

A= Pr[next observed sign statistic will be within control limit/ limits]

The probability of next observed sign statistic will be within control limits for the change in location parameter is

$$A = Pr[-c < SN_i < c],$$

and for shift in positive direction

$$A = Pr[SN_i \leq c],$$

where, 'c' is a specified constant (control limit of sign control chart) and B= 1- A.

As Davis & Woodall (2002) suggested that the following transition matrix would govern the Markov chain for the synthetic control chart.

- The row contains 'A' in first column and 'B' in second column.
- The last row contains 'A' in first column.
- In all other rows, the entry above the diagonal is 'A'.
- In all other locations, the entry is zero.

Therefore, for example, the transition probability matrix for the synthetic control chart using sign statistic when L= 4 is (Table 2).

With this Markov chain model, the ARL for the zero-state case is

$$ARL = s'(I - R)^{-1}1 \quad (10)$$

where, R is an L+1 by L+1 matrix of probabilities obtained by deleting last row and last column from the above matrix, 1 is column vector of appropriate order having all elements unity and I is an (L + 1) by (L + 1) identity matrix, s is the order (L + 1) of initial probabilities, 1 for initial state and 0 for the

Table 2. The transition probability matrix for the synthetic control chart using sign statistic when L= 4 is:

		States at time t+1					
	↓States→	0000	0001	0010	0100	1000	Signal
States at time t	0000	A	B	0	0	0	0
	0001	0	0	A	0	0	B
	0010	0	0	0	A	0	B
	0100	0	0	0	0	A	B
	1000	A	0	0	0	0	B
	Signal	0	0	0	0	0	1

rest of the cases, s'= [0, 1, 0,..., 0, 0]. Here, '01' corresponds to the initial state. For general values of L, the matrix R (the matrix of probability above with the last row and last column removed) will be an (L + 1) by (L + 1) matrix.

Since the Markov chain representation of the synthetic control chart using sign statistic has more than one absorbing states. The future behavior of the chart can be studied by using steady-state average time to signal (SSARL). If the process is running smoothly for long time, it reaches in the steady-state. The SSARL measures average number of samples required to signal when the effect of head start has disappeared.

Let R_0 be the square matrix obtained from R after dividing each element by the corresponding row sum. Let S be a row vector corresponding to the stationary probability distribution of R_0. The SSARL of the synthetic chart using sign statistic is given by

$$SSARL = S'(I - R_0)^{-1}1 \quad (11)$$

The S can be obtained by solving following equation

$S = R'_0 S$,

subject to

$$\sum_{i=1}^{n} S_i = 1$$

Finally steady-state average time to signal (SSATS) is given by,

$$SSATS = \left(SSARL - \frac{1}{2} \right)(h) \quad (12)$$

Where, sampling interval (h) is adjusted according to the desired rate of false alarms rate. The SSATS measures the average time required to signal a process shift when the effect of head start has disappeared.

We provide steady-state performance of the synthetic control chart in the following section.

4.4. *Steady-state performance of the synthetic control chart*

The objective of control charts is to quickly detect changes in the parameters of the process distribution that are produced by special causes. The ability of a control chart to detect process changes can be measured by the ATS. Thus, the ATS can provide a measure of the time required to detect a special cause when it is present at the time that monitoring starts. Any signal, given when the process is still in control, is a false alarm. In comparison study, we compare zero-state ATS with steady-state ATS of the synthetic control chart. For performance study of the synthetic chart, we consider symmetric distributions namely normal, Cauchy, double exponential distributions and asymmetric gamma distribution. ATS is computed for double exponential distribution, which is symmetric distribution with heavy tails. Cauchy distribution is used because it is symmetric distribution with extremely heavy tails. ATS values computed for each considered distributions with mean zero and variance one. In Bakir (2004) the scale parameter is set to be $\frac{1}{\sqrt{2}}$ for double exponential distribution to achieve variance equal to one. To compute SSATS of Cauchy distribution, scale parameter set to be one and shifts in location parameter. For gamma distribution parameters are set to be 4 (shape parameter) and 1/2 (scale parameter) to achieve mean zero and variance one. Control limits for each control charts are found to be such that the in-control ATS equal to the desired ATS.

Table 3 gives the zero-state and steady-state ATS profile of the synthetic control chart to detect upward shifts in the process median. For the synthetic control chart sample sizes of n = 10 is used. In-control ATS for n =10 is 1024.

The following findings are observed from Table 3.

- Steady-state ATS performance of the synthetic control chart is poor as compare to zero-state for all distributions under study.
- Steady-state performance of the synthetic control chart for double exponential distribution is better than the other distributions under study.

Following section gives the m-of-m control chart using sign statistic for monitoring location parameter.

5. The m-of-m control chart

Consider a control chart with upper control limit (UCL= k) and lower control limit (LCL= -k). Let us consider three regions for the control chart:

- The region between upper control limit and lower limit (region 1).
- The region above upper control limit (region 2).
- The region below lower control limit (region 3).

The probability of a single point falls in the regions 1, 2, 3 are denoted by pc, pu, pl respectively and these probabilities can be computed as follows:

$$pc = \Pr\left[-k < SN_i < k\right],$$
$$= \Pr\left[\frac{-k+n}{2} < T_i < \frac{k+n}{2}\right],$$

Table 3. Zero-state and steady-state ATS profile of the synthetic chart to detect upward shifts in process median (n=10, c = 9, L = 9 and ATS(0) = 1024).

$(\mu - \mu_0)$	Normal Distribution		Cauchy Distribution		Laplace Distribution		Gamma Distribution	
	0SATS	SSATS	0SATS	SSATS	0SATS	SSATS	0SATS	SSATS
0	1024.59	1024.63	1024.59	1024.63	1024.59	1024.63	1024.01	1024.06
0.1	305.76	321.91	386.88	402.54	148.50	163.01	294.28	310.43
0.2	104.90	117.83	161.35	176.20	36.80	44.65	101.63	114.40
0.3	41.40	49.79	74.79	86.04	13.45	17.51	41.48	49.87
0.4	18.75	23.91	38.52	46.59	6.52	8.70	19.67	25.00
0.5	9.69	12.81	21.93	27.66	3.86	5.07	10.65	14.03
0.6	5.64	7.52	13.67	17.78	2.61	3.30	6.47	8.63
0.7	3.64	4.76	9.22	12.21	1.94	2.33	4.31	5.70
0.8	2.55	3.21	6.65	8.87	1.52	1.74	3.10	4.00
0.9	1.90	2.28	5.06	6.74	1.24	1.37	2.36	2.94
1	1.48	1.69	4.04	5.32	1.05	1.12	1.88	2.25

Since,

$$SN_i = 2T_i - n,$$
$$pu = \Pr\left[SN_i \geq k\right],$$
$$pl = \Pr\left[SN_i \leq -k\right].$$

The m-of-m sign chart signals an out-of-control status when a sign statistic falls out-side of the control limits or m-consecutive sign statistics falls beyond the control limits. Suppose {mm} denotes the event when two successive sign statistics fall in region m. The control chart signals an out-of-control status when an event $D = \left\{\underbrace{222...2}_{m-times}, \underbrace{333...3}_{m-times}\right\}$ occurs. To design this control chart we must find appropriate control limits to keep in-control ATS at the desired level.

Now we define states of the Markov chain as follows:

State 1: One point fall between both control limits, {1}.

State 2: One point falls above upper control limit, {2}.

State 3: One point falls below lower control limit, {3}.

State 4: Two consecutive points fall above upper control limit, {22}.

State 5: Two consecutive points fall below lower control limit, {33}.

State 6: Three consecutive points fall above upper control limit, {222}.

State 7: Three consecutive pints fall below lower control limit, {333} and so on.

Finally,

State 2m: Out-of-control (absorbing) state, with associated pattern given by the set D.

The Markov chain representation of chart consist of 2m states with the first (2m – 1) of them being transient. A state is said to be transient state if and only if starting from state one, the probability of returning to state one after some finite length of time is less than one. Then the 2m × 2m transition probability matrix can be partitioned as

$$P = \begin{bmatrix} Q & (I-Q)J \\ 0 & 1 \end{bmatrix}$$

where, Q is the (2m – 1) × (2m – 1) transition probability matrix for the transient sates, I is the (2m – 1) × (2m – 1) identity matrix and J is the column vector of one of an order (2m – 1). The expected value of the run length random variable T is given by

$$E\left[T\right] = e\left(I-Q\right)^{-1} J \qquad (13)$$

where, $e_{1\times 2m-1}$ = (1,0,0,...,0) is the initial distribution. Let M_j be the expected value of the waiting time from state j until the first occurrence of D. Thus, if process is initially in-control, M_1 is the ARL. Let M = (M_1,M_2,...,M_{2m-1}) be the vector of average run lengths. By taking expectations conditional upon the result of the first subgroup these expected values can be found by solving the following linear system of equations corresponding to (I – Q)J = 1, where 1 is the column vector of one's.

$$M_1 = 1 + pc.M_1 + pu.M_2 + pl.M_3,$$
$$M_2 = 1 + pc.M_1 + pu.M_3 + pl.M_4,$$
$$M_3 = 1 + pc.M_1 + pu.M_2 + pl.M_5,$$
$$M_4 = 1 + pc.M_1 + pu.M_3 + pl.M_6,$$
$$M_5 = 1 + pc.M_1 + pu.M_2 + pl.M_7,$$
$$\vdots$$
$$M_{2m-4} = 1 + pc.M_1 + pl.M_3 + ... + pu.M_{2m-2},$$
$$M_{2m-3} = 1 + pc.M_1 + pu.M_2 + ... + pl.M_{2m-1},$$
$$M_{2m-2} = 1 + pc.M_1 + pl.M_3,$$
$$M_{2m-1} = 1 + pc.M_1 + pu.M_2.$$

By solving the above linear system of equations, the ARL M_1 for a chart with m-of-m runs rule (m > 1) is given by,

$$M_1 = \frac{\left(1-pu^m\right)\left(1-pl^m\right)}{(1-pu)(1-pl)-pu.pl\left(1-pu^{m-1}\right)\left(1-pl^{m-1}\right)-pc\left(1-pu^m\right)\left(1-pl^m\right)}$$

The 1-of-1 chart signals an out-of-control status if a sign statistic falls either above upper control limit or below a lower control limit. The 2-of-2 chart signals an out-of-control status if two consecutive sign statistics fall either above an upper control limit or below lower control limit. In other words, if two successive sign statistics fall in the region 2 or region 3, the 2-of-2 chart signals an out-of-control status. The 3-of-3 chart signals an out-of-control status if three consecutive sign statistics fall either above upper control limit or below lower control limit.

Following subsection gives steady-state average time to signal of the m-of-m control chart.

5.1. *Steady-state average time to signal*

If process is running for a long period in an in-control condition, it will reach in a steady-state condition. In order to study the long term properties of a control chart, it is appropriate to investigate the steady-state average time to signal.

Let Q_0 be a square matrix obtained from Q by imposing the condition that no signal occurs. Let $\pi^T = [\pi_1,\pi_2,...,\pi_{2m-1}]$ be the vector of steady-state probabilities for the in-control transient states. The steady-state probabilities can be obtained by solving the following equations: $\pi^T Q_0 = \pi^T$ and $\pi^T 1_{2m-1} = 1$.

Under the in-control situation $p = p_0$, let $p_1 = pc$ and $u = pc = pl$.

The SSARL can be obtained by

$$SSARL = \pi^T ARL$$

and SSATS computed using Equation 12.

Steady-state performance of the m-of-m control chart is given in following subsection.

5.2. *Steady-state performance study of the m-of-m control chart*

For efficiency comparisons, we compare the proposed m-of-m chart with the synthetic, Shewhart type $\bar{X}$ and sign control charts in terms of their out-of-control steady-state ATS and zero state ATS. The results are shown in Tables 4-11 for subgroup of size 11 under normal, double exponential, Cauchy and gamma distributions.

Following are the findings from Tables 4 to 11:

Table 4. SSATS of the m-of-m and synthetic control charts for normal distribution (n=11 and SSATS(0) =1024).

$(\mu - \mu_0)$	1-of-1 chart	2-of-2 chart	3-of-3 chart	Synthetic chart
0	1024.01	1024.15	1024.00	1024.11
0.25	278.48	88.47	73.12	173.04
0.5	57.40	16.05	14.53	21.52
0.75	16.40	5.38	6.05	5.22
1	6.19	2.81	4.01	2.00
1.25	2.92	2.00	3.43	1.03
1.5	1.64	1.73	3.28	0.67
1.75	1.07	1.64	3.25	0.53
2	0.79	1.62	3.25	0.48

Table 5. Steady-state ATS of the m-of-m and synthetic control charts for Cauchy distribution. (n=11 and SSATS(0)=1024).

$(\mu - \mu_0)$	1-of-1 chart	2-of-2 chart	3-of-3 chart	Synthetic chart
0	1024.01	1024.15	1024.00	1024.11
0.25	402.90	140.07	117.07	290.00
0.5	118.47	33.63	28.43	54.00
0.75	46.41	13.10	12.21	16.66
1	23.19	7.09	7.43	7.56
1.25	13.80	4.72	5.53	4.37
1.5	9.28	3.58	4.62	2.94
1.75	6.80	2.96	4.13	2.18
2	5.29	2.58	3.84	1.73

Table 6. Steady-state ATS of the m-of-m and synthetic control charts for double exponential distribution. (n=11 and SSATS(0)=1024).

$(\mu - \mu_0)$	1-of-1 chart	2-of-2 chart	3-of-3 chart	Synthetic chart
0	1024.01	1024.15	1024.00	1024.11
0.25	115.84	32.83	27.80	52.42
0.5	22.02	6.79	7.19	7.15
0.75	7.60	3.16	4.29	2.42
1	3.66	2.18	3.55	1.25
1.25	2.17	1.83	3.34	0.82
1.5	1.47	1.70	3.27	0.63
1.75	1.11	1.64	3.25	0.54
2	0.89	1.63	3.25	0.49

Table 7. Steady-state ATS of the m-of-m and synthetic control charts for gamma distribution. (n=11 and SSATS(0)=1024).

$(\mu - \mu_0)$	1-of-1 chart	2-of-2 chart	3-of-3 chart	Synthetic chart
0	1024.00	1024.03	1024.01	1024.00
0.25	275.40	87.15	72.03	170.04
0.5	63.25	17.52	15.69	24.01
0.75	20.98	6.41	6.88	6.61
1	9.23	3.45	4.51	2.77
1.25	5.02	2.39	3.70	1.50
1.5	3.21	1.95	3.40	0.97
1.75	2.31	1.76	3.30	0.72
2	1.82	1.67	3.26	0.59

- For small to moderate shifts the SSATS and 0SATS performance of the m-of-m chart with m=2, 3 is significantly better than the Shewhart type $\overline{X}$, sign and synthetic control charts.
- Performance of sign chart under normal distribution and double exponential distribution is better as compare to m-of-m chart with m=2,3 only for a few large shifts.
- Synthetic control chart performs better than the sign chart through-out shifts; however, its performance is better as compared to m-of-m chart with m=2, 3 only for large shifts under all distributions.
- The SSATS performance of the 3-of-3 control chart is better than the 2-of-2 control chart for all distributions only for small shifts.
- The SSATS performance of all control charts under double exponential distribution is better than the gamma, Cauchy and normal distributions to monitor process median.
- It is also observed that the SSATS values and 0SATS values are not significantly differ.

5.3. *Numerical example*

We illustrate the operations of the proposed m-of-m control chart using data generated from standard normal distribution. The data set includes 21 samples each of 11 observations. We assumed that the in-control median $\mu_0 = 0$. To have an in-control ARL equal to 1024, the upper control limits of 1-of-1 chart, 2-of-2 chart and 3-of-3 chart are 11, 8 and 6 respectively. The lower control limits of these control charts set to be zero. Table 11 gives the values of the sign statistic SN_i for 21 samples. We have constructed 1-of-1 chart, 2-of-2 chart and the 3-of-3 chart in Figure 2. The 1-of-1 chart (sign chart) signals if a sign statistic falls above upper control limit of sign chart, the 2-of-2 chart signals if when two consecutive sign statistics fall above upper control limit of the 2-of-2 chart and when three consecutive sign statistics fall above upper control limit of the 3-of-3 chart, the 3-of-3 control chart signals (Table 12).

From Figure 2, we see that no points exceed the control limits of the 1-of-1 chart and 2-of-2 chart. Consequently, one might regard the process as being in a state of statistical control. From Figure 2 it is also observed that the points 6, 7 and 8 fall above the upper control limit of the 3-of-3 chart. Therefore, the 3-of-3 chart signal at point 8.

6. Conclusions

Table 8. Zero-state ATS of the Shewhart type $\overline{X}$, m-of-m and synthetic control charts for normal distribution (n=11 and SSATS(0)=1024).

$(\mu - \mu_0)$	$\overline{X}$ chart	1-of-1 chart	2-of-2 chart	3-of-3 chart	Synthetic chart
0	1024.02	1024.01	1024.22	1024.41	1024.13
0.25	146.39	278.48	88.66	73.60	160.28
0.5	19.26	57.40	16.15	14.81	17.11
0.75	4.29	16.40	5.43	6.25	3.91
1	1.47	6.19	2.85	4.17	1.62
1.25	0.75	2.92	2.03	3.58	0.94
1.5	0.55	1.64	1.76	3.43	0.66
1.75	0.51	1.07	1.67	3.40	0.55
2	0.50	0.79	1.65	3.40	0.50

Table 9. Zero-state ATS of the Shewhart type $\overline{X}$, m-of-m and synthetic control charts for Cauchy distribution (n=11 and SSATS(0)=1024).

$(\mu - \mu_0)$	$\overline{X}$ chart	1-of-1 chart	2-of-2 chart	3-of-3 chart	Synthetic chart
0	1024.05	1024.01	1024.03	1024.41	1024.13
0.25	1024.04	402.90	140.27	117.61	276.18
0.5	1024.04	118.47	33.76	28.78	46.16
0.75	1024.03	46.41	13.19	12.46	13.00
1	1024.01	23.19	7.15	7.64	5.67
1.25	1023.99	13.80	4.78	5.71	3.29
1.5	1023.97	9.28	3.63	4.79	2.27
1.75	1023.94	6.80	3.00	4.29	1.75
2	1023.90	5.29	2.62	4.00	1.44

Table 10. Zero-state ATS of the Shewhart type $\overline{X}$, m-of-m and synthetic control charts for Laplace distribution (n=11 and SSATS(0)=1024).

$(\mu - \mu_0)$	$\overline{X}$ chart	1-of-1 chart	2-of-2 chart	3-of-3 chart	Synthetic chart
0	1024.06	1024.01	1024.03	1024.41	1024.13
0.25	218.58	115.84	32.96	28.15	44.71
0.5	32.17	22.02	6.86	7.40	5.35
0.75	6.70	7.60	3.21	4.45	1.92
1	1.96	3.66	2.21	3.70	1.10
1.25	0.87	2.17	1.86	3.48	0.78
1.5	0.58	1.47	1.73	3.42	0.63
1.75	0.51	1.11	1.68	3.40	0.55
2	0.50	0.89	1.66	3.40	0.52

Table 11. Zero-state ATS of the Shewhart type $\overline{X}$, m-of-m and synthetic control charts for gamma distribution (n=11 and SSATS(0)=1024).

$(\mu - \mu_0)$	$\overline{X}$ chart	1-of-1 chart	2-of-2 chart	3-of-3 chart	Synthetic chart
0	1024.09	1024.00	1024.19	1024.00	1024.00
0.25	249.23	275.40	87.35	72.48	157.34
0.5	38.20	63.25	17.62	15.97	19.25
0.75	12.64	20.98	6.47	7.08	4.95
1	6.19	9.23	3.50	4.68	2.16
1.25	3.79	5.02	2.43	3.85	1.28
1.5	2.66	3.21	1.98	3.55	0.89
1.75	2.06	2.31	1.79	3.45	0.70
2	1.71	1.82	1.70	3.41	0.60

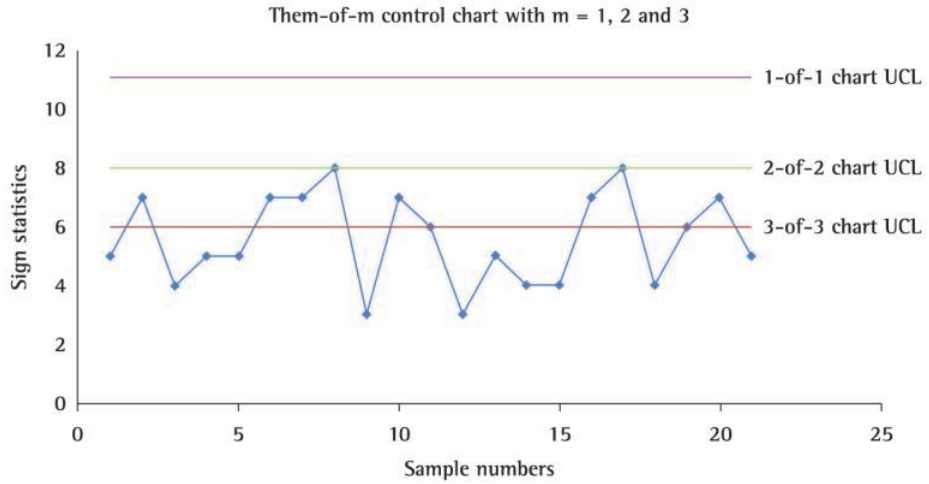


Figure 2. The m-of-m control chart with m = 1, 2, 3.

Table 12. Sample numbers and values of sign statistic.

Sample No.	Sign statistic SN_i
1	5
2	7
3	4
4	5
5	5
6	7
7	7
8	8
9	3
10	7
11	6
12	3
13	5
14	4
15	4
16	7
17	8
18	4
19	6
20	7
21	5

We have investigated the steady-state ATS of a nonparametric synthetic and the m-of-m control charts based on sign statistic. The proposed charts are used to monitor shifts in a process median. The SSATS values of proposed charts are computed by employing Markov chain approach. The steady-state performance of the m-of-m chart with m=2, 3 is significantly better than the sign chart (1-of-1 chart) and the synthetic control chart. Also, the steady-state ATS performance of the synthetic control chart is poor as compared to the zero-state ATS. The m-of-m control chart with m=2, 3 has a higher power of detecting out-of-control signal than the sign chart and the synthetic control chart.

References

Amin, R. W., & Searcy, A. J. (1991). A nonparametric exponentially weighted moving average control schemes. *Communications in Statistics: Simulation and Computation*, *20*(4), 1049-1072. http://dx.doi.org/10.1080/03610919108812996

Amin, R. W., Reynolds Junior, M. R., & Bakir, S. T. (1995). Nonparametric quality control charts based on the sign statistic. *Communications in Statistic: Theory and Methods*, *24*(6), 1597-1623. http://dx.doi.org/10.1080/03610929508831574

Bakir, S. T. (2004). A distribution-free shewhart quality control chart based on signed-ranks. *Quality Engineering, 16*(4), 613-623. http://dx.doi.org/10.1081/QEN-120038022

Bakir, S. T. (2006). Distribution-free quality control charts based on signed-rank-like statistic. *Communications in Statistics: Theory and Methods*, *35*(4), 743-757. http://dx.doi.org/10.1080/03610920500498907

Bakir, S. T., & Reynolds Junior, M. R. (1979). A nonparametric procedure for process control based on within-group ranking. *Technometrics*, *21*(2), 175-183. http://dx.doi.org/10.1080/00401706.1979.10489747

Bourke, P. D. (1991). Detecting a shift in fraction nonconforming using run length control chart with 100% inspection. *Journal of Quality Technology*, 3(2), 51-68.

Chakraborti, S., & Eryilmaz, S. (2007). A nonparametric shewhart-type signed-rank control chart based on runs. *Communications in Statistic*, *36*(2), 335-356. http://dx.doi.org/10.1080/03610910601158427

Chakraborti, S., & Van de Wiel, M. A. (2008). A nonparametric control charts based on mann-whitney statistic. *IMS Collection*, *1*, 156-172. http://dx.doi.org/10.1214/193940307000000112

Champ, W. C. (1992). Steady-state run length analysis of a shewhart control chart with supplementary

runs rules. *Communications in Statistics: Theory and Methods*, *21*(3), 765-777. http://dx.doi.org/10.1080/03610929208830813

Crosier, R. B. (1986). A new two-sided cumulative sum quality control scheme. *Technometrics*, *28*(3), 187-194. http://dx.doi.org/10.1080/00401706.1986.10488126

Davis, R. B., & Woodall, W. H. (2002). Evaluating and improving the synthetic control chart. *Journal of Quality Technology*, *34*(2), 200-208.

Ho, L. L., & Costa, A. F. B. (2011). Monitoring a wandering mean with an np chart. *Produção*, *21*(2), 254-258. http://dx.doi.org/10.1590/S0103-65132011005000027

Human, S. W., Chakraborti, S., & Smit, C. F. (2010). Nonparametric shewhart-type sign control charts based on runs. *Communications in Statistics: Theory and Methods*, *39*(11), 2046-2062. http://dx.doi.org/10.1080/03610920902969018

Khilare, S. K., & Shirke, D. T. (2010). A nonparametric synthetic control chart using sign statistic. *Communications in Statistics: Theory and Methods*, *39*(18), 3282-3293. http://dx.doi.org/10.1080/03610920903249576

Lim, T., & Cho, M. (2009). Design of control charts with m-of-m runs rules. *Quality and Reliability Engineering International*, *25*(8), 1085-1101. http://dx.doi.org/10.1002/qre.1023

McGilchrist, C. A, & Woodyer, K. D. (1975). Note on a distribution-free CUSUM technique. *Technometrics*, *17*(3), 321-325. http://dx.doi.org/10.1080/00401706.1975.10489335

Saccucci, M. S., & Lucas, J. M. (1990). Average run length for exponentially weighted moving average control schemes using the markov chain approach. *Journal of Quality Technology*, *22*(2), 154-162.

Wu, Z., & Spedding, T. A. (2000). A synthetic control chart for detecting small shifts in the process mean. *Journal of Quality Technology*, 32, 32-38.

Wu, Z., Yeo, S. H., & Spedding, T. A. (2001). A synthetic control chart for detecting fraction nonconforming increases. *Journal of Quality Technology*, 33(1), 104-111.

Stoch Environ Res Risk Assess
DOI 10.1007/s00477-015-1022-8

ORIGINAL PAPER

Estimation of confidence interval for hydrological design value for some continuous distributions under complete and censored samples

S. K. Powar · H. V. Kulkarni

Abstract The quantile of a probability distribution, known as return period or hydrological design value of a hydrological variable is the value corresponding to fixed non-exceedence probability and is very important notion in hydrology. In hydraulic engineering design and water resources management, confidence interval (CI) estimation for a population quantile is of primary interest and among other applications, is used to assess the pollution level of a contaminant in water, air etc. The accuracy on such estimation directly influences the engineering investments and safety. The two parameter Weibull, Pareto, Lognormal, Inverse Gaussian, Gamma are some commonly used probability models in such applications. In spite of its practical importance, the problem of CI estimation of a quantile of these widely applicable distributions has been less attended in the literature. In this paper, a new method is proposed to obtain a CI for a quantile of any distribution for which [or the probability distribution of any one-to-one function of the underlying random variable (RV)] generalized pivotal quantities (GPQs) exist for its parameters. The proposed method is elucidated by constructing CIs for quantiles of Weibull, Pareto, Lognormal, Extreme value distribution of type-I for minimum, Exponential and Normal distributions for complete as well as type II singly right censored samples. The empirical performance evaluation of the proposed method evinced that the proposed method has exact well concentrated coverage probabilities near the nominal level even for small uncensored samples as small as 5 and for censored samples as long as the proportion of censored observations is up to 0.70. The existing methods for Weibull distribution have poor or dispersed coverage probabilities with respect to the nominal level for complete samples. Applications of the proposed method in ground water monitoring and in the assessment of air pollution are illustrated for practitioners.

Keywords Quantile · Confidence interval · Generalized variable approach · Hydrology · Censoring

1 Introduction

The field of applications of continuous random variables (RVs) like two parameter Weibull, Pareto, Gamma, Exponential, Inverse Gaussian and Lognormal is vast and encompasses nearly all scientific disciplines. Using these distributions, data have been modeled which originate from diverse areas like the biological, environmental, health, physical and social sciences. These distributions are also widely used in meteorology and hydrology disciplines and have become standards in reliability theory for modeling time-dependent failure data. Some pertinent notable references are, Grace and Eagleson (1966), Nathan and Mcmahon (1990), Selker and Haith (1990), Power (1992), Jiang et al. (1997), Duan et al. (1998), Jandhyala et al. (1999), Seshadri (1999), Aksoy (2000), Lun and Lam (2000), Seguro and Lambert (2000), Talkner and Weber (2000), Clarke (2002), Heo et al. (2001), Tan et al. (2007), Yang et al. (2007), Krishnamoorthy and Lin (2010) and Jamdade and Jamdade (2012) among many others. The estimation problem for the confidence interval (CI) of a quantile also known as a tolerance interval (TI) for any of these distributions is a consequential problem in hydraulic

S. K. Powar (✉) · H. V. Kulkarni
Department of Statistics, Shivaji University, Kolhapur 416004, India
e-mail: sarjerao.powar@rediffmail.com

H. V. Kulkarni
e-mail: kulkarni.hemangi@gmail.com

Published online: 14 January 2015

A Simultaneous Evaluation of Adaptive Design Parameters Policies for Hotelling's T^2 Charts

Shashibhushan B. Mahadik
Department of Statistics
Shivaji University, Kolhapur, INDIA 416004

Abstract

A *complectly adaptive* (CA) Hotelling's T^2 chart, that is a T^2 chart in which all the design parameters, viz, sampling interval, sample size, control limit, and warning limit are adaptive, each taking two values, is developed. The expressions for the statistical and operational performance measures for this chart are derived using a Markov chain approach. As any adaptive T^2 chart in which one or more of the design parameters are adaptive, each taking two values, is a particular case of the CA T^2 chart, the derived expressions are directly applicable to all such charts.These expressions can be used tocompute the performancemeasures for all such charts and thus to determine the most suitable adaptive T^2 chart for a given situation.

Key words:Average number of samples to signal, average number of observations to signal, average number of switches to signal, Multivariate Statistical process control, Steady-state average time to signal.

I. INTRODUCTION

Hotelling's T^2 chart is an effective on-line process control technique used to monitor simulataneouslytwo or more quality characteristics of a process. If all the design parameters of this chart are kept fixed throughout the period of monoitoring, it is called static T^2 chart while if at least one design parameter is variableand takes a value for a trial according to the location(s) of the sample points corresponding to the previous trial(s), the chart is called adaptive T^2 chart. The general principal of choosing values of the adaptive parameters for atrial is as follows. If the last plotted point(s) indicate possibility of a shift, choose short sampling interval and/or large sample size and/or narrow in-control limits for the next trial. On the other hand, if that indicate possibility of safe or in-control region, choose long sampling interval and/or small sample size and/or wide in-control region for the next trial. It has been shown that adapting one or more design parameters of a T^2 chart increases its statistical, operational, and economic performances significantly. See, for example,Aparasi[1], Aparasi andHaro[2, 3],Faraz and Moghadam[4],Mahadik and Shirke[5], Mahadik[6-10].

Recently, Mahadik[11]has developed a*complectly adaptive* (CA) $\bar{X}$ chart, that is an $\bar{X}$ chartin which all the design parameters, viz, sampling interval, sample size, control limits, and warning limits are adaptive, each taking two values. This idea has been extended for T^2 chart in this paper.

The following sections present the general description of a CA T^2 chart,derivations of the expressions for statistical and operationalperformance measures for this chart, numerical comparisons of the performances of various adaptive T^2 charts that are the particular cases of the CA T^2 chart,andconclusions.

II. A CA T^2 CHART

Suppose the p> 1 related quality characteristics $\boldsymbol{X} = (X_1, X_2, ..., X_p)'$ to be monitored together, follow p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and known variance covariance matrix Σ. Let $\boldsymbol{\mu_0}$ be

ISSN: 2278 – 7798

the target mean vector. An occurrence of an assignable cause shifts $\boldsymbol{\mu}$ from $\boldsymbol{\mu_0}$ to $\boldsymbol{\mu_1} \neq \boldsymbol{\mu_0}$. A CA T^2 chart to monitor $\boldsymbol{\mu}$ is as described below.

The control statistic is $T_i^2 = n(i)(\overline{\boldsymbol{X}}_i - \boldsymbol{\mu_0})'\Sigma^{-1}(\overline{\boldsymbol{X}}_i - \boldsymbol{\mu_0})$, where $\overline{\boldsymbol{X}}_i$, i = 1, 2, …, is the mean vector of the i^{th} sample of size $n(i)$ drawn on $\boldsymbol{X}$. Note that when $\boldsymbol{\mu} = \boldsymbol{\mu_0}$, T_i^2 follows central chi-square distribution with p degrees of freedom, and when $\boldsymbol{\mu} = \boldsymbol{\mu_1}$, for given $n(i) = n$, it follows non-central chi-square distribution with p degrees of freedom and non-centrality parameter $n(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})'\Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}) = nd^2$, where $d = \sqrt{(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})'\Sigma^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})}$ is the Mahalanobis distance used to measure a change in the process mean vector. Let $t(i)$ be the length of sampling interval between the $(i-1)^{\text{st}}$ and i^{th}trials, i = 1, 2, …. Also, let $L(i)$ and $w(i)$ be the control and warning limits,respectively,for the i^{th} trial.

The values of $(t(i), n(i), L(i), w(i))$ can be either (t_1, n_1, L_1, w_1) or (t_2, n_2, L_2, w_2), where t_1, t_2, n_1, n_2, L_1, L_2, w_1, and w_2 are such that $t_{\max} \geq t_1 \geq t_2 \geq t_{\min}, n_{\min} \leq n_1 \leq n_2 \leq n_{\max}, \infty > L_1 \geq L_2 > 0, 0 < w_1 < L_1, 0 < w_2 < L_2$, and $w_1 \geq w_2$.Here, $t_{\max}$ and $t_{\min}$ being the longest and shortest possible sampling interval lengths, respectively, while, $n_{\min}$ and $n_{\max}$ being the smallest and largest possible sample sizes, respectively,

When T_{i-1}^2 falls below$L(i-1)$, the values of $(t(i), n(i), L(i), w(i))$, i = 2, 3, …, between (t_1, n_1, L_1, w_1) and (t_2, n_2, L_2, w_2) are chosen according to the following rule.

$$(t(i), n(i), L(i), w(i)) = \begin{cases} (t_1, n_1, L_1, w_1), & \text{if } T_{i-1}^2 \in I_1(i-1) \\ (t_2, n_2, L_2, w_2), & \text{if } T_{i-1}^2 \in I_2(i-1), \end{cases}$$

where

$$I_1(i-1) = [0, w(i-1)]$$

$$\text{and } I_2(i-1) = (w(i-1), L(i-1)).$$

The chart signals an out-of-control stateat the i^{th} trial, i = 1, 2, …, if T_i^2 falls above $L(i)$.

The values of $(t(1), n(1), L(1), w(1))$ can be chosen using an arbitrary probability distribution. In practice, it is recommended to choose the quadruplet(t_2, n_2, L_2, w_2) for that to provide additional protection against the problems that may exist initially. The trial following an out-of-control signal is again treated to be the first trial.

In the next section, expressions for performance measures for a CA T^2 chart are derived.

III. PERFORMANCE MEASURES

The measures of statistical performance of a CA T^2 chart are *steady-state average time to signal* (SSATS),*average number of samples to signal* (ANSS), and *average number of observations to signal* (ANOS). SSATS is the expected value of the time between a shift that occurs at some random time after the process starts and the time the chart signals. ANSS and ANOS are the expected values of the number of samples and the number of observations, respectively taken from the time of a shift to the time the chart signals.Themeasure of operational performance is *average number of switches to signal* (ANSW), which is the expected value of the number of switches between the quadruplets of values of sampling interval length, sample size, control limit, and warning limit from a shift to the signal.

Let SSATS_d, ANSS_d, ANOS_d, and ANSW_dbe the SSATS, ANSS, ANOS, and ANSW, respectively of a T^2 chart when the process mean vector has shifted from $\boldsymbol{\mu_0}$ to $\boldsymbol{\mu_1}$in d units. In the following,the expressions for SSATS_d, ANSS_d, and $\text{ANOS}_d$are derived using a Markov chain approach.

Henceforth,thei^{th} trial refers to the i^{th} trial after a shift when $i > 0$ and the last trial before the shift when $i = 0$.Also, T_i^2 refers to the sample point corresponding to the i^{th} trial.

ISSN: 2278 – 7798

Define the three states 1, 2, and 3 of a Markov chain corresponding to whether the sample pointcorresponding to the i^{th} trial is plotted in $I_1(i)$, $I_2(i)$, and $I_3(i)$ = [$L(i)$, ∞), respectively,i = 1, 2, ….. Note that state 3 is the absorbing state, as the control charting process is restarted when a sample point falls in region $I_3(i)$. The transition probability matrix is given by

$$\mathbf{P^d} = \begin{bmatrix} p_{11}^d & p_{12}^d & p_{13}^d \\ p_{21}^d & p_{22}^d & p_{23}^d \\ 0 & 0 & 1 \end{bmatrix},$$

where p_{jk}^d is the transition probability that j is the prior state and k is the current state, when the process mean vector has shifted by d units. For example,

$$\begin{aligned} p_{12}^d &= \Pr_d[T_i^2 \in I_2(i) \mid T_{i-1}^2 \in I_1(i-1)] \\ &= \Pr_d[T_i^2 \in I_2(i) \mid n(i) = n_1,\ L(i) = L_1,\ w(i) = w_1] \\ &= \mathrm{F}_{\lambda_1}(L_1) - \mathrm{F}_{\lambda_1}(w_1) \end{aligned}$$

where $\mathrm{F}_{\lambda_1}(\cdot)$ is the cumulative distribution function of non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\lambda_1 = n_1 d^2$.

Then, $\mathrm{SSATS_d}$ and $\mathrm{ANSS_d}$ are given by

$$\mathrm{SSATS_d} = \boldsymbol{b}'(\mathbf{I} - \mathbf{P_1^d})^{-1}\boldsymbol{t} - \mathrm{E}(U), \quad (1)$$

$$\mathrm{ANSS_d} = \boldsymbol{b}'(\mathbf{I} - \mathbf{P_1^d})^{-1}\mathbf{1},$$

and

$$\mathrm{ANOS_d} = \boldsymbol{b}'(\mathbf{I} - \mathbf{P_1^d})^{-1}\boldsymbol{n},$$

where $\mathbf{I}$ is the identity matrix of order 2, $\mathbf{P_1^d}$ is the submatrix of $\mathbf{P^d}$ that contains the probabilities associated with the transient states only, $\boldsymbol{t}' = (t_1, t_2)$, $\mathbf{1}' = (1, 1)$, $\boldsymbol{n}' = (n_1, n_2)$, and $\boldsymbol{b}' = (b_1, b_2)$, b_j being the conditional probability that T_0^2 falls in $I_j(0)$ given that it falls below $L(0)$, j = 1, 2. We note that $b_2 = 1 - b_1$. The Expression for b_1 is derived by Mahadik[7] and is as given below.

$$b_1 = \frac{\dfrac{\mathrm{F}_0(w_2)}{\mathrm{F}_0(L_2)}}{1 - \dfrac{\mathrm{F}_0(w_1)}{\mathrm{F}_0(L_1)} + \dfrac{\mathrm{F}_0(w_2)}{\mathrm{F}_0(L_2)}},$$

where $\mathrm{F}_0(.)$ is the cumulative distribution function of central chi-square distribution with p degrees of freedom.

E(U) in equation (1) is the expected value of the time U between the 0^{th} trial and the shift. Assuming that an assignable cause of a process shift occurs according to a Poisson process, it can be shown that E(U) = $\mathrm{E}[t(1)]/2 = \boldsymbol{b}'\boldsymbol{t}/2$. Hence,

$$\mathrm{SSATS_d} = \boldsymbol{b}'(\mathbf{I} - \mathbf{P_1^d})^{-1}\boldsymbol{t} - \boldsymbol{b}'\boldsymbol{t}/2.$$

Now, to derive the expression for $\mathrm{ANSW_d}$, let

$$Y_i = \begin{cases} 1, \text{ if } (T_{i-1}^2 \in I_1(i-1), T_i^2 \in I_2(i)) \\ 2, \text{ if } (T_{i-1}^2 \in I_2(i-1), T_i^2 \in I_1(i)) \\ 3, \text{ if } (T_{i-1}^2 \in I_1(i-1), T_i^2 \in I_1(i)) \\ 4, \text{ if } (T_{i-1}^2 \in I_2(i-1), T_i^2 \in I_2(i)) \\ 5, \text{ if } T_i^2 > L(i) \end{cases},$$

i = 1, 2, …

Note that { Y_i, i = 1, 2, … } is a Markov chain with transition probability matrix

$$\mathbf{Q^d} = \begin{bmatrix} 0 & p_{21}^d & 0 & p_{22}^d & p_{23}^d \\ p_{12}^d & 0 & p_{11}^d & 0 & p_{13}^d \\ p_{12}^d & 0 & p_{11}^d & 0 & p_{13}^d \\ 0 & p_{21}^d & 0 & p_{22}^d & p_{23}^d \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then, $\mathrm{ANSW_d}$ is given by

$$\mathrm{ANSW_d} = \boldsymbol{a}'(\mathbf{I_1} - \mathbf{Q_1^d})^{-1}\boldsymbol{e},$$

where, $\mathbf{I_1}$ is the identity matrix of order 4, $\mathbf{Q_1^d}$ is the submatrix of $\mathbf{Q^d}$ that contains the probabilities associated with the transient states only, $\boldsymbol{e} = (1,1,0,0)'$,and $\boldsymbol{a} = (a_1, a_2, a_3, a_4)'$, a_j being the initial probability of state j, j = 1, 2, 3, 4, given by

ISSN: 2278 – 7798

$$a_j = \Pr_d[Y_1 = j] = \begin{cases} b_1 p_{12}^d \ , & j = 1 \\ b_2 p_{21}^d \ , & j = 2 \\ b_1 p_{11}^d \ , & j = 3 \\ b_2 p_{22}^d \ , & j = 4 \end{cases}.$$

In the next section, the above derieved expressions are used to compute the performance measures for a CA T^2 chart and for all the adaptive T^2 charts which are its particular cases.

IV. PERFORMANCE COMPARISON OF THE ADAPTIVE T^2 CHARTS

In this section, we simultaneously evaluate a CA T^2 chart and its all particular cases through numerical comparisons of their statistical and operational performances. For this, we have to design all these chart such that their in-control statistical performances match. Below is described the procedure of designing a CA T^2 chart whose in-control statistical performances match to that of a given static T^2 chart. Note that this procedure is also applicable to design all the charts, which are particular cases of a CA T^2 chart,such that their in-control statistical performances match to that of the given static T^2 chart.

Let t_0, n_0, and L_0 be the sampling interval length, sample size, and control limit of a static T^2 chart. Let $SSATS_0(static)$, $ANSS_0(static)$, and $ANOS_0(static)$ be the in-control SSATS, ANSS, and ANOS, respectively of this chart. Then, we have

$$SSATS_0(static) = t_0\left(\frac{1}{1-F_0(-L_0)} - \frac{1}{2}\right)$$

$$ANSS_0(static) = \frac{1}{1-F_0(-L_0)}$$

and

$$ANOS_0(static) = \frac{n_0}{1-F_0(-L_0)}$$

Now, given t_0, n_0,and L_0, we have to choose the design parameters of a CA T^2 chart satisfying the following requirements.

$$SSATS_0(CA) = SSATS_0(static)$$

$$ANSS_0(CA) = ANSS_0(static)$$

and

$$ANOS_0(CA) = ANOS_0(static)$$

Here,$SSATS_0(CA)$, $ANSS_0(CA)$, and $ANOS_0(CA)$ are the in-control SSATS, ANSS, and ANOS, respectively of a CA T^2 chart.

Fixing any five among the design parameters (t_1, t_2, n_1, n_2, L_1, L_2, w_1, and w_2) of a CA T^2 chart, the above nonlinear equations can be solved for the remaining three parameters. This can be done, for example, using package rootSolve in R or using function 'fsolve' in Matlab.

The complete set of adaptive T^2 charts containing a CA T^2 chart and its all particular cases includes:

1. A *variable sampling interval* (VSI) T^2 chart
2. A *variable sample size* (VSS) T^2 chart
3. A *variable control limits* (VCL) T^2 chart
4. A *variable sample size and sampling interval* (VSSI) T^2 chart
5. A *variable sampling interval and control limits* (VSICL) T^2 chart
6. A *variable sampling interval and warning limits* (VSIWL) T^2 chart
7. A *variable sample size and control limits* (VSSCL) T^2 chart
8. A *variable sample size and warning limits* (VSSWL) T^2 chart
9. A *variable control limits and warning limits* (VCWL) T^2 chart
10. A *variable sample size, sampling interval, and control limits* (VSSICL) T^2 chart
11. A *variable sample size, sampling interval, and warning limits* VSSIWL) T^2 chart

ISSN: 2278 – 7798

12. A *variable sampling interval, control limits, and warning limits* (VSICWL) T^2 chart
13. A *variable sample size, control limits, and warning limits* (VSSCWL) T^2 chart
14. A CA T^2 chart

Fixing the values of t_0, n_0, and L_0 and applying the procedure described above, all the charts in above set can be designed such that their in-control statistical performances match. Table 1 shows the design parameters of one suchset of matched charts while tables 2, 3, 4, and 5, respectively show the $SSATS_d$, $ANSS_d$,$ANOS_d$,and $ANSW_d$ performancesfor these charts for various values of*d*.Such tables are useful to determine the most suitable adaptive T^2 chart for a given situation. In general, one can see from tables 2 to 5 that CA T^2 chart is the best choice if one is interested in detecting only small shifts while VSI or VSIWL T^2 chartsare the best choices if the interest is in detecting only moderate to large shifts. However, in practice, one can choose the most suitable charts taking into consideration the practical constraints in deciding which of the design parameters of the charts can be adaptive.

V. CONCLUSIONS

The expressions for the statistical and operational performance measures for a CA T^2 chart are developed. These expressions are are directly applicable to any adaptive T^2 chart in which any of the design parameters are adaptive, each taking two values. The simultaneous numerical comparisons of the performances of all such charts indicate that in general, CA T^2 chart is the best chart for detecting small shifts while VSI or VSIWL T^2 charts are the best charts for detecting moderate to large shifts. In practice, such simultaneous comparisons guide to determine the most suitable T^2 chart satisfying the practical constraints in deciding which of the design parameters of the chart can be adaptive.

Table 1: Design parameters of the matched T^2 charts

Chart	Design Parameters							
	n_1	n_2	t_1	t_2	w_1	w_2	L_1	L_2
Static	5	5	1.00	1.00	14.86	14.86	0.00	0.00
VSI	5	5	1.79	0.20	14.86	14.86	3.36	3.36
VSS	2	10	1.00	1.00	14.86	14.86	4.21	4.21
VCL	5	5	1.00	1.00	16.42	13.93	3.36	3.36
VSSI	2	10	1.48	0.20	14.86	14.86	4.21	4.21
VSICL	5	5	1.79	0.20	16.42	13.93	3.36	3.36
VSIWL	5	5	1.79	0.20	14.86	14.86	4.03	2.75
VSSCL	2	10	1.00	1.00	17.35	13.15	4.21	4.21
VSSWL	2	10	1.00	1.00	14.86	14.86	4.89	3.30
VCWL	5	5	1.00	1.00	16.42	13.93	4.04	2.75
VSSICL	2	10	1.48	0.20	16.42	13.48	4.21	4.21
VSSIWL	2	10	1.48	0.20	14.86	14.86	4.66	3.57
VSICWL	5	5	1.56	0.20	16.42	13.63	4.04	3.78
VSSCWL	2	10	1.00	1.00	16.42	13.48	4.76	3.45
CA	2	10	1.48	0.20	16.42	13.48	4.88	3.30

ISSN: 2278 – 7798

Table 2: SSATS$_d$ values for the matched T^2 charts

Chart	d							
	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
Static	126.73	48.69	17.57	6.95	1.63	0.69	0.52	0.50
VSI	118.98	38.10	10.52	3.21	0.78	0.54	0.50	0.50
VSS	120.47	33.12	8.26	3.11	1.31	0.99	0.80	0.64
VCL	123.45	44.86	15.53	6.12	1.56	0.70	0.52	0.50
VSSI	114.09	26.14	5.01	1.76	0.87	0.67	0.58	0.53
VSICL	115.93	35.15	9.37	2.89	0.77	0.54	0.50	0.50
VSIWL	117.42	36.23	9.61	2.93	0.77	0.54	0.50	0.50
VSSCL	96.98	23.09	6.24	2.66	1.31	1.05	0.88	0.71
VSSWL	119.78	31.97	8.07	3.18	1.36	1.00	0.80	0.64
VCWL	122.58	44.08	15.24	6.05	1.56	0.70	0.52	0.50
VSSICL	96.27	19.74	4.17	1.65	0.87	0.67	0.59	0.54
VSSIWL	112.70	24.66	4.75	1.83	0.93	0.69	0.58	0.53
VSICWL	115.80	35.12	9.42	2.93	0.77	0.54	0.50	0.50
VSSCWL	100.34	23.87	6.45	2.80	1.35	1.05	0.85	0.68
CA	93.85	18.15	3.99	1.78	0.96	0.71	0.60	0.54

Table 3: ANSS$_d$ values for the matched T^2 charts

Chart	d								
	0	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
Static	200	127.23	49.19	18.07	7.45	2.13	1.19	1.02	1.00
VSI	200	127.23	49.19	18.07	7.45	2.13	1.19	1.02	1.00
VSS	200	120.97	33.62	8.76	3.61	1.81	1.49	1.30	1.14
VCL	200	123.95	45.36	16.03	6.62	2.06	1.20	1.02	1.00
VSSI	200	120.97	33.62	8.76	3.61	1.81	1.49	1.30	1.14
VSICL	200	123.95	45.36	16.03	6.62	2.06	1.20	1.02	1.00
VSIWL	200	127.23	49.19	18.07	7.45	2.13	1.19	1.02	1.00
VSSCL	200	97.48	23.59	6.74	3.16	1.81	1.55	1.38	1.21
VSSWL	200	120.28	32.47	8.57	3.68	1.86	1.50	1.30	1.14
VCWL	200	123.08	44.58	15.74	6.55	2.06	1.20	1.02	1.00
VSSICL	200	102.01	25.22	7.06	3.23	1.81	1.53	1.35	1.18
VSSIWL	200	120.50	32.81	8.61	3.65	1.84	1.50	1.30	1.14
VSICWL	200	123.12	44.42	15.54	6.42	2.04	1.21	1.02	1.00
VSSCWL	200	100.84	24.37	6.95	3.30	1.85	1.55	1.35	1.18
CA	200	100.61	24.22	6.94	3.32	1.86	1.55	1.35	1.18

ISSN: 2278 – 7798

Table 4: $ANOS_d$ values for the matched T^2 charts

Chart	d								
	0	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
Static	1000	636.15	245.97	90.35	37.25	10.67	5.97	5.10	5.00
VSI	1000	636.15	245.97	90.35	37.25	10.67	5.97	5.10	5.00
VSS	1000	644.70	211.70	64.20	26.50	11.80	9.40	7.90	6.40
VCL	1000	619.77	226.82	80.16	33.12	10.30	6.02	5.12	5.01
VSSI	1000	644.70	211.70	64.20	26.50	11.80	9.40	7.90	6.40
VSICL	1000	636.15	226.80	80.20	33.10	10.30	6.00	5.10	5.00
VSIWL	1000	636.20	246.00	90.40	37.30	10.70	6.00	5.10	5.00
VSSCL	1000	518.50	146.70	47.60	22.20	11.80	10.10	8.70	7.10
VSSWL	1000	649.60	211.87	63.96	26.56	11.82	9.43	7.83	6.37
VCWL	1000	615.40	222.90	78.70	32.80	10.30	6.00	5.10	5.00
VSSICL	1000	542.80	157.30	50.20	22.90	11.70	9.90	8.40	6.80
VSSIWL	1000	648.07	211.82	64.00	26.52	11.81	9.43	7.84	6.38
VSICWL	1000	615.60	222.10	77.70	32.10	10.20	6.00	5.10	5.00
VSSCWL	1000	542.27	156.11	50.06	22.99	11.79	9.89	8.37	6.79
CA	1000	542.16	155.91	50.05	23.03	11.81	9.89	8.37	6.79

Table 5: $ANSW_d$ values for the matched T^2 charts

Chart	d								
	0	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
Static	0	0	0	0	0	0	0	0	0
VSI	99.50	62.79	22.36	6.41	1.77	0.30	0.08	0.01	0.00
VSS	93.28	54.72	13.01	2.36	0.79	0.54	0.43	0.28	0.14
VCL	99.75	61.39	20.77	5.80	1.66	0.34	0.11	0.02	0.00
VSSI	93.28	54.72	13.01	2.36	0.79	0.54	0.43	0.28	0.14
VSICL	99.75	61.39	20.77	5.80	1.66	0.34	0.11	0.02	0.00
VSIWL	79.75	50.50	17.86	4.97	1.36	0.28	0.08	0.01	0.00
VSSCL	93.70	44.30	9.22	1.91	0.78	0.58	0.50	0.37	0.21
VSSWL	73.54	42.56	9.46	1.69	0.69	0.53	0.43	0.28	0.14
VCWL	63.50	40.86	14.51	3.80	0.96	0.20	0.05	0.01	0.00
VSSICL	93.59	46.30	9.83	1.98	0.78	0.57	0.48	0.34	0.18
VSSIWL	79.80	46.37	10.52	1.88	0.72	0.53	0.43	0.28	0.14
VSICWL	93.22	58.61	20.88	6.21	1.89	0.40	0.13	0.02	0.00
VSSCWL	77.39	37.60	7.58	1.54	0.71	0.56	0.47	0.33	0.18
CA	74.06	35.86	7.16	1.47	0.69	0.56	0.47	0.33	0.18

ISSN: 2278 – 7798

REFERENCES

[1] F. Aparasi, "Hotelling's T^2 control chart with adaptive sample sizes', *International journal of Production Research*, vol. 34, pp. 2853–2862, 1996.

[2] F. Aparasi and C. Haro, "Hotelling's T^2 control chart with variable sampling intervals", *International journal of Production Research*, vol. 39, pp. 3127–3140, 2001.

[3] F. Aparasi and C. Haro, "A comparison of T^2 control charts with variable sampling schemes as opposed to MEWMA chart," *International Journal of Production Research*, vol. 41, pp. 2169-2182, 2003.

[4] A.Faraz and M. B.Moghadam, "Hotelling's T^2 control chart with two adaptive sample sizes," *Quality & Quantity*, vol. 43, pp. 903-912, 2009.

[5] S. B. Mahadik and D. T. Shirke, "A special variable sample size and sampling interval hotelling's T^2 chart," *The International Journal of Advanced Manufacturing Technology*, vol. 53, pp. 379-384, 2011.

[6] S. B. Mahadik, "Variable sampling interval Hotelling's T^2 charts with runs rules for switching between sampling interval lengths," *Quality and Reliability Engineering International*,vol. 28, pp. 131-140, 2012.

[7] S. B. Mahadik, "Hotelling's T^2 charts with variable control and warning limits," *International Journal of Quality Engineering and Technology*, vol. 3, pp. 158-167, 2012.

[8] S. B. Mahadik, "Variable sample size and sampling interval Hotelling's T^2 charts with runs rules for switching between sample sizes and sampling interval lengths," *International Journal of Reliability, quality, and Safety Engineering*, Vol. 20, 2013.

[9] S. B. Mahadik, "Hotelling's T^2 charts with variable sampling interval and warning limits", *International Journal of Quality Engineering and Technology*, Vol 3, No. 4, pp. 289-302, 2013.

[10] S. B. Mahadik, "Hotelling's T^2 charts with variable sample size, sampling interval, and warning limits", *International Journal of Science, Engineering and Technology Research*, Vol. 3, No. 1, pp. 41-54, 2014.

[11] S. B. Mahadik, "A Unified Approach to Adaptive Shewhart Control Charts", under review.

ISSN: 2278 – 7798

Shashibhushan B. Mahadikis an Assistant Professor in the Department of Statistics, Shivaji University, Kolhapur, India. Dr. Mahadik received his M. Sc., M. Phil., and Ph. D. degrees in Statistics from Shivaji University. He was Assistant Professor in D R K College of Commerce, Kolhapur and in the Department of Statistics, Solapur University, Solapur, India before joining Shivaji University. His current research interests include statistical quality control and nonparametric statistics.

ISSN: 2278 – 7798

Environ Ecol Stat
DOI 10.1007/s10651-014-0286-3

Prediction intervals for environmental events based on Weibull distribution

Ashis SenGupta · Hemangi V. Kulkarni · Uttam D. Hubale

Received: 11 July 2013 / Revised: 15 March 2014

Abstract Prediction of the occurrences of natural environmental hazards like earthquakes, rainfall etc. as well as of renewable energy like wind and solar energy are attracting increased attention of researchers. The underlying distributions are usually skewed, with gamma, lognormal and Weibull being popular distributions for modeling such phenomena. In particular the Weibull distribution has been used to model a variety of situations. This article addresses predictive inferences for future observations from a Weibull distribution based on a 'Pivot quantity'. The performance of the proposed prediction interval (PI) is empirically assessed in comparison to existing methods on the basis of coverage probabilities and expected lengths. The proposed PI has excellent performance even for small sample sizes, particularly when the shape parameter is very small. Applications of the PI to environmental phenomena are illustrated for predicting an interval for a future earthquake, for future minimum annual rainfall in next m years and for future wind energy that can be generated through wind turbines. The concept of residual PI is introduced, its theoretical closed form expression is obtained and is illustrated for the earthquake data. The predictive inference can give substantial feedback in future planning strategies in these areas.

Handling Editor: Ashis SenGupta.

A. SenGupta
Applied Statistics Unit, Indian Statistical Institute, Kolkata, India
e-mail: amsseng@gmail.com

A. SenGupta
Department of Statistics, University of California, Riverside, USA

H. V. Kulkarni (✉) · U. D. Hubale
Department of Statistics, Shivaji University, Kolhapur, India
e-mail: kulkarnih1@rediffmail.com

U. D. Hubale
e-mail: uttamhubale@gmail.com

Published online: 09 May 2014

Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v9n2p417

Nonparametric tests for testing equality of location parameters of two multivariate distributions
By Chavan, Shirke

Published: 14 October 2016

Electronic Journal of Applied Statistical Analysis
Vol. 09, Issue 02, October 2016, 417-432
DOI: 10.1285/i20705948v9n2p417

Nonparametric tests for testing equality of location parameters of two multivariate distributions

Chavan A.R.*and Shirke D. T.

Department of Statistics, Shivaji University, Kolhapur, India - 416004

Published: 14 October 2016

In this paper, we have proposed two nonparametric tests for testing the equality of location parameters of two multivariate distributions based on the notion of data depth. The proposed tests are extensions of the M-based test due to Li and Liu (2004). The performance of proposed tests has been assessed for symmetric as well as skewed multivariate distributions by simulation experiments. The tests have better performance in terms of power as compared to the M-based test and some of their competitors. The use of tests is illustrated with real life data.

keywords: Data depth, DD plot, Multivariate nonparametric tests, Location parameter, Permutation test.

1 Introduction

In several situations comparison between two data sets is required for number of reasons. The comparison can be based on the locations of these data sets. If multivariate data follow multivariate normal distribution then the task is easy as well known tests are available in the literature. However, if data do not follow multivariate normal distribution or we have no information about underlying distribution, nonparametric multivariate statistical methods are used to analyze data. One of the multivariate nonparametric statistical methods is based on the notion of the statistical data depth function, which was first introduced by Tukey (1975).

*Corresponding author: chavanatul2190@gmail.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

A data depth is a device for finding the location of multivariate data point with respect to a given data cloud. Larger depths are associated with more central points. Data depth gives a natural center-outward ranking to a multivariate data points with respect to data cloud. With the help of such rankings, Li and Liu (2004) proposed two depth-based nonparametric tests for multivariate location difference viz. T-based test and M-based test. These tests are developed using the Depth Depth (DD) plot (Liu et al., 1999). Dovoedo and Chakraborti (2015) have reported an extensive simulation study to evaluate the performance of these two tests for well known family of multivariate skewed distributions as well as multivariate symmetric distributions and compared performance of these tests for four popular affine-invariant depth functions, namely Mahalanobis depth, Spatial depth, Halfspace depth and Simplicial depth. We briefly discuss few of these in this article.

Several nonparametric tests have been proposed to deal with the multivariate two sample location problems as well as multi-sample location problems based on the concept of data depth. See Rousson (2002), Li et al. (2011), Chenouri and Small (2012) among others. Many of these methods are use permutation test to calculate the p-value.

In this paper we have proposed two nonparametric tests for testing equality of location parameters of two multivariate distributions based on the data depth, which are purely nonparametric. These tests are extensions of the M-based test introduced by Li and Liu (2004). Li and Liu (2004) use the most deepest point of two data clouds. We instead, consider some pre-specified number of most deepest points of the data clouds under comparison and construct tests based on these points. The performance of the proposed tests has been assessed by simulation experiments. The proposed tests give better performance in terms of power as compared to the M-based test and T-based test for symmetric as well as skewed multivariate distributions.

The rest of the paper is organized as follows. In section 2, we briefly discuss the notion of data depth, various data depth functions with their properties and DD plot. In section 3, we review the existing T-based and M-based tests of multivariate locations proposed by Li and Liu (2004). We describe the two new proposed nonparametric tests for testing the equality of locations using data depth in section 4. In section 5, we report simulation studies to compare performance of proposed tests with existing tests. In section 6, we apply the proposed tests to real life data. Section 7 contains some concluding remarks.

2 Statistical Data Depth Functions, Its Properties and DD Plot

2.1 Data Depth

Let $(X_1, X_2, ..., X_m)$ be a data set (cloud), where each $X_i \in \mathbb{R}^p$ is assumed to follow a continuous distribution with cumulative distribution function (CDF) $F(.)$, $i = 1, 2, ..., m$. Let $D(x, F)$ be the depth of a point x with respect to F. A data depth is a function defined from $\mathbb{R}^p$ to $[0, \infty)$. Notion of data depth can be used to obtain the location of a given data points with respect to a data cloud. It measures the centrality of a given

data point with respect to a given data cloud. The deepest point using notion of data depth has the largest depth. Data depth gives a natural center-outward ranking to a data points with respect to data cloud. Such rankings were used for testing difference in location or scale parameters of two or more multivariate distributions, constructing nonparametric control charts, outlier detection and classification problem etc.

Tukey (1975) has first invented the word depth for picturing data. In literature, many different notions of data depth functions were proposed for capturing different probabilistic properties of multivariate data. Among them, the most popular choices of data depth functions are Mahalanobis depth (Mahalanobis, 1936), Simplicial depth (Liu, 1990), majority depth (Singh, 1991), half-space depth (Tukey, 1975), projection depth (Donoho and Gasko, 1992) etc. Some of these depth functions are reviewed in the following.

- **Mahalanobis Depth**

The Mahalanobis depth of a point $x \in \mathbb{R}^p$ with respect to F on $\mathbb{R}^p$ is defined as,

$$MHD(x, F) = \frac{1}{(x-\mu_F)' \Sigma_F^{-1} (x-\mu_F)},$$

where μ_F is a location parameter or center and Σ_F is the variance covariance matrix or dispersion matrix of F. The sample version of Mahalanobis depth can be obtained by replacing μ_F by $\bar{X}$ (sample mean) and Σ_F by S (sample variance covariance matrix).

- **Simplicial Depth**

The simplicial depth of a point $x \in \mathbb{R}^p$ with respect to F on $\mathbb{R}^p$ is defined as,

$$SD(x, F) = Pr_F(s[X_1, X_2, ..., X_{p+1}] \ni x),$$

where $X_1, X_2, ..., X_{p+1}$ are independent and identically distributed observations from F and $s[X_1, X_2, ..., X_{p+1}]$ is a closed simplex whose vertices are $X_1, X_2, ..., X_{p+1}$. The Sample version of simplicial depth can be obtained by replacing F by F_m in this expression. That is,

$$SD(x, F_m) = \binom{m}{p+1}^{-1} \sum_* I(x \epsilon s[X_{i1}, X_{i2}, ..., X_{ip+1}]),$$

where $(*)$ runs over all possible subsets of $X_1, X_2, ..., X_m$ of size $(p+1)$. Larger the depth $SD(x, F_m)$ indicates x is contained in more simplices generated from the sample.

- **Tukey's Halfspace Depth**

Tukey's halfspace depth of a point $x \in \mathbb{R}^p$ with respect to probability measure P on $\mathbb{R}^p$ is defined as the minimum probability mass carried by any closed half space containing x, that is,

$$HSD(x, F) = \inf_H \{P(H) : H \text{ is a closed halfspace containing x } \},$$

The sample version of $HSD(x, F)$ is obtained by replacing F by F_m. If $k = 1$ then $HSD(x, F) = min\{F(x), 1 - F(x^-)\}$.

2.2 Properties of Depth Function

A depth function $D(x, F)$ is a non-negative function lies between $[0, \infty)$. According to Zuo and Serfling (2000), the depth function should satisfy the following four properties.

1. **Affine-invariance**: Suppose $x \in \mathbb{R}^p$ be a any given data point. Let A be any invertible matrix and $b \in \mathbb{R}^p$, then depth of a point $Ax + b$ with respect to F is equal to the depth of a point with respect to F. That is, $D(Ax + b, F) = D(x, F)$.

2. **Maximality at a center**: If F is centrally symmetric about $x_0 \in \mathbb{R}^p$, then depth of x_0 is the largest depth among all data points. That is,

$$D(x_0, F) \geq D(x, F) \text{ for any } x \in \mathbb{R}^p$$

3. **Monotonicity relative to any deepest point**: If $D(x_0, F) \geq D(x, F)$ for any $x \in \mathbb{R}^p$, then $D(x_0 + \lambda(x - x_0), F)$ is monotone non-increasing over $[0, \infty)$ for $\lambda \in [0, 1]$.

4. **Vanishing at infinity**: If $||x|| \longrightarrow \infty$ then $D(x, F) \longrightarrow 0$, where $||x||$ is the Euclidean norm in $\mathbb{R}^p$.

In the following section, we describe DD plot.

2.3 Depth-Depth Plot (DD Plot)

Let $(X_1, X_2, ..., X_m)$ and $(Y_1, Y_2, ..., Y_n)$ be two random samples from two continuous distributions F and G respectively, where X_i, $Y_j \in \mathbb{R}^p$, $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$. Let $D(x, F)$ and $D(x, G)$ be the depths of a point $x \in Z$ with respect to F and G respectively, where $Z = X \cup Y$. Let

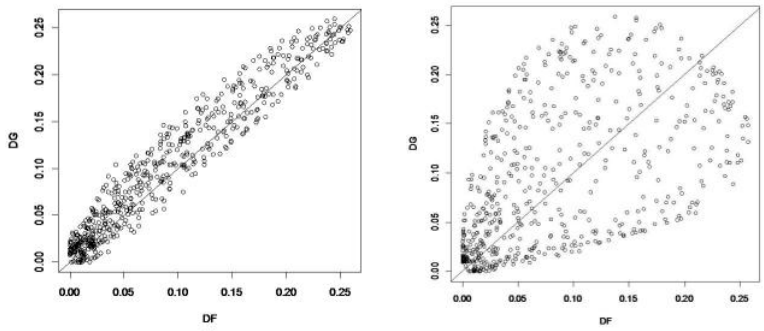
$$DD(F, G) = \{(D(x, F), D(x, G)), \quad \forall x \in Z\}.$$

The empirical version of $DD(F, G)$ based on the above described random samples is given by,

$$DD(F_m, G_n) = \{(D(x, F_m), D(x, G_n)), \quad \forall x \in Z\}.$$

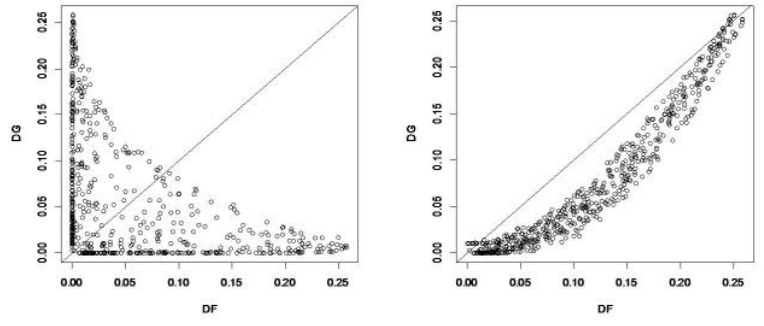
DD plot is a two-dimensional graph, which is the plot of points in the set $DD(F_m, G_n)$. The DD plot can be used as a convenient diagnostic tool for graphical comparison of two multivariate samples. Difference in locations or scales or skewness or kurtosis are associated with different patterns observed on the DD plots. If $F = G$ then the points on the empirical DD Plot should fall on a 45^0 line segment. This is illustrated in Figure 1(a), which is the DD plot of two multivariate samples drawn from the biariate normal distribution with mean vector $\underline{\mu} = \underline{0}$ and dispersion matrix I_2, where I_2 is the identity matrix of order two. That is $N_2(\underline{0}, I_2)$. The departure of F from G will indicate departure of points from 45^0 line segment and Figure 1(b), Figure 2(a), Figure 2(b) and

Figure 3 reveal different patterns of DD plot that indicate the location differences, large location differences, scale differences and skewness differences (both location and scale differences) respectively. From Figure 1(b), the DD plot has a leaf-shaped figure with the cusp lying on the diagonal line towards the upper right corner and the leaf steam at the lower left corner point (0,0) when there is a shift in location parameters of two multivariate samples. In each of these Figures, we plot DD plot of DG against DF where F and G are chosen appropriately, where DF and DG are the depth of the points with respect to F and G respectively. We use Simplicial depth as a depth function to plot the DD plot in figure 1, 2 and 3. The study reported here is based on Simplicial depth function. The DD plots have been plotted using 'depth' package available in R (R Core Team, 2016).



(a) (b)

Figure 1: DD plots of (a) $F = G = N_2(0, I_2)$ and (b) $F = N_2(0, I_2)$ and $G = N_2(0.5, I_2)$.



(a) (b)

Figure 2: DD plots of (a) $F = N_2(0, I_2)$ and $G = N_2(1.5, I_2)$ and (b) $F = N_2(0, I_2)$ and $G = N_2(0, 0.5I_2)$.

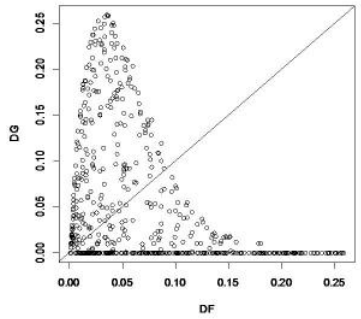


Figure 3: DD plot of $F = N_2(0, I_2)$ and $G = N_2(1, 0.1I_2)$.

In the following section, we describe T-based and M-based tests due to Li and Liu (2004).

3 T-based and M-based Tests

Li and Liu (2004) have proposed the T-based and the M-based tests for testing the equality of location parameters of two multivariate distributions by observing the DD plot introduced by Li and Liu (2004). These tests are completely nonparametric in nature.

Let $X = (X_1, X_2, ..., X_m)$ and $Y = (Y_1, Y_2, ..., Y_n)$, $X_i \in \mathbb{R}^p$, $Y_j \in \mathbb{R}^p$, $i = 1, 2, ..., m$, $j = 1, 2, ..., n$, be two data vectors observed from the distributions with CDF F and G respectively. Moreover, we assume that F and G are identical except for a possible location shift.

Let μ_1 and μ_2 be the location parameters of F and G respectively. The problem under consideration is to test

$$H_0 : \mu_1 = \mu_2 \quad \text{Vs} \quad H_1 : \mu_1 \neq \mu_2.$$

It is equivalent to test

$$H_0 : \theta = 0 \quad \text{Vs} \quad H_1 : \theta \neq 0,$$

where $\theta = \mu_1 - \mu_2$. That is θ is the shift in location parameters of two multivariate distributions.

3.1 The T-based Test

In the presence of location shift in two distribution, the DD plot has a leaf shaped figure (Figure 1(b), Figure 2(a)) with the leaf stem anchoring at the lower left corner point $(0, 0)$ and the cusp lying on the diagonal line pointing towards the upper right corner. On the basis of this observation, Li and Liu (2004) constructed the test statistic which

is the distance between the origin (0,0) and the cusp point. Li and Liu (2004) suggested the following procedure to calculate the distance between the cusp point and the origin (0,0).

For (a_1, b_1) and (a_2, b_2) in $\in \mathbb{R}^2$, define

$(a_1, b_1) \geq (a_2, b_2)$ if $a_1 \geq a_2$ and $b_1 \geq b_2$,

$(a_1, b_1) < (a_2, b_2)$ otherwise.

Define the set Q as

$$Q = \{z \in X \cup Y : \textit{there does not exist } w \in X \cup Y \textit{ s.t.}$$
$$(D(w, F_m), D(w, G_n)) \geq (D(z, F_m), D(z, G_n))\}.$$

Then the cusp point is the point $(D(z_c, F_m), D(z_c, G_n))$ that satisfies $z_c \in Q$ and $|D(z_c, F_m) - D(z_c, G_n)| \leq |D(z, F_m) - D(z, G_n)|$ for all $z \in Q$. Let $T = (D(z_c, F_m) + D(z_c, G_n))/2$. The distance between the origin (0,0) and the cusp point is approximately $\sqrt{2}T$. Li and Liu (2004) used T as a test statistic instead of using $\sqrt{2}T$ and smaller the value of T indicates the larger shift in location. The p-value of the test is obtained by using the Fisher's permutation test. Let

$$P_B^T = \frac{\sum_{i=1}^{B} I_{(T_i^* \leq T_{obs})}}{B},$$

where $I(.)$ is the indicator function, T_{obs} is the observed value of test statistic T calculated from the original combined sample, B is the number of times the combined sample $X \cup Y$ is permuted and T_i^* is the value of test statistic T corresponding to i^{th} permuted combined sample, $i = 1, 2, ..., B$.

3.2 The M-based Test

Li and Liu (2004) developed another test for testing the equality of location parameters of two multivariate distributions based on the deepest point. In the theory of data depth, the location parameter is the point having maximum depth. Therefore if the two distributions F and G are identical then they should have the same deepest point. If there is a shift in location then the deepest point corresponding to the distribution F would not be the deepest point corresponding to the distribution G. In fact, the deepest point of F will have a smaller depth value with respect to G. M-based test statistic due to Li and Liu (2004) is given by,

$$M = min\{D(v_1, F_m), D(u_1, G_n)\},$$

where v_1 is the deepest point of $X \cup Y$ corresponding to G_n, and u_1 is the deepest point of $X \cup Y$ corresponding to F_m. Here larger the location difference, smaller the value of M. The p-value of the test is obtained by using the Fisher's permutation test. Let

$$P_B^M = \frac{\sum_{i=1}^{B} I_{(M_i^* \leq M_{obs})}}{B},$$

where $I(.)$ is the indicator function, M_{obs} is the observed value of test statistic M calculated from the original combined sample, B is the number of times the combined sample $X \cup Y$ is permuted and M_i^* is the value of test statistic M corresponding to i^{th} permuted combined sample, $i = 1, 2, ..., B$.

4 Proposed Tests

In the M-based test, Li and Liu (2004) consider only single deepest point for constructing the M-based test statistic. The test based on single deepest point considers a single data point. There is scope for improving the performance of this test by incorporating few more data points while constructing the test. This can be achieved by considering more than one deepest point. We propose the following two test statistic which are based on k ($k \geq 2$) deepest points for above hypothesis testing problem which can be considered as extensions of the previously discussed M-based test.

Suppose the set U consists of the k deepest points in $X \cup Y$ with respect to F_m and the set V consists of the k deepest points in $X \cup Y$ with respect to G_n. Then we define two test statistic as follows,

- **M_1-based test statistic**

$$M_1 = min\{\tfrac{1}{k}\sum_{i=1}^{k} D(u_i, G_n), \tfrac{1}{k}\sum_{i=1}^{k} D(v_i, F_m)\},$$

- **M_2-based test statistic**

$$M_2 = \tfrac{1}{k}\sum_{i=1}^{k}(\min_i(D(u_i, G_n), D(v_i, F_m))),$$

where u_i is the i^{th} point of the set U and v_i is the i^{th} point of the set V. Here for both of these two test statistic, larger the location difference, smaller the value of M_1 as well as of M_2. Therefore we propose two tests based on the above defined two statistic. Each test rejects H_0 for smaller value of the corresponding statistic.

The p-value of the proposed tests are obtained by using the Fisher's permutation test. Let

$$P_B^{M_1} = \frac{\sum_{i=1}^{B} I_{(M_{1i}^* \leq M_{1obs})}}{B},$$

where $I(.)$ and B are defined as earlier, M_{1obs} is the observed value of test statistic M_1 calculated from the original combined sample and M_{1i}^* is the value of test statistic M_1 corresponding to i^{th} permuted combined sample, $i = 1, 2, ..., B$. Similarly, we can calculate the p-value for test statistic M_2.

5 Performance of Tests

We have carried out extensive simulation study to assess the performance of two proposed tests, T-based, M-based and Hotelling T^2 tests for a bivariate data. The performance of proposed tests has been evaluated in terms of power for two Bivariate symmetric distributions (Bivariate normal, Bivariate Cauchy) as well as two Bivariate skewed distributions with pattern 1 and pattern 2 (Bivariate skew normal; Azzalini, 2005), bivariate skew-t

distribution (Azzalini and Capitanio, 2003). In the simulation study, the number of observations generated from each distribution F and G are taken to be m=n=100 and the original sample is permuted B=500 times. The power of T-based, M-based, Hotelling T^2, M_1-based and M_2-based tests are obtained by the proportion of the simulated p-values less than equal to the level of significance $\alpha = 0.05$. Here 1000 simulations are used for reporting the power and also results are reported for various values of k=2,3,4,5. Distributions used in the simulation study are listed in Table-1.

Table 1: Distributions used in the simulation study

Distribution	Parameters
Symmetric normal	$N_2(\xi, \Omega = I)$
Symmetric cauchy	$Cauchy(\xi, \Omega = I)$
Skew-normal Pattern 1	$SN_2(\xi, \Omega = I, a = (10, 4)^T)$
Skew-normal Pattern 2	$SN_2(\xi, \Omega = I, a = (4, 10)^T)$
Skew-t Pattern 1	$ST_2(\xi, \Omega = I, a = (10, 4)^T, v = 1)$
Skew-t Pattern 2	$ST_2(\xi, \Omega = I, a = (10, 4)^T, v = 3)$

The parameter ξ denotes the location parameter, Ω denotes the dispersion parameter, a denotes the shape parameter (or skewness parameter) and v denotes the degrees of freedom. From all these distributions, the first random sample of size 100 is generated with parameter $\xi = (0, 0)^T$ and dispersion parameter Ω is an identity matrix of order 2 and second random sample of size 100 is generated with parameter $\xi = (\mu, \mu)^T$ and dispersion parameter Ω is an identity matrix of order 2. Details regarding shape parameter a and degrees of freedom v are provided in Table-1. We provide powers of all these discussed tests for different values of $\mu = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5$. R-software is used for simulation studies.

Table-8 provides powers for T-based, M-based, Hotelling T^2 and proposed tests when F is bivariate Cauchy distribution with parameters $((0, 0), I_2)$ and G is bivariate normal distribution with parameters $((\mu_1, \mu_2), I_2)$ with sample sizes m=n=100 and Table-9 provides powers for T-based, M-based, Hotelling T^2 and proposed tests when F is trivariate Cauchy distribution with parameters $((0, 0, 0), I_3)$ and G is trivariate normal distribution with parameters $((\mu_1, \mu_2, \mu_3), I_3)$ with sample sizes m=n=50.

Table 2: Power comparison of T-based, M-based, Hotelling T^2 and proposed tests when underlying distribution is bivariate normal with sample sizes $m = n = 100$ for simplicial depth function.

	μ	0.0	0.1	0.2	0.3	0.4	0.5
	T-based	0.046	0.094	0.267	0.505	0.773	0.950
	M-based	0.046	0.106	0.282	0.547	0.810	0.954
	Hotelling T^2	0.059	0.142	0.413	0.769	0.957	0.995
k=2	M_1-based	0.051	0.099	0.310	0.567	0.846	0.972
	M_2-based	0.052	0.091	0.317	0.584	0.847	0.966
k=3	M_1-based	0.048	0.108	0.324	0.598	0.857	0.973
	M_2-based	0.055	0.108	0.334	0.613	0.864	0.974
k=4	M_1-based	0.046	0.110	0.317	0.609	0.851	0.976
	M_2-based	0.048	0.107	0.324	0.633	0.864	0.979
k=5	M_1-based	0.045	0.113	0.337	0.614	0.859	0.984
	M_2-based	0.045	0.110	0.337	0.628	0.867	0.983

Table 3: Power comparison of T-based, M-based, Hotelling T^2 and proposed tests when underlying distribution is bivariate cauchy with sample sizes $m = n = 100$ for simplicial depth function.

	μ	0.0	0.1	0.2	0.3	0.4	0.5
	T-based	0.050	0.094	0.171	0.351	0.561	0.793
	M-based	0.058	0.090	0.169	0.366	0.568	0.801
	Hotelling T^2	0.019	0.023	0.026	0.033	0.038	0.045
k=2	M_1-based	0.057	0.093	0.189	0.396	0.587	0.818
	M_2-based	0.064	0.101	0.183	0.386	0.601	0.822
k=3	M_1-based	0.059	0.092	0.175	0.382	0.595	0.821
	M_2-based	0.061	0.093	0.185	0.378	0.609	0.819
k=4	M_1-based	0.057	0.097	0.185	0.393	0.601	0.816
	M_2-based	0.059	0.098	0.191	0.389	0.611	0.817
k=5	M_1-based	0.062	0.086	0.170	0.391	0.601	0.828
	M_2-based	0.059	0.091	0.181	0.385	0.608	0.823

Table 4: Power comparison of T-based, M-based, Hotelling T^2 and proposed tests when underlying distribution is bivariate skew-normal distribution, pattern 1 with sample sizes $m = n = 100$ for simplicial depth function.

	μ	0.0	0.1	0.2	0.3	0.4	0.5
	T-based	0.051	0.156	0.517	0.905	0.995	1.000
	M-based	0.051	0.176	0.587	0.914	0.992	0.999
	Hotelling T^2	0.047	0.262	0.783	0.995	1.000	1.000
k=2	M_1-based	0.052	0.181	0.641	0.946	0.998	1.000
	M_2-based	0.049	0.189	0.659	0.952	0.998	1.000
k=3	M_1-based	0.049	0.189	0.644	0.959	1.000	1.000
	M_2-based	0.055	0.200	0.671	0.964	1.000	1.000
k=4	M_1-based	0.046	0.190	0.656	0.967	1.000	1.000
	M_2-based	0.048	0.201	0.684	0.972	1.000	1.000
k=5	M_1-based	0.044	0.202	0.667	0.962	0.999	1.000
	M_2-based	0.049	0.216	0.686	0.973	1.000	1.000

Table 5: Power comparison of T-based, M-based, Hotelling T^2 and proposed tests when underlying distribution is bivariate skew-normal distribution, pattern 2 with sample sizes $m = n = 100$ for simplicial depth function.

	μ	0.0	0.1	0.2	0.3	0.4	0.5
	T-based	0.047	0.162	0.535	0.898	0.997	1.000
	M-based	0.053	0.202	0.603	0.935	0.995	1.000
	Hotelling T^2	0.054	0.262	0.791	0.989	1.000	1.000
k=2	M_1-based	0.044	0.220	0.654	0.951	0.998	1.000
	M_2-based	0.044	0.231	0.670	0.954	0.998	1.000
k=3	M_1-based	0.042	0.220	0.671	0.955	1.000	1.000
	M_2-based	0.047	0.224	0.688	0.962	1.000	1.000
k=4	M_1-based	0.042	0.218	0.668	0.957	1.000	1.000
	M_2-based	0.044	0.237	0.685	.968	1.000	1.000
k=5	M_1-based	0.049	0.214	0.673	0.957	1.000	1.000
	M_2-based	0.054	0.219	0.699	0.963	1.000	1.000

Table 6: Power comparison of T-based, M-based, Hotelling T^2 and proposed tests when underlying distribution is bivariate skew-t distribution, pattern 1 with sample sizes $m = n = 100$ for simplicial depth function.

	μ	0.0	0.1	0.2	0.3	0.4	0.5
	T-based	0.040	0.119	0.353	0.740	0.938	0.991
	M-based	0.049	0.147	0.451	0.811	0.961	0.999
	Hotelling T^2	0.040	0.128	0.281	0.561	0.801	0.928
k=2	M_1-based	0.052	0.137	0.483	0.847	0.976	1.000
	M_2-based	0.050	0.158	0.499	0.854	0.976	1.000
k=3	M_1-based	0.060	0.170	0.513	0.867	0.982	0.999
	M_2-based	0.052	0.177	0.527	0.882	0.985	0.999
k=4	M_1-based	0.055	0.162	0.528	0.882	0.981	1.000
	M_2-based	0.053	0.168	0.536	0.889	0.986	1.000
k=5	M_1-based	0.055	0.155	0.520	0.903	0.988	1.000
	M_2-based	0.051	0.170	0.548	0.905	0.989	1.000

Table 7: Power comparison of T-based, M-based, Hotelling T^2 and proposed tests when underlying distribution is bivariate skew-t distribution, pattern 2 with sample sizes $m = n = 100$ for simplicial depth function.

	μ	0.0	0.1	0.2	0.3	0.4	0.5
	T-based	0.053	0.080	0.194	0.371	0.603	0.799
	M-based	0.046	0.082	0.251	0.482	0.748	0.911
	Hotelling T^2	0.016	0.017	0.023	0.031	0.036	0.048
k=2	M_1-based	0.054	0.097	0.272	0.541	0.804	0.940
	M_2-based	0.052	0.092	0.274	0.537	0.805	0.944
k=3	M_1-based	0.055	0.096	0.284	0.575	0.819	0.948
	M_2-based	0.060	0.093	0.295	0.589	0.822	0.950
k=4	M_1-based	0.051	0.095	0.308	0.584	0.841	0.952
	M_2-based	0.055	0.085	0.300	0.593	0.840	0.955
k=5	M_1-based	0.047	0.111	0.314	0.604	0.846	0.960
	M_2-based	0.051	0.103	0.306	0.611	0.852	0.960

Table 8: Power comparison of T-based, M-based, Hotelling T^2 and proposed tests when $F : Cauchy((0,0), I_2)$ and $G : N_2((\mu_1, \mu_2), I_2)$ with sample sizes $m = n = 100$ for simplicial depth function.

	(μ_1, μ_2)	(0.1,0)	(0,0.2)	(0.1,0.2)	(0.3,0.3)
	T-based	0.057	0.078	0.074	0.160
	M-based	0.087	0.107	0.135	0.277
	Hotelling T^2	0.019	0.028	0.032	0.049
k=2	M_1-based	0.101	0.145	0.166	0.346
	M_2-based	0.092	0.151	0.171	0.360
k=3	M_1-based	0.115	0.151	0.185	0.402
	M_2-based	0.102	0.153	0.185	0.410
k=4	M_1-based	0.126	0.177	0.206	0.438
	M_2-based	0.114	0.167	0.195	0.441
k=5	M_1-based	0.129	0.190	0.221	0.469
	M_2-based	0.113	0.177	0.196	0.457

Table 9: Power comparison of T-based, M-based, Hotelling T^2 and proposed tests when $F : Cauchy((0,0,0), I_3)$ and $G : N_3((\mu_1, \mu_2, \mu_3), I_3)$ with sample sizes $m = n = 50$ for simplicial depth function.

	(μ_1, μ_2, μ_3)	(0.0, 0.0, 0.1)	(0.0, 0.2, 0.0)	(0.0, 0.1, 0.2)	(0.1, 0.2, 0.3)
	T-based	0.062	0.064	0.078	0.103
	M-based	0.096	0.096	0.138	0.173
	Hotelling T^2	0.026	0.029	0.031	0.051
k=2	M_1-based	0.107	0.143	0.151	0.221
	M_2-based	0.084	0.124	0.134	0.218
k=3	M_1-based	0.097	0.134	0.151	0.215
	M_2-based	0.088	0.116	0.135	0.214
k=4	M_1-based	0.107	0.120	0.148	0.197
	M_2-based	0.095	0.098	0.143	0.198
k=5	M_1-based	0.096	0.110	0.134	0.192
	M_2-based	0.087	0.106	0.123	0.187

It is clear from the power comparison Table-2 to Table-9 that the proposed M_1-based and M_2-based tests give better performance in terms of power as compared to the T-based, M-based and Hotelling T^2 tests for skewed multivariate distributions as well as

multivariate cauchy distribution with a simplicial depth function. Proposed tests also give comparable results to Hotelling T^2, when the underlying distribution is bivariate normal. As such there is no criterion defined to choose an optimal value of k. However $k = 5$ appears to be reasonably good choice for majority of distributions. Between M_1-based and M_2-based tests, we recommend M_2-based test, as it has more power than M_1-based test for most of the distributions.

6 Application to Real Life Data

We consider Iris dataset (Fisher, 1936), which contains 150 observations each 50 for setosa, versicolor and virginica with four variables sepal length, sepal width, petal length and petal width. These are three populations corresponding to setosa, versicolor and virginica respectively. We select only two populations namely setosa and versicolor for illustration. The location parameters consists of values of sepal length, sepal width, petal length and petal width in the respective populations.

We are interested in testing equality of location parameters of these two populations. Multivariate normality test for setosa and versicolor data based on Shapiro test gives p-value 0.07906 and 0.00574 respectively. Therefore, sepal length, sepal width, petal length and petal width corresponding to versicolor population do not follow four variate normal distribution and Hotelling T^2 test is not appropriate in this case. Therefore, we use proposed tests to evaluate whether there is shift in location parameters of distribution of setosa and versicolor. The p-values for the proposed tests based on $B = 500$ permutations are reported in the following Table.

Table 10: T-based, M-based, M_1-based and M_2-based p-values for the Iris dataset based on $B = 500$ permutations using simplicial depth function

	Test	p-value
	T-based	0.000
	M-based	0.148
k=2	M_1-based	0.034
	M_2-based	0.036
k=3	M_1-based	0.014
	M_2-based	0.018
k=4	M_1-based	0.006
	M_2-based	0.008
k=5	M_1-based	0.002
	M_2-based	0.004

It is clear from the Table-10 that all the p-values of the proposed and T-based tests indicates that setosa and versicolor populations do not have same location but M-based test fails to conclude that setosa and versicolor populations do not have same location.

7 Conclusion

In this paper, we use data depth approach for comparing location parameters of two multivariate distributions. The proposed tests are purely nonparametric tests. They have a better performance in terms of power as compared to the existing M-Based and T-based test for symmetric as well as skewed multivariate distributions. Notion of data depth is useful for testing location and/or scale of two multivariate distributions.

Acknowledgement

The authors would like to thank both referee and the co-editor for their helpful comments. The authors would also like to thank University Grants Commission, New Delhi for providing financial assistance to carry out the research work under Special Assistance Programme (F.520/8/DRS-I/2016(SAP-I)).

References

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188.

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.

Chenouri, S. and Small, C. G. (2012). A nonparametric multivariate multisample test based on data depth. *Electronic Journal of Statistics*, 6:760–782.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, pages 1803–1827.

Dovoedo, Y. and Chakraborti, S. (2015). Power of depth-based nonparametric tests for multivariate locations. *Journal of Statistical Computation and Simulation*, 85(10):1987–2006.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Li, J., Ban, J., and Santiago, L. S. (2011). Nonparametric tests for homogeneity of species assemblages: a data depth approach. *Biometrics*, 67(4):1481–1488.

Li, J. and Liu, R. Y. (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science*, pages 686–696.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh). *The annals of statistics*, 27(3):783–858.

Mahalanobis, P. (1936). Mahalanobis distance. In *Proceedings National Institute of Science of India*, volume 49, pages 234–256.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rousson, V. (2002). On distribution-free tests for the multivariate two-sample location-scale model. *Journal of multivariate analysis*, 80(1):43–57.

Singh, K. (1991). A notion of majority depth. *Unpublished document*.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of statistics*, pages 461–482.

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

Robust Linearized Ridge M-estimator for Linear Regression Model

N. H. Jadhav & D. N. Kashid

To cite this article: N. H. Jadhav & D. N. Kashid (2016) Robust Linearized Ridge M-estimator for Linear Regression Model, Communications in Statistics - Simulation and Computation, 45:3, 1001-1024, DOI: 10.1080/03610918.2014.911898

To link to this article: https://doi.org/10.1080/03610918.2014.911898

Accepted author version posted online: 11 Sep 2014.
Published online: 11 Nov 2015.

Communications in Statistics—Simulation and Computation®, 45: 1001–1024, 2016

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2014.911898

Taylor & Francis
Taylor & Francis Group

Robust Linearized Ridge M-estimator for Linear Regression Model

N. H. JADHAV[1] AND D. N. KASHID[2]

[1]Department of Statistics, D. R. K. College of Commerce, Kolhapur, Maharashtra, India
[2]Department of Statistics, Shivaji University, Kolhapur, Maharashtra, India

In the multiple linear regression, multicollinearity and outliers are commonly occurring problems. They produce undesirable effects on the ordinary least squares estimator. Many alternative parameter estimation methods are available in the literature which deals with these problems independently. In practice, it may happen that the multicollinearity and outliers occur simultaneously. In this article, we present a new estimator called as Linearized Ridge M-estimator which combats the problem of simultaneous occurrence of multicollinearity and outliers. A real data example and a simulation study is carried out to illustrate the performance of the proposed estimator.

Keywords Linearized ridge regression estimator; M-estimator; Mean square error; Multicollinearity; Outlier.

Mathematics Subject Classification 62J05; 62J07.

1. Introduction

Consider the multiple linear regression model

$$Y = X\beta + \varepsilon \tag{1.1}$$

where Y is an $n \times 1$ vector of observations on the response variable, X is a known $n \times p$ matrix of regressor variables, β is a $p \times 1$ vector of unknown regression coefficients and ε is an $n \times 1$ vector of errors with $E(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2 I$ and σ^2 is an unknown error variance. Without loss of generality, we assume that the variable Y and X are standardized in such a way that $X'Y$ denotes the correlation vector between the response variable and regressor variables and $X'X$ has the form of correlation matrix.

It is well known that, when $\varepsilon \sim N\left(0, \sigma^2 I\right)$, then the optimal estimator of the regression parameters is the ordinary least squares estimator (OLSE) (Montgomery et al., 2010). It is

Received November 7, 2012; Accepted March 31, 2014
Address correspondence to N. H. Jadhav, Department of Statistics, D. R. K. College of Commerce, Kolhapur 416002, Maharashtra, India; E-mail: dnk_stats@unishivaji.ac.in
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

denoted by

$$\hat{\beta}_{OLSE} = (X'X)^{-1}X'Y \tag{1.2}$$

The OLSE is widely used in regression analysis due to its computational ease. However, in the presence of multicollinearity, the OLSE gives misleading information. To overcome such a problem, several methods are available in the literature. The ordinary ridge regression estimator (ORRE) proposed by Hoerl and Kennard (1970a, b) is one of the most popular biased estimators. It is given by

$$\hat{\beta}_{ORRE} = (X'X + kI)^{-1}X'X\hat{\beta}_{OLSE} \tag{1.3}$$

where $k > 0$ is a ridge or shrinkage parameter. However, $\hat{\beta}_{ORRE}$ is a nonlinear function of k. To resolve such a problem, Liu (1993) proposed a new biased estimator of β called Generalized Liu estimator (GLE)

$$\hat{\beta}_{GLE} = (X'X + I)^{-1}(X'X + D)\hat{\beta}_{OLSE} \tag{1.4}$$

where $D = \text{diag}(d_1, d_2, \ldots, d_p)$, $0 < d_j < 1$, $j = 1, 2, \ldots, p$ (see Akdeniz and Kaciranlar, 1995). When $d_1 = d_2 = \ldots = d_p = d$, the $\hat{\beta}_{GLE}$ reduces to the Liu estimator (LE) (see Liu, 1993) and it is given by $\hat{\beta}_{LE} = (X'X + I)^{-1}(X'X + dI)\hat{\beta}_{OLSE}$, where d is a Liu parameter. The advantage of the LE over the ORRE is that the $\hat{\beta}_{LE}$ is a linear function of d. Therefore, it is easier to choose d in $\hat{\beta}_{LE}$ than to choose k in $\hat{\beta}_{ORRE}$. Some authors like Kaciranlar et al. (1999), Akdeniz and Erol (2003), Alheety and Kibriya (2009) defined the LE for $d \in R$ and the GLE for each $d_j \in R, \forall j = 1, 2, \ldots, p$.

Motivated by the work of Liu (1993), Liu and Gao (2011) proposed a linearized ridge regression estimator (LRRE) to combat the problem of multicollinearity. It is given by

$$\hat{\beta}_{LRRE} = (X'X + I)^{-1}(X'X + QDQ')\hat{\beta}_{OLSE} \tag{1.5}$$

where $D = \text{diag}(d_1, d_2, \ldots, d_p)$, $d_j \in R, \forall j = 1, 2, \ldots, p$ and $Q = (q_1, q_2, \ldots, q_p)$ is an orthogonal matrix such that $Q'X'XQ = \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, $\lambda_1, \lambda_2, \ldots, \lambda_p \geq 0$ are the eigenvalues of $X'X$ and $q_1, q_2, \ldots, q_p$ are the corresponding eigenvectors. Range on the value taken by the diagonal elements of the shrinkage matrix D differentiates the LRRE from the GLE.

Gao and Liu (2011) considered a well-known class of estimators (see Groß, 2003; Hocking et al., 1976; Obenchain, 1975) known as generalized shrinkage estimator (GSE). This class contains OLSE, ORRE, LE, LRRE and many other shrinkage estimators. Gao and Liu (2011) shows that the MSE of the LRRE is not larger than the MSE of any other estimator in the class of GSE. However, by substituting the optimal values of the shrinkage parameter in the MSE, one can obtain the lower bound of the MSE for any estimator in the class.

Another important problem in regression analysis which has addressed by many authors is the presence of outliers in the data. The OLSE is sensitive to the presence of outliers in response variable (Y) (Hampel et al., 1986; Huber, 1964; Huber, 1981). To handle this problem, various robust estimators are put forwarded in the literature like M-estimator (ME), least trimmed squares estimator (LTSE), least median squares estimator (LMSE) (see Rousseeuw and Leroy, 1987). The ME is the most popular robust estimator for the

presence of outliers in the response variable (Y) and it is obtained by minimizing

$$\sum_{i=1}^{n} \rho\left(\frac{Y_i - x_i'\beta}{s}\right) \tag{1.6}$$

where$\rho(\cdot)$ is any robust criterion function and s is an estimate of scale parameter (see Birkes and Dodge, 1993; Groß, 2003; Huber and Ronchetti, 2009 Maronna et al., 2006 Tiku and Akkaya, 2004). To obtain the estimate of β, partially differentiate Eq. (1.6) with respect to each parameter and equate to zero, we get p nonlinear equations of the form

$$\sum_{i=1}^{n} \varphi\left(\frac{Y_i - x_i'\beta}{s}\right) x_{ij} = 0,\ j = 1, 2, \ldots, p \tag{1.7}$$

where $\varphi(\cdot)$ is partial derivative of ρ with respect to β and x_{ij} denote the jth entry in the ith row of matrix X. The p equations obtained in Eq. (1.7) are solved iteratively. In this article, Huber's ρ function (Huber, 1964) is used as a robust criterion function. The OLSE is used as an initial estimates of regression parameters and the initial weight matrix W^0 is set to an identity matrix of order n. For the lth iteration, the diagonal weight matrix, W^l, with diagonal entries w_i^l, $i = 1, 2, \ldots, n$ is obtained as

$$w_i^l = \begin{cases} \frac{t}{\left|\left(Y_i - x_i'\hat{\beta}_{ME}^{(l-1)}\right)/S^{(l-1)}\right|} & \text{if} \quad \left|Y_i - x_i'\hat{\beta}_{ME}^{(l-1)}\right| > t \\ 1 & \text{if} \quad \left|Y_i - x_i'\hat{\beta}_{ME}^{(l-1)}\right| \le t \end{cases} \tag{1.8}$$

where t =1.345 and $\hat{\beta}_{ME}^{(l-1)}$ denote the ME of β at $(l-1)$th iteration. The estimate of scale parameter at $(l-1)$th iteration $\left(S^{(l-1)}\right)$ is obtained by using the formula $S^{(l-1)} = 1.4826$ median $|e_i^{(l-1)} -$ median $(e_i^{(l-1)})|$ where $e_i^{(l-1)} = Y_i - x_i'\hat{\beta}_{ME}^{(l-1)}$. At convergence, the iterative reweighted least square estimator (See Montgomery et al., 2010) is known as ME and is given by

$$\hat{\beta}_{ME} = (X'WX)^{-1}X'WY \tag{1.9}$$

where W is a weight matrix with diagonal entries w_i, $i = 1, 2, \ldots, n$ obtained at convergence of iterative reweighted least square estimator.

Several methods are available in the literature which deals with the problem of multicollinearity and outliers in the data separately. However, very few methods tackle the problem of simultaneous occurrence of multicollinearity and outliers. Silvapulle (1991) proposed a ridge M-estimator (RME) as a robust version of ORRE by shrinking the ME with the robust estimate of the shrinkage parameter k. It is defined as

$$\hat{\beta}_{RME} = (X'X + kI)^{-1}X'X\hat{\beta}_{ME} \tag{1.10}$$

This estimator is also a nonlinear function of shrinkage parameter$k$. Arslan and Billor (2000) proposed an alternative class of Liu-type M-estimators (LME) to handle the problem of multicollinearity and outliers simultaneously. It is given by

$$\hat{\beta}_{LME} = (X'X + I)^{-1}\left(X'X + dI\right)\hat{\beta}_{ME} \tag{1.11}$$

where $0 < d < 1$. Jadhav and Kashid (2011) proposed a robust version of jackknifed ridge regression estimator known as jackknifed ridge M-estimator (JRME). It is given by

$$\hat{\beta}_{JRME} = (I - k^2 Q'(X'X + kI)^{-2} Q)\hat{\beta}_{ME} \tag{1.12}$$

where k is a shrinkage parameter to be replaced by its robust estimate (Jadhav and Kashid, 2011). This $\hat{\beta}_{JRME}$ is also a nonlinear and complicated function of shrinkage parameter k. In this article, we proposed a robust version of LRRE. The objective of this proposed estimator is to combats the simultaneous occurrence of multicollinearity and outliers in the data.

The remaining article is organized as follows. In Section 2, we propose a linearized ridge M-estimator (LRME) which combats the simultaneous occurrence of multicollinearity and outliers in data. Also, the asymptotic MSE of the LRME is obtained in this section. In Section 3, the superiority of the LRME over the other estimators is presented. Also, a robust choice of the shrinkage matrix D with an iterative form of the LRME is obtained. In Section 4, a numerical example is presented and Section 5 covers an extensive simulation study to illustrate the performance of estimators through estimated MSE (EMSE) sense. Article ends with some concluding remarks.

2. Proposed Estimator—Linearized Ridge M-estimator

In this section, we propose a linearized ridge M-estimator (LRME) of unknown regression parameters β of regression model given in Eq. (1.1). It is defined as

$$\hat{\beta}_{LRME} = (X'X + I)^{-1}(X'X + QDQ')\hat{\beta}_{ME} \tag{2.1}$$

where Q is the matrix of eigenvectors $(q_1, q_2, \ldots, q_p)$ corresponding to eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of $X'X$ matrix and D is the diagonal matrix of shrinkage parameters $(d_1, d_2, \ldots, d_p)$ where $d_j \in R, \forall j = 1, 2, \ldots, p$.

Gao and Liu (2011) studied the properties of the LRRE and recommend to use not only theoretically but also in practice. However, this estimator is not robust to outliers in Y, because it is obtained by shrinking a non-robust estimator (OLSE) with the shrinkage matrix $\left(X'X + I\right)^{-1}\left(X'X + QDQ'\right)$. Therefore, we define a new estimator which shrinks the ME with the same shrinkage quantity. Thus, the proposed estimator will become a stable estimator for the presence of both multicollinearity and outliers in the data. Using the same motivation, we have studied the properties of the proposed estimator.

For simplicity, we use a canonical form of regression model for the further discussion and study. The regression model given in Eq. (1.1) can be written in canonical form as

$$Y = Z\alpha + \varepsilon \tag{2.2}$$

where $Z = XQ$ and $\alpha = Q'\beta$. Then the LRME of α can be written as

$$\hat{\alpha}_{LRME} = (\Lambda + I)^{-1}(\Lambda + D)\hat{\alpha}_{ME} \tag{2.3}$$

where $\hat{\alpha}_{ME}$ is the ME of α in the canonical form. Note that, because of the relation$\alpha = Q'\beta$, any estimator $\hat{\alpha}$ of α has a corresponding $\hat{\beta} = Q\hat{\alpha}$ and $\text{MSE}(\hat{\beta}) = \text{MSE}(\hat{\alpha})$ (see Sakallioglu and Kaciranlar, 2008). Hence, it is sufficient to consider only a canonical form.

Before studying the properties like bias, variance and MSE of the proposed estimator, one should consider the properties of the ME given in the following remark.

Remark 2.1. Birkes and Dodge (1993) noted that 'the distribution of the ME ($\hat{\alpha}_{ME}$) of α cannot be specified exactly, but for large n, under certain assumptions the distribution is approximately normal with mean vector α and covariance matrix Ω'. Arslan and Billor (2000) studied the performance of LME using the asymptotic properties of ME. The asymptotic unbiased estimate of Ω given by Arslan and Billor (2000) is $\hat{A}^2\Lambda^{-1}$ where $\hat{A}^2 = s^2(n-p)^{-1}\sum_{i=1}^{n}[\psi(r_i/s)]^2/[\frac{1}{n}\sum_{i=1}^{n}\psi'(r_i/s)]^2$ and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ is a matrix of eigenvalues of $Z'Z$ (see for details Huber and Ronchetti, 2009).

Hence, considering the Remark 2.1, we obtain the expressions of bias, covariance, and MSE of the LRME for large sample size as follows.

2.1. *Bias*

The bias of the LRME is given by

$$\begin{aligned} bias(\hat{\alpha}_{LRME}) &= E(\hat{\alpha}_{LRME}) - \alpha \\ &= E\left[(\Lambda + I)^{-1}(\Lambda + D)\hat{\alpha}_{ME}\right] - \alpha \\ &= (\Lambda + I)^{-1}(\Lambda + D)\alpha - \alpha \\ &= \alpha + (\Lambda + I)^{-1}(D - I)\alpha - \alpha \\ &= (\Lambda + I)^{-1}(D - I)\alpha \end{aligned} \tag{2.4}$$

2.2. *Covariance*

The covariance of the LRME is expressed as

$$\begin{aligned} cov(\hat{\alpha}_{LRME}) &= cov((\Lambda + I)^{-1}(\Lambda + D)\hat{\alpha}_{ME}) \\ &= (\Lambda + I)^{-1}(\Lambda + D)\,\text{cov}(\hat{\alpha}_{ME})(\Lambda + D)(\Lambda + I)^{-1} \\ &= (\Lambda + I)^{-1}(\Lambda + D)\,\Omega\,(\Lambda + D)(\Lambda + I)^{-1} \end{aligned} \tag{2.5}$$

2.3. *MSE*

The MSE of the LRME is as follows

$$\begin{aligned} MSE(\hat{\alpha}_{LRME}) &= tr(cov(\hat{\alpha}_{LRME})) + [bias(\hat{\alpha}_{LRME})]'[bias(\hat{\alpha}_{LRME})] \\ &= tr\left\{(\Lambda + I)^{-1}(\Lambda + D)\,\Omega\,(\Lambda + D)(\Lambda + I)^{-1}\right\} \\ &\quad + \alpha'(D - I)(\Lambda + I)^{-2}(D - I)\alpha \\ MSE(\hat{\alpha}_{LRME}) &= \sum_{j=1}^{p}\frac{(\lambda_j + d_j)^2}{(\lambda_j + 1)^2}\Omega_{jj} + \sum_{j=1}^{p}\frac{(d_j - 1)^2}{(\lambda_j + 1)^2}\alpha_j^2 \end{aligned} \tag{2.6}$$

where Ω_{jj} is jth diagonal element of Ω.

3. Superiority of the LRME

The MSE criterion is widely used as a measure of closeness between the estimates and the true values of the parameter. We use the MSE of the estimators for comparison of the performance of various estimators. In the following subsections, we have obtained the expressions for the difference in MSE of LRME and LRRE, ME, RME, JRME and LME. The conditional superiority of LRME over the other estimators is developed. It is observed that, under some conditions, the MSE of LRME is smaller than the MSE of other estimators.

3.1. *In the Presence of Multicollinearity*

In this subsection, we compare the MSE of LRME with the MSE of the LRRE and the condition under which the LRME shows smaller MSE than that of the LRRE is obtained.

Theorem 3.1. *If* $\Omega_{jj} < \sigma^2\lambda_j^{-1}$ *for every* j, *then* $MSE(\hat{\alpha}_{LRME}) < MSE(\hat{\alpha}_{LRRE})$ *for all* $d_j \in$, $j = 1, 2, \ldots, p$.

Proof. The MSE of the LRRE is given by

$$\begin{aligned} MSE(\hat{\alpha}_{LRRE}) &= tr(cov(\hat{\alpha}_{LRRE})) + [bias(\hat{\alpha}_{LRRE})]'[bias(\hat{\alpha}_{LRRE})] \\ &= \sigma^2 tr\left\{(\Lambda + I)^{-1}(\Lambda + D)\Lambda^{-1}(\Lambda + D)(\Lambda + I)^{-1}\right\} \\ &\quad + \alpha'(D - I)(\Lambda + I)^{-2}(D - I)\alpha \\ &= \sigma^2 \sum_{j=1}^{p} \frac{(\lambda_j + d_j)^2}{\lambda_j(\lambda_j + 1)^2} + \sum_{j=1}^{p} \frac{(d_j - 1)^2}{(\lambda_j + 1)^2}\alpha_j^2 \end{aligned} \tag{3.1}$$

Using Eqs. (2.6) and (3.1), the difference between the MSE of LRRE and LRME can be given by

$$MSE(\hat{\alpha}_{LRRE}) - MSE(\hat{\alpha}_{LRME}) = \sum_{j=1}^{p} \frac{(\lambda_j + d_j)^2}{(\lambda_j + 1)^2}\left[\sigma^2\lambda_j^{-1} - \Omega_{jj}\right] \tag{3.2}$$

Hence,

$$MSE(\hat{\alpha}_{LRRE}) - MSE(\hat{\alpha}_{LRME}) > 0$$

whenever

$$\sigma^2\lambda_j^{-1} > \Omega_{jj}, \text{ for all } j = 1, 2, \ldots, p.$$

When the problem of multicollinearity is severe, some of the λ_j's are too small. Consequently, the above condition gets satisfied whenever the datasets with multicollinearity and outlying observations are present in the data and the MSE of the LRME is smaller than the MSE of the LRRE.

3.2. *In the Presence of Outlier*

In this subsection, we compare the MSE of the LRME with the MSE of the ME. The condition, under which the MSE of the LRME is smaller than that of the ME is obtained and reported in the following theorem.

Theorem 3.2. *If* $\frac{(1+2\lambda_j+d_j)}{(1-d_j)} > \frac{\alpha_j^2}{\Omega_{\text{jj}}}$ *for all* $j = 1, 2, \ldots, p$, *then* $MSE(\hat{\alpha}_{LRME}) < MSE(\hat{\alpha}_{ME})$.

Proof. The difference between the MSE of ME and LRME can be given by

$$\begin{aligned} &MSE(\hat{\alpha}_{ME}) - MSE(\hat{\alpha}_{LRME}) \\ &= \sum_{j=1}^{p} \Omega_{jj} - \sum_{j=1}^{p} \frac{(\lambda_j + d_j)^2}{(\lambda_j + 1)^2} \Omega_{jj} + \sum_{j=1}^{p} \frac{(d_j - 1)^2}{(\lambda_j + 1)^2} \alpha_j^2 \\ &= \sum_{j=1}^{p} \left[1 - \frac{(\lambda_j + d_j)^2}{(\lambda_j + 1)} \right] \Omega_{\text{jj}} + \sum_{j=1}^{p} \frac{(d_j - 1)^2}{(\lambda_j + 1)^2} \alpha_j^2 \end{aligned}$$

By some simplifications, it leads to,

$$\begin{aligned} &MSE(\hat{\alpha}_{ME}) - MSE(\hat{\alpha}_{LRME}) \\ &= \sum_{j=1}^{p} \frac{(1 - d_j)}{(\lambda_j + 1)^2} \left\{ (1 + 2\lambda_j + d_j)\Omega_{\text{jj}} - (1 - d_j)\alpha_j^2 \right\} \end{aligned} \tag{3.3}$$

In order to make$MSE(\hat{\alpha}_{ME}) > MSE(\hat{\alpha}_{LRME})$, we have $\frac{(1+2\lambda_j+d_j)}{(1-d_j)} > \frac{\alpha_j^2}{\Omega_{\text{jj}}}$ for all j. Hence the proof. □

3.3. *In the Presence of Multicollinearity and Outlier*

Three estimators namely, the RME, JRME and LME are considered in this subsection for the purpose of comparison of MSE's of these estimators with that of the LRME. The MSE expressions of these estimators are as follows.

$$MSE(\hat{\alpha}_{RME}) = tr(cov(\hat{\alpha}_{RME})) + [bias(\hat{\alpha}_{RME})]' [bias(\hat{\alpha}_{RME})]$$

$$= \sum_{j=1}^{p} \frac{\lambda_j^2}{(\lambda_j + k)^2} \Omega_{\text{jj}} + \sum_{j=1}^{p} \frac{k^2}{(\lambda_j + k)^2} \alpha_j^2 \tag{3.4}$$

$$MSE(\hat{\alpha}_{JRME}) = tr(cov(\hat{\alpha}_{JRME})) + [bias(\hat{\alpha}_{JRME})]' [bias(\hat{\alpha}_{JRME})]$$

$$= \sum_{j=1}^{p} \left(1 - \frac{k^2}{(\lambda_j + k)^2} \right)^2 \Omega_{\text{jj}} + \sum_{j=1}^{p} \frac{k^4}{(\lambda_j + k)^4} \alpha_j^2 \tag{3.5}$$

$$MSE\left(\hat{\alpha}_{LME}\right) = tr\left(cov\left(\hat{\alpha}_{LME}\right)\right) + \left[bias\left(\hat{\alpha}_{LME}\right)\right]' \left[bias\left(\hat{\alpha}_{LME}\right)\right]$$

$$= \sum_{j=1}^{p} \frac{(\lambda_j + d)^2}{(\lambda_j + 1)^2} \Omega_{jj} + \sum_{j=1}^{p} \frac{(d-1)^2}{(\lambda_j + 1)^2} \alpha_j^2 \tag{3.6}$$

Based on these MSE expressions, we compare the MSE of the LRME with the MSE of these estimators. The conditions under which the LRME shows smaller MSE than that of the RME, JRME and LME are obtained in Theorem 3.3 as follows.

Theorem 3.3. ***(i)*** *If* $d_j < 1 - \frac{k(\lambda_j+1)}{(\lambda_j+k)}$ *for every* j, *then* $MSE\left(\hat{\alpha}_{LRME}\right) < MSE\left(\hat{\alpha}_{RME}\right)$.

(ii) *If* $d_j < 1 - \frac{k^2(\lambda_j+1)}{(\lambda_j+k)^2}$ *for every* j, *then* $MSE\left(\hat{\alpha}_{LRME}\right) < MSE\left(\hat{\alpha}_{JRME}\right)$.

(iii) *If each* d_j, $j = 1, 2, \ldots, p$ *satisfy any one of the following condition*

(a) $\frac{2(\alpha_j^2 - \lambda_j \Omega_{jj})}{(\Omega_{jj} + \alpha_j^2)} - d < d_j < d$

(b) $d < d_j < \frac{2(\alpha_j^2 - \lambda_j \Omega_{jj})}{(\Omega_{jj} + \alpha_j^2)} - d$ *then* $MSE\left(\hat{\alpha}_{LRME}\right) < MSE\left(\hat{\alpha}_{LME}\right)$.

Proof. Proof of the part (i)

The difference between the MSE of RME and LRME is given by

$$MSE\left(\hat{\alpha}_{RME}\right) - MSE\left(\hat{\alpha}_{LRME}\right)$$

$$= \sum_{j=1}^{p} \left[\frac{\lambda_j^2}{\left(\lambda_j + k\right)^2} - \frac{\left(\lambda_j + d_j\right)^2}{\left(\lambda_j + 1\right)^2} \right] \Omega_{jj} + \sum_{j=1}^{p} \left[\frac{k^2}{\left(\lambda_j + k\right)^2} - \frac{\left(d_j - 1\right)^2}{\left(\lambda_j + 1\right)^2} \right] \alpha_j^2 \tag{3.7}$$

After simplifying (3.7), one can easily find that the $MSE\left(\hat{\alpha}_{RME}\right) - MSE\left(\hat{\alpha}_{LRME}\right) > 0$ when $d_j < 1 - \frac{k(\lambda_j+1)}{(\lambda_j+k)}$ for all $j = 1, 2, \ldots, p$.

Proof of part (ii)

Consider the difference between the MSE of JRME and LRME as

$$MSE\left(\hat{\alpha}_{JRME}\right) - MSE\left(\hat{\alpha}_{LRME}\right)$$

$$= \sum_{j=1}^{p} \left[\left(1 - \frac{k^2}{\left(\lambda_j + k\right)^2} \right)^2 - \frac{\left(\lambda_j + d_j\right)^2}{\left(\lambda_j + 1\right)^2} \right] \Omega_{jj} + \sum_{j=1}^{p} \left[\frac{k^4}{\left(\lambda_j + k\right)^4} - \frac{\left(d_j - 1\right)^2}{\left(\lambda_j + 1\right)^2} \right] \alpha_j^2 \tag{3.8}$$

After some simplification, we observe that, $MSE\left(\hat{\alpha}_{JRME}\right) > MSE\left(\hat{\alpha}_{LRME}\right)$ if $d_j < 1 - \frac{k^2(\lambda_j+1)}{(\lambda_j+k)^2}$ for all j.

Note that, $\frac{k(\lambda_j+1)}{(\lambda_j+k)} > \frac{k^2(\lambda_j+1)}{(\lambda_j+k)^2}$. It clearly indicates that the MSE of LRME is less than MSE of JRME whenever the MSE of LRME is less than MSE of RME.

Proof of part (iii)

The difference between the MSE of LME and LRME can be given by

$$\begin{aligned} &MSE\left(\hat{\alpha}_{LME}\right) - MSE\left(\hat{\alpha}_{LRME}\right) \\ &= \sum_{j=1}^{p}\left[\left(\lambda_j + d\right)^2 - \left(\lambda_j + d_j\right)^2\right]\frac{\Omega_{jj}}{\left(\lambda_j+1\right)^2} \\ &+ \sum_{j=1}^{p}\left[(d-1)^2 - \left(d_j - 1\right)^2\right]\frac{\alpha_j^2}{\left(\lambda_j+1\right)^2} \\ &= \sum_{j=1}^{p}\left\{\left(d + d_j + 2\lambda_j\right)\Omega_{jj} + \left(d + d_j - 2\right)\alpha_j^2\right\}\frac{\left(d-d_j\right)}{\left(\lambda_j+1\right)^2} \end{aligned} \tag{3.9}$$

$$MSE\left(\hat{\alpha}_{LME}\right) - MSE\left(\hat{\alpha}_{LRME}\right) > 0$$

if $(d + d_j + 2\lambda_j)\Omega_{jj} + (d + d_j - 2)\alpha_j^2 > 0$ and $(d - d_j) > 0$ for some or all j or $(d + d_j + 2\lambda_j)\Omega_{jj} + (d + d_j - 2)\alpha_j^2 < 0$ and $(d - d_j) < 0$ for remaining j.

After rearrangement of the terms, it can be written as

$$d_j\left(\Omega_{jj} + \alpha_j^2\right) + \left(d + 2\lambda_j\right)\Omega_{jj} + (d-2)\,\alpha_j^2 > 0$$

and $d > d_j$ for some or all j or

$$d_j\left(\Omega_{jj} + \alpha_j^2\right) + \left(d + 2\lambda_j\right)\Omega_{jj} + (d-2)\,\alpha_j^2 < 0$$

and $d < d_j$ for remaining j.

By simplification, we get

$$d_j > \frac{2\left(\alpha_j^2 - \lambda_j\Omega_{jj}\right)}{\left(\Omega_{jj} + \alpha_j^2\right)} - d$$

and $d_j < d$ for some or all j or

$$d_j < \frac{2\left(\alpha_j^2 - \lambda_j\Omega_{jj}\right)}{\left(\Omega_{jj} + \alpha_j^2\right)} - d$$

and $d_j > d$ for remaining j.

This implies that, the $MSE\left(\hat{\alpha}_{LME}\right) - MSE\left(\hat{\alpha}_{LRME}\right) > 0$ if

$$\frac{2\left(\alpha_j^2 - \lambda_j\Omega_{jj}\right)}{\left(\Omega_{jj} + \alpha_j^2\right)} - d < d_j < d$$

for some or all j or

$$\frac{2\left(\alpha_j^2 - \lambda_j\Omega_{jj}\right)}{\left(\Omega_{jj} + \alpha_j^2\right)} - d < d_j < d$$

for remaining j.

This completes the proof of Theorem 3.3.

Remark 3.1 To obtain the LME, Arslan and Billor (2000) proposed a robust choice of d as

$$\hat{d}_M = 1 - \hat{A}^2 \left[\sum_{j=1}^{p} \frac{1}{\lambda_j(\lambda_j+1)} \Big/ \sum_{j=1}^{p} \frac{\alpha^2_{ME_j}}{(\lambda_j+1)^2} \right] \tag{3.10}$$

where $\hat{A}^2 = s^2 (n-p)^{-1} \sum_{i=1}^{n} [\psi(r_i/s)]^2 / [\frac{1}{n} \sum_{i=1}^{n} \psi'(r_i/s)]^2$, $s = 1.4826$ median $|e_i -$ median $(e_i)|$, $e_i = (Y_i - Z_i' \hat{\alpha}_{ME})$ (see Huber and Ronchetti, 2009). Note that, it is not guaranteed that the value of $\hat{d}_M$ always lie between 0 and 1.

Below, we obtain the optimal value of D by minimizing the MSE of LRME. Also, an iterative computational procedure is given to obtain the iterative LRME.

3.4. *Robust Choice of D*

The optimal values of $d_1, d_2, \ldots, d_p$ are those which minimizes the MSE of LRME. To obtain the optimal value of d_j, $j = 1, 2, \ldots, p$, we use the standard procedure.

Consider

$$g\left(d_1, d_2, \ldots, d_p\right) = MSE\left(\hat{\alpha}_{LRME}\right)$$

$$= \sum_{j=1}^{p} \frac{\left(\lambda_j + d_j\right)^2}{\left(\lambda_j+1\right)^2} \Omega_{jj} + \sum_{j=1}^{p} \frac{\left(d_j - 1\right)^2}{\left(\lambda_j+1\right)^2} \alpha_j^2 \tag{3.11}$$

Differentiate Eq. (3.11) with respect to d_j and equate to zero, it follows that,

$$\frac{\partial g\left(d_1, d_2, \ldots, d_p\right)}{\partial d_j} = \frac{2\left(\lambda_j + d_j\right)}{\left(\lambda_j+1\right)^2} \Omega_{jj} + \frac{2\left(d_j - 1\right)}{\left(\lambda_j+1\right)^2} \alpha_j^2$$

Therefore,

$$\frac{\partial g\left(d_1, d_2, \ldots, d_p\right)}{\partial d_j} = 0 \Rightarrow d_j = \frac{\alpha_j^2 - \lambda_j \Omega_{jj}}{\Omega_{jj} + \alpha_j^2} \quad j = 1, 2, \ldots, p \tag{3.12}$$

Moreover,

$$\frac{\partial^2 g\left(d_1, d_2, \ldots, d_p\right)}{\partial d_i \partial d_j} = \begin{cases} \frac{2\left(\Omega_{ii} + \alpha_i^2\right)}{(\lambda_i+1)^2} & when\ i = j \\ 0 & when\ i \neq j \end{cases}$$

Hence, $\frac{\partial^2 g(d_1, d_2, \ldots, d_p)}{\partial d_i \partial d_j} \geq 0 \, \forall \, i = j = 1, 2, \ldots, p$. This implies that $g(d_1, d_2, \ldots, d_p)$ is minimum at $d_j = \frac{\alpha_j^2 - \lambda_j \Omega_{jj}}{\Omega_{jj} + \alpha_j^2}$, $j = 1, 2, \ldots, p$.

After simplifying the expression of d_j given in Eq. (3.12), one can easily get

$$d_j = 1 - \frac{1 + \lambda_j}{1 + \left(\alpha_j^2 / \Omega_{jj}\right)} \tag{3.13}$$

- From the expression given in Eq. (3.13), it is found that the value of each d_j should be less than 1 as $\frac{1+\lambda_j}{1+(\alpha_j^2/\Omega_{jj})}$ has lower bound 0.
- If $\lambda_j < \alpha_j^2/\Omega_{jj}$, then the corresponding d_j lies between 0 and 1.
- If $\lambda_j > \alpha_j^2/\Omega_{jj}$, then the corresponding $d_j < 0$.

Unfortunately, the value of d_j depends on the unknown model parameters and so, for the practical purpose, we need to replace these unknowns with their suitable estimates. Hence, the estimator of d_j is obtained as

$$\hat{d}_j = \frac{\hat{\alpha}_{MEj}^2 - \lambda_j \hat{\Omega}_{jj}}{\hat{\Omega}_{jj} + \hat{\alpha}_{MEj}^2} \quad j = 1, 2, \ldots, p \tag{3.14}$$

An iterative method is also used to obtain the estimates of iterative LRME. In brief, we explain the procedure as follows.

Consider,

$$\hat{\beta}_{LRME}^{(0)} = \hat{\beta}_{LRME}|_{D=D^{(0)}} \tag{3.15}$$

with $D^{(0)} = \text{diag}(\hat{d}_1^{(0)}, \hat{d}_2^{(0)}, \ldots \hat{d}_p^{(0)})$, where $\hat{d}_j^{(0)} = \frac{\hat{\alpha}_{MEj}^2 - \lambda_j \hat{\Omega}_{jj}}{\hat{\Omega}_{jj} + \hat{\alpha}_{MEj}^2}$ and $\hat{\Omega}_{jj} = \hat{A}^2 \lambda_j^{-1}$. We continue to update β with $\hat{\beta}_{LRME}^{(0)} = Q\hat{\alpha}_{LRME}^{(0)}$ to get a new updated LRME. After iterating analogically, we get an iterative LRME as

$$\hat{\beta}_{LRME}^{(l)} = \hat{\beta}_{LRME}|_{D=D^{(l)}} \tag{3.16}$$

where, $D^{(l)} = \text{diag}(\hat{d}_1^{(l)}, \hat{d}_2^{(l)}, \ldots \hat{d}_p^{(l)})$ and $\hat{d}_j^{(l)} = \frac{(\hat{\alpha}_{LRMEj}^{(l-1)})^2 - \lambda_j \hat{\Omega}_{jj}}{\hat{\Omega}_{jj} + \left(\hat{\alpha}_{LRMEj}^{(l-1)}\right)^2}$, $i = 1, 2, \ldots, p, l = 1, 2, \ldots$

In order to implement the LRME in Section 4 and Section 5, we have used the estimate of d_j given in Eq. (3.14). Using the simulation study, it is found that the iterative LRME has a very fast convergence rate.

4. Numerical Example

To illustrate the theoretical results and to evaluate the performance of various estimators, a real data set on tobacco blends given by Myers (1990) is used. Arslan and Billor (2000) analyzed this data to study the performance of LME and the other estimators. This data contains 30 observations on four regressor variables X_1, X_2, X_3, and X_4 with the response variable Y that measure the amount of heat evolved from tobacco during the smoking process.

It is observed that, the variance inflation factor (VIF) values for this data are 324.1412, 45.1728, 173.2577, and 138.1753. It reveals the severe problem of multicollinearity. Also, two outliers in response variable (Y) are pointed out by Arslan and Billor (2000). Hence, the tobacco blends data suffers from the simultaneous occurrence of outliers and multicollinearity. For this dataset, the estimated MSE (EMSE) of each estimator is obtained by replacing all unknown parameters in the corresponding theoretical MSE expression of that estimator. For example, the EMSE of the LRME is obtained by using the expression given

Table 1
Estimates and EMSE of estimators

Estimators	$\hat{\alpha}_{OLSE}$	$\hat{\alpha}_{LRRE}$	$\hat{\alpha}_{ME}$	$\hat{\alpha}_{RME}$	$\hat{\alpha}_{JRME}$	$\hat{\alpha}_{LME}$	$\hat{\alpha}_{LRME}$
$\hat{\alpha}_1$	0.4857	0.4848	0.4888	0.4888	0.4888	0.4635	0.4883
$\hat{\alpha}_2$	−0.6727	−0.5574	−0.6500	−0.6500	−0.6500	−0.4858	−0.5714
$\hat{\alpha}_3$	−1.0746	−0.8784	−1.2319	−1.2319	−1.2319	−0.9172	−1.1143
$\hat{\alpha}_4$	1.4436	1.0547	0.8841	0.8840	0.8841	0.6572	0.5488
EMSE	1.1032	0.7521	0.6960	0.6960	0.6960	0.4844	0.4197

in Eq. (2.6) as

$$\text{EMSE} = \sum_{j=1}^{p} \frac{\left(\lambda_j + \hat{d}_j\right)^2}{\left(\lambda_j + 1\right)^2} \hat{A}^2/\lambda_j + \sum_{j=1}^{p} \frac{\left(\hat{d}_j - 1\right)^2}{\left(\lambda_j + 1\right)^2} \hat{\alpha}^2_{LRME_j} \tag{4.1}$$

where $\hat{d}_j$ is given in Eq. (3.14). We compare the LRME with the OLSE, LRRE, ME, RME, JRME and LME in EMSE sense. The estimates of different estimators with their EMSE are shown in Table 1.

From Table 1 it can be concluded that:

- The EMSE of LRME is smaller than the EMSE of other estimators. It reveals that the LRME shows largest reduction in EMSE.
- The estimates of the OLSE and LRRE reveals that the presence of outliers and multicollinearity affect the estimates of regression parameters.
- The estimates of unknown regression parameters and EMSE for ME, RME, and JRME are equal. Hence, the performance of ME, RME, and JRME is same for this data.
- The conditions obtained in Section 3 for the superiority of LRME hold for this data.

5. Simulation Study

In this section, we present a simulation study to evaluate the performance of proposed estimator. To achieve the required degree of multicollinearity, the following simulation design proposed by McDonald and Galarneau (1975) is used to generate regressor variables as

$$x_{ij} = \left(1 - \rho^2\right)^{1/2} \zeta_{ij} + \rho \zeta_{i(p+1)} \quad , i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, p \tag{5.1}$$

where ζ_{ij}'s are independent standard normal pseudo-random numbers, ρ^2 is the correlation between any two regressor variables. The ($p =$) 4 regressor variables are considered and n observations on the response variable Y are generated using the regression model

$$Y = 10 + 4X_1 + X_2 + 6X_3 + 2X_4 + \varepsilon. \tag{5.2}$$

where, $\varepsilon \sim N_n(0, \sigma^2 I)$. Note that, the choice of model given above is arbitrary and for sake of illustration, it is used here. The outlier observations are introduced artificially in the response variable by using the procedure (see Jadhav and Kashid, 2011) given as follows.

Table 2
ASEVAR and ASESB of estimators

			Without outlier			With one outlier		
ρ	$\hat{\beta}$		$\sigma^2 = 1$	$\sigma^2 = 25$	$\sigma^2 = 100$	$\sigma^2 = 1$	$\sigma^2 = 25$	$\sigma^2 = 100$
					$n = 30$			
0.9	OLSE	ASEVAR	0.0043	0.0919	0.2567	0.4777	0.5102	0.5862
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.0031	0.0269	0.0615	0.0903	0.0906	0.0989
		ASESB	0.0002	0.0045	0.0113	0.0180	0.0183	0.0201
	ME	ASEVAR	0.0044	0.0932	0.2611	0.5526	0.6663	0.8471
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.0042	0.0566	0.1013	0.4697	0.4179	0.4002
		ASESB	0.0000	0.0038	0.0137	0.0002	0.0011	0.0022
	JRME	ASEVAR	0.0044	0.0870	0.2108	0.5468	0.6263	0.7272
		ASESB	0.0000	0.0004	0.0044	0.0000	0.0001	0.0005
	LME	ASEVAR	0.0040	0.0736	1.6662	8.54E+04	6.10E+05	4.90E+05
		ASESB	0.0002	0.1180	12.8339	5.75E+07	1.60E+09	1.29E+09
		$\hat{d}_M^{\#}$	1000	703	493	554	281	153
	$\text{LME}_{\text{d\#}}$	ASEVAR	0.0040	0.0291	0.0653	0.0078	0.0137	0.0253
		ASESB	0.0002	0.0119	0.0261	0.0027	0.0054	0.0088
	LRME	ASEVAR	0.0032	0.0276	0.0637	0.0087	0.0106	0.0126
		ASESB	0.0002	0.0046	0.0116	0.0015	0.0019	0.0023
0.99	OLSE	ASEVAR	0.0355	0.7854	2.2499	4.5109	4.8326	5.6789
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.0148	0.1595	0.4375	0.7033	0.7806	0.8827
		ASESB	0.0021	0.0309	0.0861	0.1471	0.1615	0.1833
	ME	ASEVAR	0.0363	0.8001	2.2999	7.9859	6.6982	9.1633
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.0285	0.2333	0.5597	4.8477	2.5814	3.2081
		ASESB	0.0010	0.0409	0.1137	0.0041	0.0125	0.0191
	JRME	ASEVAR	0.0356	0.5734	1.4964	7.4119	5.2847	6.8620
		ASESB	0.0000	0.0175	0.0571	0.0003	0.0037	0.0077
	LME	ASEVAR	0.0179	13.8958	123.6060	2.50E+07	3.52E+07	9.25E+06
		ASESB	0.0071	231.4154	1.01E+04	1.35E+11	3.17E+11	5.98E+10
		$\hat{d}_M^{\#}$	943	473	417	444	254	165
	$\text{LME}_{\text{d\#}}$	ASEVAR	0.0173	0.1613	0.4594	0.0316	0.0838	0.1605
		ASESB	0.0049	0.0663	0.1872	0.0163	0.0322	0.0571
	LRME	ASEVAR	0.0150	0.1660	0.4553	0.0442	0.0602	0.0770
		ASESB	0.0022	0.0320	0.0893	0.0088	0.0116	0.0145
0.999	OLSE	ASEVAR	0.3526	7.7153	22.3229	45.0429	49.2971	55.4750
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.0777	1.4699	4.2749	7.7741	7.9657	8.5516
		ASESB	0.0149	0.2943	0.8544	1.5928	1.6400	1.7886
	ME	ASEVAR	0.3596	7.8720	22.6974	57.4998	55.7978	86.9774
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.1388	1.7702	5.0567	31.8218	19.3437	29.1870
		ASESB	0.0185	0.3873	1.0907	0.0487	0.1195	0.1931
	JRME	ASEVAR	0.2916	4.8864	13.7915	51.7203	41.1841	63.7751
		ASESB	0.0057	0.2024	0.5841	0.0050	0.0438	0.0845
	LME	ASEVAR	3.3599	1.03E+03	1.59E+03	4.37E+07	7.46E+07	7.75E+07
		ASESB	31.6896	1.25E+05	1.02E+05	1.22E+11	3.91E+11	3.85E+11
		$\hat{d}_M^{\#}$	542	437	439	451	235	159

(*Continued on next page*)

Table 2
ASEVAR and ASESB of estimators (*Continued*)

			Without outlier			With one outlier		
ρ	$\hat{\beta}$		$\sigma^2 = 1$	$\sigma^2 = 25$	$\sigma^2 = 100$	$\sigma^2 = 1$	$\sigma^2 = 25$	$\sigma^2 = 100$
	$LME_{d\#}$	ASEVAR	0.0790	1.4560	4.3420	0.3583	0.6813	1.5305
		ASESB	0.0324	0.6037	1.7790	0.1771	0.2885	0.6172
	LRME	ASEVAR	0.0812	1.5321	4.4095	0.4531	0.5816	0.7829
		ASESB	0.0155	0.3040	0.8685	0.0907	0.1171	0.1529
0.9999	OLSE	ASEVAR	3.5141	76.7675	223.5224	466.4939	491.7869	568.5576
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.6830	15.0387	42.5911	76.7899	80.8770	81.6369
		ASESB	0.1355	2.9669	8.5278	15.8614	16.7343	17.1806
	ME	ASEVAR	3.5869	77.9804	228.0065	579.8510	605.2126	860.2789
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.8545	17.6523	50.5052	281.8503	242.1621	266.4587
		ASESB	0.1790	3.9347	10.9946	0.5047	1.2503	2.2243
	JRME	ASEVAR	2.3093	47.8383	139.7290	495.1777	469.8509	603.1857
		ASESB	0.0917	2.0920	5.9069	0.0574	0.4371	0.9622
	LME	ASEVAR	273.6035	4.52E+03	2.56E+04	7.62E+08	5.29E+08	5.25E+09
		ASESB	2.17E+04	4.10E+05	3.43E+06	2.92E+12	1.59E+12	8.88E+13
		$\hat{d}_M^{\#}$	447	451	433	444	270	173
	$LME_{d\#}$	ASEVAR	0.6892	15.5898	42.0141	3.5634	7.4437	16.3258
		ASESB	0.2837	6.3166	17.3429	1.7641	2.9721	6.3012
	LRME	ASEVAR	0.7033	15.6570	44.5410	4.6181	5.8584	8.5160
		ASESB	0.1380	3.0616	8.8885	0.9305	1.1422	1.6547
					$n = 50$			
0.9	OLSE	ASEVAR	0.0024	0.0510	0.1442	0.2459	0.2778	0.3239
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.0019	0.0191	0.0369	0.0495	0.0533	0.0544
		ASESB	0.0001	0.0029	0.0065	0.0096	0.0105	0.0110
	ME	ASEVAR	0.0024	0.0522	0.1473	0.6509	0.2398	0.3333
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.0024	0.0380	0.0670	0.6153	0.1802	0.1664
		ASESB	0.0000	0.0018	0.0080	0.0000	0.0005	0.0013
	JRME	ASEVAR	0.0024	0.0509	0.1276	0.6501	0.2346	0.2970
		ASESB	0.0000	0.0001	0.0016	0.0000	0.0000	0.0003
	LME	ASEVAR	0.0023	0.0229	0.3581	3.70E+05	3.11E+04	3.82E+05
		ASESB	0.0001	0.0113	1.1358	4.54E+08	2.37E+07	1.475E+09
		$\hat{d}_M^{\#}$	1000	888	548	715	397	159
	$LME_{d\#}$	ASEVAR	0.0023	0.0220	0.0423	0.0037	0.0090	0.0128
		ASESB	0.0001	0.0076	0.0169	0.0010	0.0034	0.0058
	LRME	ASEVAR	0.0019	0.0195	0.0383	0.0052	0.0073	0.0083
		ASESB	0.0001	0.0030	0.0067	0.0008	0.0012	0.0015
0.99	OLSE	ASEVAR	0.0200	0.4351	1.2635	2.3310	2.5749	3.0369
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.0101	0.0942	0.2541	0.3958	0.4335	0.4932
		ASESB	0.0013	0.0181	0.0500	0.0810	0.0893	0.1019
	ME	ASEVAR	0.0205	0.4459	1.2903	5.2302	2.2697	3.3113
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.0177	0.1447	0.3128	3.9827	0.9198	0.9796
		ASESB	0.0004	0.0248	0.0689	0.0007	0.0070	0.0121
	JRME	ASEVAR	0.0204	0.3419	0.8582	5.1094	1.8725	2.3867

(*Continued on next page*)

Table 2
ASEVAR and ASESB of estimators (*Continued*)

			Without outlier			With one outlier		
ρ	$\hat{\beta}$		$\sigma^2=1$	$\sigma^2=25$	$\sigma^2=100$	$\sigma^2=1$	$\sigma^2=25$	$\sigma^2=100$
		ASESB	0.0000	0.0087	0.0332	0.0000	0.0022	0.0050
	LME	ASEVAR	0.0129	5.4266	48.6563	2.21E+07	6.85E+06	2.69E+06
		ASESB	0.0026	70.5478	2.09E+03	7.66E+10	5.74E+10	1.17E+10
		$\hat{d}_M^{\#}$	999	514	415	599	247	143
	$\text{LME}_{\text{d\#}}$	ASEVAR	0.0129	0.0941	0.2764	0.0155	0.0410	0.1010
		ASESB	0.0026	0.0399	0.1135	0.0067	0.0180	0.0384
	LRME	ASEVAR	0.0103	0.0978	0.2623	0.0173	0.0304	0.0437
		ASESB	0.0013	0.0186	0.0510	0.0032	0.0059	0.0081
0.999	OLSE	ASEVAR	0.1979	4.3103	12.3912	22.7979	25.6685	30.2683
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.0473	0.8126	2.3843	3.9627	4.2328	4.8167
		ASESB	0.0089	0.1616	0.4748	0.8180	0.8673	1.0072
	ME	ASEVAR	0.2023	4.4211	12.7255	61.4494	21.9338	28.0873
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.0902	0.9416	2.6421	36.9321	6.8520	7.1481
		ASESB	0.0104	0.2211	0.6480	0.0115	0.0660	0.1165
	JRME	ASEVAR	0.1751	2.7213	7.7349	56.6948	15.8073	18.6450
		ASESB	0.0022	0.1161	0.3527	0.0010	0.0303	0.0575
	LME	ASEVAR	0.7087	1.57E+03	3.82E+02	7.66E+09	4.87E+07	1.09E+08
		ASESB	2.9466	1.71E+05	1.90E+04	5.83E+14	2.85E+11	1.79E+12
		$\hat{d}_M^{\#}$	585	421	416	508	210	130
	$\text{LME}_{\text{d\#}}$	ASEVAR	0.0468	0.8578	2.4414	0.1197	0.3767	0.8127
		ASESB	0.0201	0.3588	1.0334	0.0583	0.1516	0.3588
	LRME	ASEVAR	0.0484	0.8504	2.4603	0.1362	0.2683	0.4089
		ASESB	0.0091	0.1676	0.4860	0.0267	0.0531	0.0798
0.9999	OLSE	ASEVAR	1.9786	42.2700	123.1427	230.3123	258.2842	295.7860
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.3855	7.9769	23.8499	39.2900	42.0808	46.2238
		ASESB	0.0775	1.5890	4.7206	8.0856	8.6971	9.6184
	ME	ASEVAR	2.0264	43.2754	125.4728	591.0670	272.3938	295.3610
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.4734	8.9636	26.5415	326.8068	83.0566	78.3865
		ASESB	0.1050	2.1803	6.4689	0.1210	0.6522	1.3010
	JRME	ASEVAR	1.3251	26.0572	75.8385	527.6826	194.5027	199.1760
		ASESB	0.0527	1.1907	3.4579	0.0172	0.3101	0.6333
	LME	ASEVAR	296.5514	6.58E+03	8.66E+03	1.04E+10	1.03E+09	1.48E+08
		ASESB	1.86E+04	1.46E+06	6.88E+05	1.20E+14	6.52E+12	5.19E+11
		$\hat{d}_M^{\#}$	429	416	435	486	196	135
	$\text{LME}_{\text{d\#}}$	ASEVAR	0.3953	8.4799	24.5826	1.2531	4.0254	10.0493
		ASESB	0.1676	3.6180	10.2417	0.6019	1.6802	3.8774
	LRME	ASEVAR	0.4022	8.4387	24.8891	1.3506	2.4498	4.6511
		ASESB	0.0798	1.6785	4.8816	0.2647	0.4788	0.9137
					$n=100$			
0.9	OLSE	ASEVAR	0.0011	0.0246	0.0684	0.1048	0.1213	0.1494
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.0010	0.0118	0.0207	0.0244	0.0258	0.0275
		ASESB	0.0000	0.0015	0.0035	0.0044	0.0048	0.0055
	ME	ASEVAR	0.0012	0.0252	0.0700	0.2669	0.0603	0.1042

(*Continued on next page*)

Table 2
ASEVAR and ASESB of estimators (*Continued*)

			Without outlier			With one outlier		
ρ	$\hat{\beta}$		$\sigma^2 = 1$	$\sigma^2 = 25$	$\sigma^2 = 100$	$\sigma^2 = 1$	$\sigma^2 = 25$	$\sigma^2 = 100$
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.0012	0.0211	0.0411	0.2638	0.0510	0.0620
		ASESB	0.0000	0.0006	0.0034	0.0000	0.0002	0.0006
	JRME	ASEVAR	0.0012	0.0250	0.0656	0.2669	0.0599	0.0978
		ASESB	0.0000	0.0000	0.0003	0.0000	0.0000	0.0001
	LME	ASEVAR	0.0011	0.0148	0.0481	1.39E+05	1.33E+02	3.61E+02
		ASESB	0.0000	0.0036	0.0632	1.59E+08	1.44E+04	6.48E+04
		$\hat{d}_M^{\#}$	1000	992	693	857	701	290
	$\text{LME}_{\text{d\#}}$	ASEVAR	0.0011	0.0149	0.0227	0.0018	0.0051	0.0059
		ASESB	0.0000	0.0035	0.0093	0.0003	0.0016	0.0025
	LRME	ASEVAR	0.0010	0.0120	0.0209	0.0029	0.0054	0.0053
		ASESB	0.0000	0.0016	0.0035	0.0004	0.0008	0.0009
0.99	OLSE	ASEVAR	0.0095	0.2059	0.6013	0.9727	1.1313	1.4078
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.0060	0.0491	0.1233	0.1732	0.2013	0.2334
		ASESB	0.0005	0.0092	0.0244	0.0350	0.0407	0.0483
	ME	ASEVAR	0.0098	0.2109	0.6142	2.1696	0.6537	1.0007
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.0091	0.0880	0.1572	1.9566	0.2738	0.2630
		ASESB	0.0001	0.0120	0.0347	0.0001	0.0040	0.0069
	JRME	ASEVAR	0.0097	0.1804	0.4308	2.1625	0.5589	0.7049
		ASESB	0.0000	0.0027	0.0158	0.0000	0.0008	0.0031
	LME	ASEVAR	0.0079	0.7719	7.9268	3.43E+07	3.63E+05	1.22E+05
		ASESB	0.0008	3.6105	112.355	3.44E+11	8.62E+08	1.47E+08
		$\hat{d}_M^{\#}$	1000	569	453	712	305	183
	$\text{LME}_{\text{d\#}}$	ASEVAR	0.0079	0.0495	0.1247	0.0064	0.0263	0.0320
		ASESB	0.0008	0.0215	0.0539	0.0020	0.0102	0.0145
	LRME	ASEVAR	0.0061	0.0508	0.1288	0.0066	0.0182	0.0252
		ASESB	0.0006	0.0096	0.0251	0.0010	0.0033	0.0049
0.999	OLSE	ASEVAR	0.0938	2.0446	5.8689	9.8479	11.2917	13.9890
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	LRRE	ASEVAR	0.0279	0.4019	1.1838	1.7077	1.9036	2.2727
		ASESB	0.0049	0.0802	0.2377	0.3502	0.3947	0.4730
	ME	ASEVAR	0.0966	2.1035	6.0232	27.2867	6.3359	8.7946
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.0574	0.4627	1.2555	18.0036	1.5887	1.8645
		ASESB	0.0045	0.1161	0.3361	0.0024	0.0363	0.0612
	JRME	ASEVAR	0.0908	1.3643	3.7526	26.0038	4.2453	5.4684
		ASESB	0.0004	0.0598	0.1822	0.0001	0.0179	0.0352
	LME	ASEVAR	0.0940	76.2581	1.42E+03	3.60E+09	6.91E+06	2.68E+07
		ASESB	0.1472	3.16E+03	3.06E+05	8.88E+13	2.38E+10	3.87E+11
		$\hat{d}_M^{\#}$	746	419	432	446	212	179
	$\text{LME}_{\text{d\#}}$	ASEVAR	0.0293	0.4190	1.1617	0.0324	0.1832	0.2643
		ASESB	0.0118	0.1814	0.5003	0.0146	0.0825	0.1128
	LRME	ASEVAR	0.0287	0.4257	1.2378	0.0310	0.1256	0.2074
		ASESB	0.0050	0.0846	0.2471	0.0059	0.0247	0.0409
0.9999	OLSE	ASEVAR	0.9260	20.3834	58.2605	94.9885	109.2081	140.0409
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

(*Continued on next page*)

Table 2
ASEVAR and ASESB of estimators (*Continued*)

			Without outlier			With one outlier		
ρ	$\hat{\beta}$		$\sigma^2 = 1$	$\sigma^2 = 25$	$\sigma^2 = 100$	$\sigma^2 = 1$	$\sigma^2 = 25$	$\sigma^2 = 100$
	LRRE	ASEVAR	0.1814	3.9007	11.3386	16.7707	18.9919	24.1932
		ASESB	0.0365	0.7811	2.2750	3.4200	3.9128	4.9548
	ME	ASEVAR	0.9524	20.9470	59.7002	377.4422	63.1850	101.3495
		ASESB	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	RME	ASEVAR	0.2360	4.0166	11.8750	230.5320	12.5682	21.1476
		ASESB	0.0514	1.0717	3.2486	0.0310	0.3707	0.6008
	JRME	ASEVAR	0.6636	12.4937	36.2461	344.7947	37.8859	62.7410
		ASESB	0.0237	0.6162	1.8237	0.0080	0.2034	0.3322
	LME	ASEVAR	14.1861	1.48E+03	1.95E+03	1.18E+11	1.60E+07	3.63E+09
		ASESB	237.0883	2.59E+05	1.06E+05	6.42E+15	2.67E+10	3.18E+14
		$\hat{d}_M^{\#}$	436	400	437	367	172	175
	$\text{LME}_{\text{d}\#}$	ASEVAR	0.1733	4.0012	10.9630	0.3091	1.9067	2.2597
		ASESB	0.0787	1.7132	4.7788	0.1396	0.8588	0.9917
	LRME	ASEVAR	0.1886	4.0459	11.9105	0.2551	1.2843	2.0431
		ASESB	0.0378	0.8065	2.3676	0.0495	0.2582	0.4066

$\hat{d}_M^{\#}$ denote the number of times $\hat{d}_M$ belongs to (0, 1) and the $\text{LME}_{\text{d}\#}$ denote the ASEVAR and ASESB of LME based on the values of those d's which satisfy the condition $\hat{d}_M \in (0, 1)$.

Let e_i^*, $i = 1, 2, \ldots, n$ be the absolute value of i^{th} residual obtained from the least squares fit. Next, arrange these values in increasing order of their magnitude such that $e_{(1)}^* \leq e_{(2)}^* \leq \ldots \leq e_{(n)}^*$. We introduce one outlier by multiplying actual value of Y corresponding to $e_{(n)}^*$ by twenty. On the similar line, two outliers are introduced in the response variable Y corresponding to $e_{(n)}^*$, $e_{(n-1)}^*$ and so on.

5.1. *Comparison of Estimators*

In this subsection, the above simulation experiment is replicated 1000 times for $\rho = 0.9$, 0.99, 0.999, 0.9999, sample size(n) = 30, 50, 100 and error variance $\left(\sigma^2\right) = 1$, 25, 100 respectively. For each combination of n, ρ, and σ^2, the average of sum of estimated variances (ASEVAR) and average of sum of estimated squared bias (ASESB) is obtained by replacing the unknown parameters by their suitable estimates in the corresponding MSE expression given in Section 2 and Section 3. For example, the ASEVAR and ASESB of LRME is obtained by using the expression given in Eq. (2.6) as

$$\text{ASEVAR} = \frac{1}{1000} \sum_{i=1}^{1000} \sum_{j=1}^{p} \frac{\left(\lambda_j + \hat{d}_j\right)^2}{\left(\lambda_j + 1\right)^2} \hat{A}^2 / \lambda_j \tag{5.3}$$

and

$$\text{ASESB} = \frac{1}{1000} \sum_{i=1}^{1000} \sum_{j=1}^{p} \frac{\left(\hat{d}_j - 1\right)^2}{\left(\lambda_j + 1\right)^2} \hat{\alpha}_{LRME_j}^2 \tag{5.4}$$

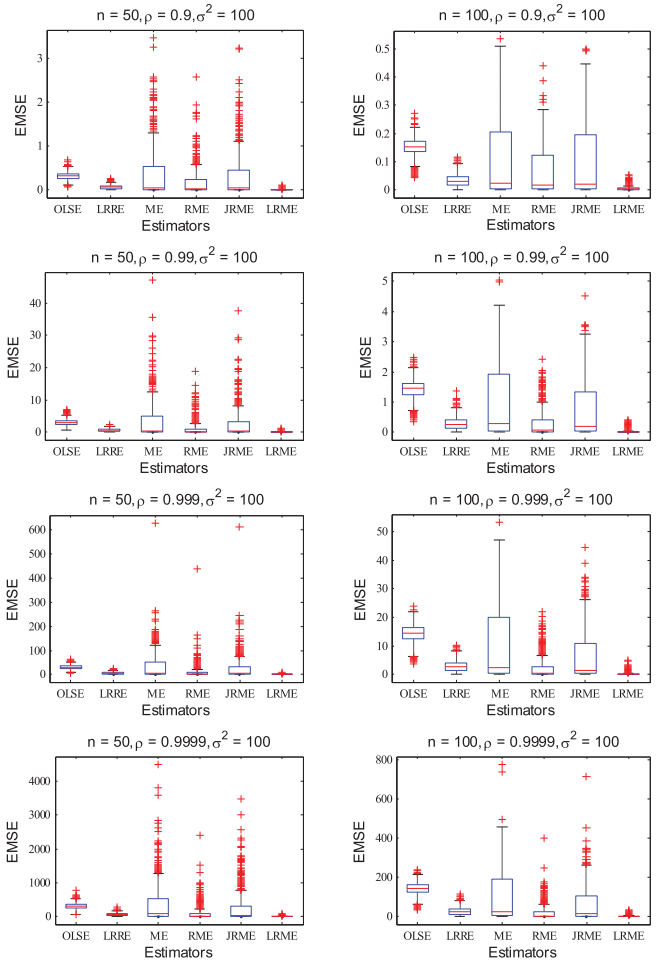


Figure 1. Box plots of OLSE, LRRE, ME, RME, JRME, and LRME with 1 outlier.

Using the same technique, we obtain the ASEVAR and ASESB of OLSE, LRRE, ME, RME, JRME and LME and are reported in Table 2. Note that, the LME is obtained without considering the range bound of the estimate of d.

Table 2 indicate that,

- For without and with one outlier case and for any combination of n and σ^2, as degree of multicollinearity (ρ) increases, the ASEVAR of each estimator goes on increases.
- For each replication with low degree of multicollinearity ($\rho = 0.9$) and smaller error variance ($\sigma^2 = 1$), the value of $\hat{d}_M$ lies in 0 and 1. But slight increase in ρ or σ^2 affects the $\hat{d}_M$ estimator and the frequency of $\hat{d}_M$ to lie in (0, 1) reduces.
- For without and with one outlier case, for any combination of ρ and σ^2, as sample size increases, the ASEVAR of the OLSE, LRRE, $\text{LME}_{\text{d\#}}$ and LRME goes on decreases. But for any combination of n and ρ, as error variance increases, ASEVAR of the OLSE, LRRE, $\text{LME}_{\text{d\#}}$ and LRME goes on increases.

Table 3
Descriptive statistics of EMSE's of estimators for $\sigma^2 = 100$

ρ	Estimator	Mean	Median	SD	QD
			$n = 50$		
0.9	OLSE	0.3207	0.3176	0.0877	0.0526
	LRRE	0.0671	0.0611	0.0446	0.0314
	ME	0.3240	0.0553	0.4527	0.2595
	RME	0.1586	0.0344	0.2479	0.1136
	JRME	0.2865	0.0497	0.4093	0.2231
	LME	5.36E+07	0.5074	8.70E+08	4.08E+03
	$LME_{d\#}$	0.0190	0.0070	0.0253	0.0109
	LRME	0.0092	0.0037	0.0155	0.0030
0.99	OLSE	3.0845	3.0578	0.9258	0.5896
	LRRE	0.5814	0.5039	0.4258	0.2926
	ME	3.1349	0.4052	4.6581	2.4810
	RME	0.8580	0.1124	1.6689	0.5206
	JRME	2.1862	0.2692	3.5392	1.6251
	LME	9.42E+11	17.5531	2.35E+13	6.85E+04
	$LME_{d\#}$	0.1392	0.0319	0.2315	0.0714
	LRME	0.0548	0.0137	0.1238	0.0163
0.999	OLSE	29.8001	29.5320	8.7529	5.4040
	LRRE	5.7969	5.1320	4.2649	2.9787
	ME	30.9913	6.7067	44.7174	24.9266
	RME	8.0859	1.0130	20.1042	4.4535
	JRME	20.8155	3.9466	35.8031	14.9738
	LME	5.06E+09	392.7401	6.21E+10	2.11E+06
	$LME_{d\#}$	1.5744	0.3573	2.3581	1.1895
	LRME	0.5464	0.1032	1.2958	0.1495
0.9999	OLSE	304.6856	300.4153	95.1655	57.9261
	LRRE	61.3254	53.8122	42.3041	28.7499
	ME	337.1994	87.1232	488.3154	258.6752
	RME	81.8823	10.1306	170.5768	43.4390
	JRME	219.6707	46.6097	361.7709	152.1898
	LME	1.48E+11	3.82E+03	1.33E+12	2.54E+07
	$LME_{d\#}$	12.9488	5.1100	17.3153	8.7172
	LRME	5.1813	1.1514	11.1360	1.7610
			$n = 100$		
0.9	OLSE	0.1519	0.1543	0.0326	0.0183
	LRRE	0.0338	0.0305	0.0218	0.0154
	ME	0.1043	0.0238	0.1154	0.1011
	RME	0.0636	0.0164	0.0726	0.0606
	JRME	0.0983	0.0226	0.1090	0.0954
	LME	8.13E+04	0.0344	1.20E+06	63.7344
	$LME_{d\#}$	0.0075	0.0023	0.0115	0.0025
	LRME	0.0062	0.0030	0.0088	0.0025

(*Continued on next page*)

Table 3
Descriptive statistics of EMSE's of estimators for $\sigma^2 = 100$ (*Continued*)

ρ	Estimator	Mean	Median	SD	QD
0.99	OLSE	1.4168	1.4483	0.3332	0.1793
	LRRE	0.2826	0.2472	0.2046	0.1405
	ME	0.9768	0.2914	1.0931	0.9431
	RME	0.2712	0.0644	0.3833	0.2008
	JRME	0.7038	0.2099	0.8284	0.6494
	LME	1.60E+07	3.9899	2.73E+08	7.72E+03
	$LME_{d\#}$	0.0629	0.0187	0.0949	0.0341
	LRME	0.0294	0.0080	0.0572	0.0091
0.999	OLSE	14.0989	14.3240	3.3405	1.9837
	LRRE	2.9123	2.6666	2.0161	1.4101
	ME	9.8044	2.2517	11.0366	9.7139
	RME	2.0669	0.3490	3.2743	1.3404
	JRME	6.1067	1.4803	7.6695	5.2710
	LME	1.56E+10	54.5738	4.25E+11	1.19E+05
	$LME_{d\#}$	0.3624	0.1324	0.6303	0.1544
	LRME	0.2591	0.0689	0.5157	0.0923
0.9999	OLSE	139.9973	142.6241	33.8215	19.5044
	LRRE	27.4107	24.4642	19.9365	13.8322
	ME	98.7998	25.2994	114.6741	92.8359
	RME	20.5262	3.3328	34.9397	12.5489
	JRME	60.3414	15.1822	78.5162	52.3301
	LME	5.45E+09	588.5886	9.89E+10	7.64E+05
	$LME_{d\#}$	4.4088	1.4234	7.2430	1.8673
	LRME	2.6433	0.7175	5.2797	0.9979

- The LRRE consistently shows smaller ASEVAR than that of the other estimators (except $LME_{d\#}$) for without outlier case with any combination of n, ρ and σ^2.
- For one outlier case, the ASEVAR and corresponding AMSE of the LRME is smaller than that of the other estimators (except $LME_{d\#}$). But, for larger error variance ($\sigma^2 = 25, 100$), the ASEVAR and corresponding average EMSE (AEMSE) of the LRME is smaller than that of the $LME_{d\#}$.

5.2. *Comparison of EMSE of Estimators*

The above simulation experiment is repeated 1000 times for one outlier case with $n = 50$ and 100, $\rho = 0.9$, 0.99, 0.999, and 0.9999 and $\sigma^2 = 100$. The EMSE's of the OLSE, LRRE, ME, RME, JRME, LME, and LRME are calculated by adding SEVAR and SESB of corresponding estimators. The box plots of 1000 EMSE's of the estimators are obtained and shown in the following Figure 1. Some of the EMSE values of the LME are too high, so the box plot of EMSE's of the LME is not shown in Figurey 1.

Figure 1 clearly shows that, the performance of the LRME is superior than the performance of the OLSE, LRRE, ME, RME and JRME for all values of n and ρ. It also reveals that, the LRME consistently shows smaller EMSE as compare to the other estimators.

Table 4
Ratio of AEMSE of estimators in the presence of multicollinearity and increasing number of outliers

	With One Outlier			With Two Outliers			With Three Outliers		
σ^2	1	25	100	1	25	100	1	25	100
					$\rho = 0.9$				
LRME/OLSE	0.0244	0.0306	0.0303	0.0196	0.0163	0.0107	0.0216	0.0161	0.0093
LRME/LRRE	0.1015	0.1332	0.1498	0.0886	0.0759	0.0558	0.0955	0.0751	0.0492
LRME/ME	0.0092	0.0354	0.0294	0.0169	0.0250	0.0123	0.0490	0.0305	0.0135
LRME/RME	0.0098	0.0470	0.0584	0.0180	0.0338	0.0234	0.0534	0.0407	0.0253
LRME/JRME	0.0092	0.0362	0.0330	0.0169	0.0257	0.0136	0.0492	0.0312	0.0150
LRME/LME	1.32E-11	3.58E-10	6.64E-12	1.18E-11	3.22E-10	3.78E-11	2.09E-10	2.07E-11	3.17E-11
$\hat{d}_M^{\#}$	715	397	159	562	311	93	517	279	111
LRME/ $\text{LME}_{\text{d\#}}$	1.2766	0.6855	0.5269	0.8676	0.7536	0.5606	0.7391	0.6386	0.4267
					$\rho = 0.99$				
LRME/OLSE	0.0088	0.0141	0.0171	0.0120	0.0088	0.0072	0.0169	0.0109	0.0068
LRME/LRRE	0.0430	0.0694	0.0870	0.0580	0.0445	0.0396	0.0786	0.0542	0.0372
LRME/ME	0.0039	0.0160	0.0156	0.0105	0.0123	0.0080	0.0248	0.0212	0.0099
LRME/RME	0.0051	0.0392	0.0522	0.0139	0.0266	0.0239	0.0333	0.0478	0.0263
LRME/JRME	0.0040	0.0194	0.0217	0.0108	0.0143	0.0107	0.0256	0.0250	0.0128
LRME/LME	2.68E-13	6.32E-13	4.44E-12	1.37E-14	3.22E-11	2.30E-13	3.30E-16	2.01E-11	5.27E-11
$\hat{d}_M^{\#}$	599	247	143	533	248	118	488	274	112
LRME/ $\text{LME}_{\text{d\#}}$	0.9234	0.6153	0.3716	0.7745	0.6172	0.4719	0.7425	0.5622	0.3530

(*Continued on next page*)

Table 4
Ratio of AEMSE of estimators in the presence of multicollinearity and increasing number of outliers (*Continued*)

	With One Outlier			With Two Outliers			With Three Outliers		
σ^2	1	25	100	1	25	100	1	25	100
					$\rho = 0.999$				
LRME/OLSE	0.0071	0.0125	0.0161	0.0113	0.0088	0.0066	0.0168	0.0100	0.0058
LRME/LRRE	0.0341	0.0630	0.0839	0.0538	0.0434	0.0348	0.0784	0.0522	0.0313
LRME/ME	0.0027	0.0147	0.0174	0.0084	0.0135	0.0063	0.0253	0.0192	0.0070
LRME/RME	0.0044	0.0465	0.0673	0.0132	0.0372	0.0198	0.0407	0.0540	0.0190
LRME/JRME	0.0029	0.0203	0.0261	0.0089	0.0176	0.0088	0.0272	0.0251	0.0092
LRME/LME	2.79E-16	1.13E-12	2.72E-13	2.16E-14	5.71E-12	1.82E-13	2.04E-13	2.01E-12	1.12E-11
$\hat{d}_M^{\#}$	508	210	130	501	251	132	515	245	97
LRME/ $\text{LME}_{\text{d\#}}$	0.9152	0.6084	0.4172	0.7348	0.6050	0.5110	0.7719	0.5177	0.3731
					$\rho = 0.9999$				
LRME/OLSE	0.0070	0.0113	0.0188	0.0113	0.0084	0.0063	0.0158	0.0108	0.0057
LRME/LRRE	0.0341	0.0577	0.0997	0.0542	0.0409	0.0332	0.0725	0.0544	0.0313
LRME/ME	0.0027	0.0108	0.0188	0.0082	0.0118	0.0066	0.0271	0.0191	0.0082
LRME/RME	0.0049	0.0350	0.0698	0.0160	0.0327	0.0208	0.0509	0.0499	0.0259
LRME/JRME	0.0031	0.0150	0.0279	0.0096	0.0158	0.0092	0.0306	0.0245	0.0114
LRME/LME	1.34E-14	4.49E-13	1.07E-11	2.34E-16	3.26E-13	7.60E-13	2.01E-14	4.72E-13	1.39E-12
$\hat{d}_M^{\#}$	486	196	135	532	254	121	492	287	97
LRME/ $\text{LME}_{\text{d\#}}$	0.8708	0.5133	0.3996	0.7722	0.6329	0.5522	0.7372	0.5934	0.3453

$\hat{d}_M^{\#}$ denote the number of times $\hat{d}_M$ belongs to (0, 1) and the $\text{LME}_{\text{d\#}}$ denote the AEMSE of LME based on the values of those d's which satisfy the condition $\hat{d}_M \in (0, 1)$.

In addition, the descriptive statistics like mean, median, standard deviation (SD) and quartile deviation (QD) of the 1000 EMSE's of the OLSE, LRRE, ME, RME, JRME, LME, and LRME are obtained and reported in the Table 3.

Table 3 clearly indicates that the mean, median, SD and QD of the LRME is consistently smaller than that of the other estimators for all combinations of n and ρ. Hence, the performance of the LRME is good as compare to other estimators.

5.3. *Comparison of EMSE for More Than One Outlier*

The same simulation experiment used in the subsections 5.1 and 5.2 is repeated 1000 times for all possible combinations of $\rho = 0.9, 0.99, 0.999, 0.9999$ and $\sigma^2 = 1, 25, 100$. For each combination of ρ and σ^2 with $n = 50$, the EMSE of the OLSE, LRRE, ME, RME, JRME, LME, and LRME is computed by introducing one, two and three outliers in the response variable. The AEMSE ratio of the LRME over the all remaining estimators is obtained and reported in Table 4.

From Table 4 it clearly seems that, for any degree of multicollinearity with different number of outliers, the AEMSE ratio of the LRME over the OLSE, LRRE, ME, RME, JRME and LME is less than one. Hence, the performance of the LRME is better as compare to the other estimators for more than one outlier case.

6. Conclusion

In this article, we have introduced a new estimator for regression parameters to deal with the problem of multicollinearity and outliers in the data. Some conditions are obtained theoretically to study the superiority of LRME over different estimators in the MSE sense. A Numerical example on real dataset is illustrated to support the superiority of proposed estimator. Also, an extensive simulation study is carried out to evaluate the performance of the LRME. It indicates that, the performance of the LRME is better than the OLSE, LRRE, ME, RME, JRME and LME when the multicollinearity and outliers simultaneously present in the data.

Acknowledgments

The authors are very grateful to the anonymous reviewers for so many valuable comments and constructive suggestions which resulted in the present version.

Funding

This work was partially supported by University Grants Commission, New Delhi, India under Major Research Project Scheme.

References

Akdeniz, F., Erol, H. (2003). Mean squared error matrix comparisons of some biased estimators in linear regression. *Communications in Statistics - Theory and Methods* 32(12):2389–2413.

Akdeniz, F., Kaciranlar, S. (1995). On the almost unbiased generalized Liu estimator and unbiased estimation of the bias and MSE, *Communications in Statistics—Theory and Methods* 24(7):1789–1797.

Alheety, M. I., Kibria, B. M. G. (2009). On the Liu and almost unbiased Liu estimators in the presence of multicollinearity with heteroscedastic or correlated errors. *Surveys in Mathematics and its Applications* 4:155–167.

Arslan, O., Billor, N. (2000). Robust Liu estimator for regression based on an M-estimator. *Journal of Applied Statistics* 27(1):39–47.

Birkes, D., Dodge, Y. (1993). *Alternative Methods of Regression*. New York: Wiley.

Gao, F., Liu, X. Q. (2011). Linearized ridge regression estimator under the mean square error criterion in a linear regression model. *Communications in Statistics - Simulation and Computation* 40:1434–1443.

Groß, J. (2003). *Linear Regression*. Lecture Notes in Statistics. Berlin: Springer Verlag.

Hampel, F. R., Ronchetti, E. M., Rousseeuvw, P. J., Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Function*. New York: Wiley.

Hocking, R. R., Speed, F. M., Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics* 18(4):425–437.

Hoerl, A. E., Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12:55–67.

Hoerl, A. E., Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics* 12:69–82.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35(1):73–101.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Huber, P. J., Ronchetti, E. M. (2009). *Robust Statistics*. 2nd ed. New York: Wiley.

Jadhav, N. H., Kashid, D. N. (2011). A jackknifed ridge M-estimator for regression model with multicollinearity and outliers. *Journal of Statistical Theory and Practice* 5(4):659–673.

Kaciranlar, S., Sakallioglu, S., Akdeniz, F., Styan, G. P. H., Werner, H. J. (1999). A new biased estimator in linear regression and detailed analysis of the widely-analysed dataset on Portland Cement. *Sankhya B* 61:443–459.

Liu, K. J. (1993). A new class of biased estimate in linear regression. *Communications in Statistics - Theory and Methods* 22:393–402.

Liu, X. Q., Gao, F. (2011). Linearized ridge regression estimator in linear regression. *Communications in Statistics—Theory and Methods* 40(12):2182–2192.

Maronna, R. A., Martin, D. R., Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. New York: Wiley.

McDonald, G. C., Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association* 70(350):407–416.

Montgomery, D. C., Peck, E. A., Vining, G. G. (2010). *Introduction to Linear Regression Analysis. 3rd* ed. New York: Wiley.

Myers, R. H. (1990). *Classical and Modern Regression with Applications. 2nd* ed. Boston: MA Duxbury.

Obenchain, R. L. (1975). Ridge analysis following a preliminary test of the shrunken hypothesis *Technometrics* 17(4):431–441.

Rousseeuw, P. J., Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Sakallioglu, S., Kaçiranlar, S. (2008). A new biased estimator based on ridge estimation. *Stat Papers* 49:669–689.

Silvapulle, M. J. (1991). Robust ridge regression based on an M-estimator. *Australian Journal of Statistics* 33:319–333.

Tiku, M. L., Akkaya, A. D. (2004). *Robust Estimation and Hypothesis Testing*. New Delhi: New Age International (P) Ltd.

Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 33

5-1-2016

Variable Selection in Regression using Multilayer Feedforward Network

Tejaswi S. Kamble
Shivaji University, Kolhapur, Maharashtra, India, tejustat@gmail.com

Dattatraya N. Kashid
Shivaji University, Kolhapur, Maharashtra, India., dnk_stats@unishivaji.ac.in

Variable Selection in Regression using Multilayer Feedforward Network

Cover Page Footnote
We thank the editor and anonymous referees for their valuable suggestions which led to the improvement of this article. First author would like to thank University Grant Commission, New Delhi, INDIA for financial support under Rajiv Gandhi National Fellowship scheme vide letter number F.14-2(SC)/2010(SA-III).

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 670-689.

ISSN 1538 − 9472

Variable Selection in Regression using Multilayer Feedforward Network

Tejaswi S. Kamble
Shivaji University
Kolhapur, Maharashtra, India

Dattatraya N. Kashid
Shivaji University
Kolhapur, Maharashtra, India

The selection of relevant variables in the model is one of the important problems in regression analysis. Recently, a few methods were developed based on a model free approach. A multilayer feedforward neural network model was proposed for developing variable selection in regression. A simulation study and real data were used for evaluating the performance of proposed method in the presence of outliers, and multicollinearity.

Keywords: Subset selection, artificial neural network, multilayer feedforward network, full network model and subset network model.

Introduction

The objective of regression analysis is to predict the future value of response variable for the given values of predictor variables. In the regression model, the inclusion of a large number of predictor variables leads to the problems such as i) decrease in prediction accuracy, and ii) increase in cost of the data collection (Miller, 2002). To improve the prediction accuracy of the regression model, one approach is to retain only a subset of relevant predictor variables in the model, and eliminate the irrelevant predictor variables. The problem of choosing an appropriate relevant set from a large number of predictor variables is called subset selection or variable selection in regression.

In traditional regression analysis, the form of the regression model must be first specified, then fitted to the data. However, if a pre-specified form of the model is itself wrong, another model must be used. Searching for a correct model for the given data becomes difficult when complexity is present in the data. A better alternative approach in the above situation would be to estimate a function or model from the data. Such an approach is called Statistical Learning; Artificial

Ms. Kamble is a Junior Research Fellow in the Department of Statistics. Email her at tejustat@gmail.com. Dr. Kashid is a Professor in the Department of Statistics. Email him at dnk_stats@unishivaji.ac.in.

Neural Network (ANN) and Support Vector Machine (SVM) are statistical learning techniques.

ANNs have recently received a great deal to attention in many fields of study, such as pattern reorganization, marketing research etc. ANN is important because of its potential use in prediction and classification problems. Usually, ANN is used for prediction when form of the regression model is not specified. In this article, ANN is used for selection of relevant predictor variables in the model.

Mallows's C_p (Mallows, 1973) and S_p statistics (Kashid and Kulkarni, 2002), along with other existing variable selection methods, are suitable under certain assumptions with prior knowledge about the data. When no prior knowledge about the data is available, ANN is an attractive variable selection method (Castellano and Fanelli, 2000), because ANN is a data-based approach. ANN is used in this study for obtaining predicted values of the subset regression model. The criteria C_p and S_p are based on prediction values of subset models. Therefore, we propose modification in C_p and S_p based on predicted values of the ANN model.

Mallows's C_p (Mallows, 1973) is defined by

$$C_p = \frac{RSS_p}{\sigma^2} + (n - 2p) \tag{1}$$

where p is the number of parameters in the subset regression model with $p-1$ regressors, RSS_p is the residual sum of squares of the subset model, n is the number of data points used for fitting the subset regression model, and σ^2 is replaced by its suitable estimates, usually based on the full model. In this study, the following cases are used.

Case 1

A simulation design proposed by McDonald and Galarneau (1975) is used for introducing multicollinearity in the regressor variables. It is given by

$$X_{ij} = \left(1 - \rho^2\right)^{\frac{1}{2}} Z_{ij} + \rho Z_{i(J+1)}, \; i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, J$$

where Z_{ij} are independent standard normal pseudo-random numbers of size n, and ρ^2 is the correlation between any two predictor variables. The response variable Y is generated by using the following regression model with $n = 30$ and $\rho = 0.999$:

$$Y_i = 1 + 4X_{i1} + 5X_{i2} + 0X_{i3} + \varepsilon_i,\ i = 1,2,...,30$$

where $\varepsilon_i \sim N(0,1)$. To identify the degree of multicollinearity, the variance inflation factor (VIF) is used (Montgomery, Peck, and Vining, 2006). For this data, the VIFs for the variables are 339.6, 572.5 and 350.1. These VIFs indicates the presence of severe multicollinearity in the data. We compute the value of the C_p statistic $C_p(M)$ and report the results in Table 1.

Case 2

Data generated in Case 1 is used, and one outlier is introduced by multiplying the actual Y corresponding to the maximum absolute residual by 25. The value of the response variable $Y = 8.2235$ is replaced by $Y = 205.5878$. The value of the C_p statistic $C_p(MO)$ is computed and reported in Table 1.

Case 3

The following nonlinear regression model is generated using the above $X_i, i = 1,2,3$ and ε_i which are generated in Case 1. The nonlinear regression model is

$$Y = \exp\left(1 + 4X_{i1} + 5X_{i2} + 0X_{i3}\right) + \varepsilon_i,\ i = 1,2,...,30$$

The values of the C_p statistic $C_p(NL)$ are computed for the nonlinear regression model and reported in Table 1.

Table 1. Values of $C_p(M)$, $C_p(MO)$, and $C_p(NL)$.

Regressors in subset model	P	$C_p(M)$	$C_p(MO)$	$C_p(NL)$
X_1	2	1.8617	3.0077	2.0726
X_2	2	2.2565	2.2510	1.0605
X_3	2	3.2585	1.9152	2.3498
X_1X_2	3	2.2237	2.8740	2.0059
X_1X_3	3	3.8518	3.2340	3.8492
X_2X_3	3	4.1730	3.4448	3.0179
$X_1X_2X_3$	4	4.0000	4.0000	4.0000

As seen in Table 1, the criterion C_p selects the wrong subset models for all the above-cited cases. The statistic fails to select the correct model in the presence of a) multicollinearity alone, b) both multicollinearity and outlier, and c)

nonlinear regression, because OLS estimation does not perform well in each case. Consequently, variable selection methods based on OLS estimator fail to select the correct model.

Regression Model and Neural Network Model

In general, the regression model is defined as

$$\mathbf{Y} = f\left(X, \boldsymbol{\beta}\right) + \boldsymbol{\varepsilon} \tag{2}$$

where f is any function of predictor variables $X_1, X_2, \ldots, X_{k-1}$ and unknown regression coefficients $\boldsymbol{\beta}$. If f is a non-linear function, then regression parameters are estimated by using nonlinear least squares method (or some other method). If f is linear, the regression model can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of response variables, $\mathbf{X}$ is a matrix of order $n \times k$ with 1's in the first column, $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression coefficients and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors which are independent and identically distributed $N(0,\sigma^2\mathbf{I})$. The least squares estimator of $\boldsymbol{\beta}$ is given by (Montgomery et al., 2006)

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

The predicted value of the regression model is obtained by the fitted equation

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

The prediction accuracy of the regression model depends on the selection of an appropriate model, which means the form of the function (f) must be specified before the regression analysis. If form of the model is not known, then one of the most appropriate alternative methods to handle this situation is artificial neural network.

Multilayer Feedforward Network (MFN)

The MFN can approximate any measurable function to any desired degree of accuracy (Hornik, Stinchcombe, and White, 1989). This MFN model consists of an input layer, an output layer, and one or more hidden layer(s). We represent the architecture of MFN with one hidden layer consisting of J hidden nodes, and a single node in an output layer, as shown in Figure 1. A vector $\mathbf{X} = [X_0, X_1, \ldots, X_{k-1}]'$ is the vector of k units in the input layer and $\mathbf{Y}$ is the output of the network.

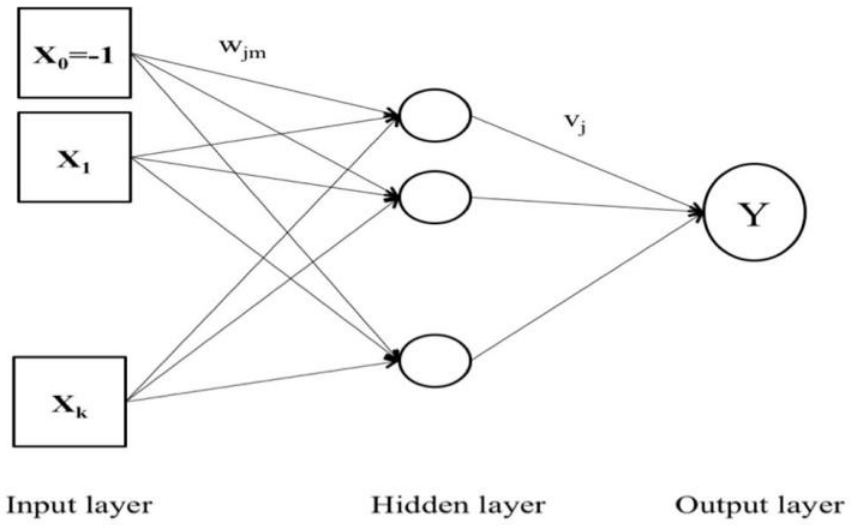


Figure 1. Multilayer feedforward network

From Figure 1, each input signal is connected to each node in the hidden layer with weight w_{jm}, $m = 0,1,2,3,\ldots,k-1$, $j = 1,2,\ldots,J$, and hidden nodes are connected to a node in the output layer with weight v_j, $j = 1,2,\ldots,J$. The final output Y_i for the i^{th} data point is given by

$$Y_i = g_2\left(\sum_{j=1}^{J} V_j g_1\left(\sum_{m=0}^{k-1} w_{jm} X_{im}\right)\right) \quad i = 1,2,\ldots,n$$

where g_1 and g_2 denote activation functions used in the hidden layer and output layer respectively; it is not necessary that g_1 and g_2 are the same activation functions. The above network model can be written as

$$\mathbf{Y} = f\left(X, \boldsymbol{\beta}\right) \tag{4}$$

where $\boldsymbol{\beta} = (v_1, \ldots, v_J, \boldsymbol{w}_0, \boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{k-1})$, $\boldsymbol{w}_m = (w_{1m}, w_{2m}, \ldots, w_{Jm})$, $m = 0,1,2,\ldots,k-1$ and $f(X,\boldsymbol{\beta})$ is a nonlinear function of the inputs $X_0, X_1, X_2, \ldots, X_{k-1}$ and the weight vector $\boldsymbol{\beta}$. If we add an error term in the above model (4), then it becomes a regression model as in Equation 2, where $\boldsymbol{\varepsilon}$ is the random error.

The next step in ANN modeling is training the network. The purpose of training the network is to obtain weights in a neural network model using the training data. Various training methods or algorithms are available in the literature. The robust back-propagation method (see Kasko, 1992) is one such. First, two types of MFN models must be defined, namely the full MFN model and the subset MFN model, for proposing modification in C_p and S_p statistics.

Full MFN and subset MFN model

A full MFN model is constructed with input units $X_1, X_2, \ldots, X_{k-1}$ and bias node $X_0 = -1$. The MFN model in Equation 4 is a full MFN model. The network weights are obtained by training the network and the network output vector based on a full MFN model, as

$$\hat{\mathbf{Y}} = f\left(X, \hat{\boldsymbol{\beta}}\right) \tag{5}$$

where $\hat{\boldsymbol{\beta}}$ is the estimated weight vector.

A subset MFN model is constructed with a subset of input units $X_A = (X_0, X_1, X_2, \ldots, X_{p-1})'$ of size $p(p \leq k)$ in the input layer. The subset network model is given by

$$\mathbf{Y} = f\left(X_A, \boldsymbol{\beta}_A\right) \tag{6}$$

where X and $\boldsymbol{\beta}$ are partitioned as $X = [X_A : X_B]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_A : \boldsymbol{\beta}_B]$. Similarly, the network output vector based on subset MFN model is

$$\hat{\mathbf{Y}} = f\left(X_A, \hat{\boldsymbol{\beta}}_A\right) \tag{7}$$

where $\hat{\boldsymbol{\beta}}_A$ is the estimated weight vector.

To implement the training procedure using network training algorithm, we need to select the number of hidden layers in the MFN and the number of hidden nodes in that hidden layer. This is discussed in the next section.

Selection of Hidden Layer and Hidden Nodes

The selection of learning rate parameter, initial weights and number of hidden layers in the MFN model and the number of hidden nodes in each hidden layer is an important task. The number of hidden layers is determined first. The network begins as a one-hidden-layer network (Lawrence, 1994). If the one-hidden-layer MFN network does not sufficient for training the network, then more hidden layers are added. In the MFN model, theoretically a single hidden layer is sufficient, because any continuous function defined on a compact set in R^n can be approximated by a multilayer ANN with one hidden layer with sigmoid activation function (Cybenko, 1989). Based on this result, we consider the single hidden layer MFN model with sigmoid activation function.

The choice of number of hidden neurons in the hidden layer is also a considerable problem, and it depends on the data. Research has proposed various methods for selection of hidden nodes in the hidden layer (see Chang-Xue, Zhi-Guang and Kusiak, 2005), as follows:

- $H_1 = 2I + 1$ (Hecht-Nelson, 1987)
- $H_2 = (I + O)/2$ (Lawrence and Fredrickson, 1998)
- $n/10 - I - O \leq H_3 \leq n/2 - I - O$ (Lawrence and Fredrickson, 1998)
- $H_4 = I\log_2 n$ (Marchandani and Cao, 1989)
- $H_5 = O(I + 1)$ (Lipmann, 1987)

Here, I is the number of inputs, O is the number of output neurons, and n is the number of training data points.

Variable Selection Methods and Proposed Methods

In the classical linear regression, several variable selection procedures have been suggested by the researchers. Most methods are based on least squares (LS) parameter estimation procedure. The variable selection methods based on LS estimates of $\boldsymbol{\beta}$ fail to select the correct subset model in the presence of outlier, multicollinearity, or nonlinear relationship between $\mathbf{Y}$ and X. Here, we modified existing subset selection methods using MFN model for prediction.

It is demonstrated that the Mallows's C_p statistic does not work well when assumptions are violated. Researchers have suggested some other methods for variable selection (see Ronchetti and Staudte, 1994; Sommer and Huggins, 1996). Also Kashid and Kulkarni (2002) have suggested a more general criterion, the S_p statistic for variable selection in cases of clean and outlier data. It can be defined as

$$S_p = \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2}{\sigma^2} + \left(k - 2p\right) \tag{8}$$

where $\hat{Y}_{ik}$ is the predicted value of the full model, $\hat{Y}_{ip}$ is the predicted value of the subset model based on M-estimator of the regression parameters, and k and p are the number of parameters in the full and subset model respectively. The σ^2 is replaced by its suitable estimates, which usually consists of the full model.
The subset selection procedure is same for both the methods. The S_p statistic is equivalent to the C_p statistic when LS method is used for estimating regression coefficients. The following suggests modification in both criteria using the complicity measure.

MC$_p$ and *MS*$_p$ Criteria

In a modified version of the C_p and S_p statistics, the network output (estimated values of response **Y**) is obtained by using the single hidden layer with a single output MFN model.

The network outputs $\hat{Y}_{ik} = f\left(\mathbf{X}_i, \hat{\boldsymbol{\beta}}\right)$ and $\hat{Y}_{ip} = f\left(\mathbf{X}_{iA}, \hat{\boldsymbol{\beta}}_A\right)$ denote outputs based on full MFN and subset MFN model, respectively. The residual sum of squares for the full and subset network models are defined as

$$RSS_k = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_{ik}\right)^2, \text{ and}$$
$$RSS_p = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_{ip}\right)^2$$

The modified version of C_p and S_p are denoted as MC_p and MS_p. They are defined by

$$MC_p = \frac{RSS_p}{\sigma^2} + C(n, p), \text{ and} \tag{9}$$

$$MS_p = \frac{\sum_{i=1}^{n} \left(\hat{Y}_{ik} - \hat{Y}_{ip} \right)^2}{\sigma^2} + C(n, p) \tag{10}$$

where n is the number of data points and p is the number of inputs including bias node (X_o). $\hat{Y}_{ik}$ and $\hat{Y}_{ip}$ are the predicted values of Y based on the full and subset MFN models, respectively, $C(n,p)$ is the penalty term, and σ^2 is replaced by its suitable estimate if it is unknown. The motivation for proposing modified versions of C_p and S_p are as follows.

In criterion MC_p, we use two types of measures. The first term measures the discrepancy between the desired output and network output based on the subset MFN model. The smaller this value is, the closer to the desired output it is; the smallest value of this measure is smallest for the full model. Therefore, it is difficult to select the correct model by minimizing criterion. So, we add a complicity measure called the penalty function, comprised of only p, only n, or both n and p.

In the second criterion MS_p, we use sum of squared difference between network output of the full and subset MFN models. The smallest value indicates that a prediction based on the subset MFN model is as accurate as the full MFN model. When full MFN model is itself the correct model, this value is zero. It is difficult to select the correct model using the minimizing criterion. Therefore we added the penalty function similar to criterion defined in (9) and used the same logic for the selection of subset. The selection procedure for both methods is as follows.

Step I: Compute the MC_p for all possible subsets.
Step II: Select the subset corresponding to the minimum value of MC_p. Use the same procedure for MS_p.

Choice of Estimator of σ^2

An estimator of σ^2 is required to implement the MC_p and MS_p criteria. In the literature of regression, various estimators of σ^2 are available. What follows are estimators of σ^2 used in MC_p and MS_p based on full network output, and a study of the effect of these estimators on the value of MC_p and MS_p.

1. $\hat{\sigma}_1^2 = \dfrac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_{ik}\right)^2}{n-k}$

2. $\hat{\sigma}_2^2 = \left(1.4826\text{median}\left|r_i - \text{median}\left(r_i\right)\right|\right)^2$

3. $\hat{\sigma}_3^2 = \left(1.4826\text{median}\left|r_i\right|\right)^2$

where n is the number of data points, k is the number of inputs in the full MFN model including bias node $r_i = Y_i - \hat{Y}_{ik}$, and $\hat{Y}_{ik}$ is the network output for the i^{th} data point based on the full MFN model.

Performances of MC_p and MS_p

To evaluate the performance of MC_p and MS_p, we have used single hidden layer MFN model and robust back-propagation training method with sigmoid activation function in the hidden layer and output layer. In robust back-propagation, we use an error suppressor function $s(e)$ by replacing the scalar squared error e (Kasko, 1992), because $s(e) = e^2$ is not robust. The following error suppressor functions are used in this study.

1. $E_1 = s(e) = \max(-c, \min(c,e))$ (Huber function)
   (where c = 1.345 is bending constant)

2. $E_2 = s(e) = 2e/(1+e^2)$ (Cauchy function)

3. $E_2 = s(e) = \tanh(e/2)$ (Hyperbolic tangent function)

The learning rate parameter (η) is selected by trial and error, and the number of hidden nodes in hidden layer is selected using the selection methods given earlier. The following seven penalty functions are used for computing MS_p and MC_p; some are available in the literature (Sakate and Kashid, 2014).

1. $P_1 = 2p$

2. $P_2 = p\log(n+2)$

3. $P_3 = 2p + \dfrac{2(p+1)(p+2)}{n-p-2}$

4. $P_4 = p(\log n + 1)$

5. $P_5 = \dfrac{2pn}{n-p-1}$

6. $P_6 = 2p + \dfrac{2p(p+1)}{n-p-1}$

7. $P_7 = p\log n$

The performance of the proposed methods is measured for different combinations of penalty functions $(P_l)\, l = 1,2,\ldots,7$, selection methods of hidden nodes in the hidden layer $(H_m)\, m = 1,2,\ldots,5$, and error suppressor functions $(E_o)\, o = 1,2,3$; these are denoted by (P_l, H_m, E_o). Three simulation designs are used for the evaluation of the performance of MS_p and MC_p.

Simulation Design A

The performance of proposed modified versions of $S_p(MS_p)$ and $C_p(MC_p)$ are evaluated using the following models with two error distributions.

Model I: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, where $\boldsymbol{\beta} = (1,5,10,0)$,

Model II: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$, where $\boldsymbol{\beta} = (1,5,10,0,0)$

The regressor variables were generated from $U(0,1)$ and the error term was generated from $N(0,1)$ and Laplace (0,1). The response variable Y was generated using Models I and II for sample sizes 20 and 30, respectively. This experiment is repeated 100 times and ability of these methods to select the correct model is measured using learning parameter $(\eta) = 0.1$ and $\hat{\sigma}_1^2$. The results are reported in Tables 2 through 5.

Table 2. Model selection ability of MS_p and MC_p in 100 replications for Model I of size 20

Error distribution	Error suppressor function		H_1		H_2		H_3		H_4		H_5	
		P_n	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p
Normal	Huber	P_1	79	66	84	77	72	75	73	64	77	71
		P_2	86	81	92	82	81	87	84	77	87	84
		P_3	88	86	94	90	90	92	89	86	93	89
		P_4	88	85	94	88	88	90	87	81	90	87
		P_5	86	81	92	85	82	87	85	79	88	85
		P_6	86	81	92	85	82	87	85	79	88	85
		P_7	85	79	92	82	79	87	82	77	87	84
	Cauchy	P_1	78	58	77	32	76	52	67	57	63	69
		P_2	91	71	85	35	83	72	79	68	80	76
		P_3	93	79	85	34	86	77	87	80	84	83
		P_4	92	74	85	36	84	77	84	74	83	81
		P_5	91	71	85	36	83	72	79	69	82	76
		P_6	91	71	85	36	83	72	79	69	82	76
		P_7	91	70	85	35	82	72	79	66	79	75
	Hyperbolic Tangent	P_1	79	66	74	77	75	79	75	79	77	83
		P_2	86	81	86	84	85	87	85	87	86	91
		P_3	88	86	91	89	87	90	87	90	92	91
		P_4	88	85	88	86	86	89	86	89	89	91
		P_5	86	81	86	84	85	88	85	88	87	91
		P_6	86	81	86	84	85	88	85	88	87	91
		P_7	85	79	85	84	85	87	85	87	85	91
Laplace	Huber	P_1	69	67	75	66	75	69	77	34	78	66
		P_2	83	81	86	80	87	73	89	36	79	79
		P_3	86	86	91	84	89	80	94	35	80	81
		P_4	87	83	88	82	89	76	93	36	81	81
		P_5	84	81	86	80	87	73	91	36	80	79
		P_6	84	81	86	80	87	73	91	36	80	79
		P_7	81	81	86	77	85	73	88	35	79	79
	Cauchy	P_1	74	54	77	52	68	67	70	51	71	62
		P_2	83	75	81	60	80	77	80	66	78	74
		P_3	86	85	86	67	84	80	85	76	80	81
		P_4	86	84	84	65	82	79	84	72	79	78
		P_5	84	77	82	60	80	77	82	67	78	74
		P_6	84	77	82	60	80	77	82	67	78	74
		P_7	83	74	80	60	79	77	79	65	75	73
	Hyperbolic Tangent	P_1	70	67	76	69	85	76	85	76	82	63
		P_2	83	81	82	82	90	85	90	85	88	75
		P_3	86	86	87	88	92	89	92	89	93	75
		P_4	87	84	86	87	92	88	92	88	93	78
		P_5	84	81	83	83	90	85	90	85	88	76
		P_6	84	81	83	83	90	85	90	85	88	76
		P_7	82	81	82	82	90	84	90	84	87	74

Table 3. Model selection ability of MS_p and MC_p in 100 replications for Model I of size 30

Error distribution	Error suppressor function		H_1		H_2		H_3		H_4		H_5	
		P_n	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p
Normal	Huber	P_1	78	72	78	74	71	69	76	62	74	72
		P_2	89	81	89	88	83	85	90	74	90	92
		P_3	93	87	92	92	92	87	94	96	92	94
		P_4	88	77	84	84	78	82	92	72	85	80
		P_5	87	77	82	82	77	79	92	66	80	79
		P_6	87	77	82	82	77	79	92	66	80	78
		P_7	89	81	88	88	83	85	90	74	88	92
	Cauchy	P_1	72	59	74	71	77	59	76	52	70	50
		P_2	85	73	81	88	84	74	86	68	86	76
		P_3	94	82	87	93	88	81	94	80	94	80
		P_4	80	66	83	83	83	69	84	62	80	68
		P_5	79	65	82	79	81	68	84	60	80	66
		P_6	79	65	82	79	81	68	84	61	80	66
		P_7	84	73	81	88	84	74	86	68	86	68
	Hyperbolic Tangent	P_1	83	74	82	71	78	74	74	62	78	76
		P_2	89	82	93	88	92	87	82	72	90	88
		P_3	94	87	96	92	94	91	86	68	96	92
		P_4	85	81	91	81	88	83	86	72	84	83
		P_5	85	81	88	79	86	82	82	70	85	82
		P_6	85	81	88	79	86	82	82	71	84	82
		P_7	88	92	93	88	91	86	82	74	90	86
Laplace	Huber	P_1	73	56	77	70	72	54	80	58	78	62
		P_2	82	75	91	85	91	80	80	78	88	80
		P_3	89	81	92	87	90	84	86	86	90	86
		P_4	82	70	85	81	82	75	81	70	90	76
		P_5	81	66	84	77	82	72	81	64	91	72
		P_6	81	66	84	77	82	73	81	65	84	72
		P_7	82	74	91	85	88	80	80	72	88	80
	Cauchy	P_1	62	33	74	47	77	66	76	56	77	60
		P_2	78	43	83	66	86	78	86	66	85	76
		P_3	87	58	87	73	90	80	92	80	87	84
		P_4	75	40	81	58	84	77	80	62	84	70
		P_5	73	38	80	56	82	75	78	62	84	66
		P_6	73	38	80	56	82	75	78	62	84	66
		P_7	77	43	83	64	86	78	86	66	84	74
	Hyperbolic Tangent	P_1	72	77	72	71	78	68	78	60	82	50
		P_2	85	90	89	84	85	86	82	78	96	76
		P_3	88	93	91	89	90	88	86	86	97	84
		P_4	82	87	84	83	84	83	78	78	94	70
		P_5	82	86	83	80	82	80	78	78	94	62
		P_6	82	86	83	80	82	80	78	78	94	62
		P_7	84	90	89	84	85	87	80	80	98	76

Table 4. Model selection ability of MS_p and MC_p in 100 replications for Model II of size 20

Error distribution	Error suppressor function		H_1		H_2		H_3		H_4		H_5	
		P_n	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p
Normal	Huber	P_1	60	33	60	43	62	50	62	38	68	60
		P_2	79	53	77	59	72	72	76	60	74	72
		P_3	85	68	83	78	82	82	85	72	78	85
		P_4	82	64	83	65	83	78	80	78	76	80
		P_5	80	57	79	60	72	74	76	64	74	76
		P_6	80	57	79	60	72	74	76	64	74	76
		P_7	77	53	76	59	72	70	76	58	74	72
	Cauchy	P_1	54	40	51	24	60	22	48	32	60	43
		P_2	68	40	72	46	70	38	76	49	70	56
		P_3	72	43	80	68	82	50	80	56	76	65
		P_4	71	45	75	64	80	46	80	52	76	63
		P_5	69	51	73	46	70	38	78	49	78	58
		P_6	69	63	73	46	70	38	78	49	78	58
		P_7	66	50	71	42	68	38	74	49	70	56
	Hyperbolic Tangent	P_1	63	42	69	60	50	50	61	44	68	70
		P_2	74	72	78	72	68	74	88	65	84	84
		P_3	82	85	82	78	74	82	88	78	94	86
		P_4	79	83	82	74	74	78	88	78	90	86
		P_5	75	76	78	74	70	78	88	78	89	85
		P_6	75	76	79	74	70	76	88	68	88	84
		P_7	72	70	79	74	66	70	89	68	80	84
Laplace	Huber	P_1	40	44	54	32	56	35	68	48	41	40
		P_2	62	58	68	52	67	56	76	72	62	60
		P_3	76	66	88	78	74	75	74	65	70	74
		P_4	70	65	72	63	76	73	82	76	64	70
		P_5	65	59	68	52	66	60	76	72	60	60
		P_6	65	59	68	52	66	60	76	72	61	60
		P_7	58	58	67	50	66	54	76	70	60	56
	Cauchy	P_1	59	29	50	32	52	32	44	22	44	49
		P_2	61	40	64	48	74	50	56	45	64	62
		P_3	64	53	65	56	78	60	58	53	73	72
		P_4	65	50	64	52	76	58	56	52	67	68
		P_5	64	43	65	48	74	50	56	48	64	64
		P_6	64	43	65	48	75	50	56	48	64	64
		P_7	61	40	62	44	75	46	54	43	62	58
	Hyperbolic Tangent	P_1	54	44	58	44	56	35	52	38	60	60
		P_2	78	60	78	70	67	57	60	53	74	72
		P_3	74	66	84	76	74	74	61	56	87	81
		P_4	74	66	83	76	78	76	62	54	83	80
		P_5	72	60	78	70	66	60	61	52	74	74
		P_6	72	60	78	70	66	60	61	52	74	74
		P_7	70	60	78	78	66	54	61	50	72	76

Table 5. Model selection ability of MS_p and MC_p in 100 replications for Model II of size 30

Error distribution	Error suppressor function		H_1		H_2		H_3		H_4		H_5	
		P_n	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p
Normal	Huber	P_1	69	36	64	55	64	30	72	46	66	46
		P_2	82	77	83	64	76	60	84	70	84	66
		P_3	83	87	86	73	78	80	86	76	84	88
		P_4	80	66	80	63	76	43	82	64	80	64
		P_5	78	85	72	60	74	40	78	60	78	62
		P_6	78	58	72	61	74	39	78	60	77	62
		P_7	83	77	82	64	75	60	84	70	80	66
	Cauchy	P_1	45	25	51	44	52	30	52	23	44	34
		P_2	68	58	65	68	71	60	72	40	62	52
		P_3	79	68	74	74	78	66	79	58	78	62
		P_4	56	51	64	64	68	44	66	32	54	42
		P_5	57	38	64	64	66	45	65	30	46	42
		P_6	57	38	64	64	66	44	64	30	46	42
		P_7	66	54	64	68	70	58	65	40	62	52
	Hyperbolic Tangent	P_1	68	36	70	57	52	53	72	44	56	35
		P_2	82	76	80	78	70	69	84	72	76	62
		P_3	82	86	80	86	80	82	86	76	86	80
		P_4	80	66	78	72	70	74	81	64	68	52
		P_5	76	60	76	68	66	69	80	62	68	48
		P_6	76	60	76	69	66	69	79	62	68	48
		P_7	82	76	81	76	70	69	84	70	32	63
Laplace	Huber	P_1	56	36	54	48	52	56	48	52	52	36
		P_2	86	50	72	70	74	84	70	74	76	70
		P_3	92	54	78	74	84	92	74	80	84	70
		P_4	74	46	66	64	69	80	66	72	70	50
		P_5	74	46	64	64	62	70	64	72	66	46
		P_6	74	46	63	64	62	70	64	72	66	46
		P_7	86	50	72	68	74	84	68	74	76	70
	Cauchy	P_1	32	36	60	24	50	34	40	21	36	21
		P_2	52	60	80	42	60	62	74	45	56	48
		P_3	64	74	86	48	74	70	84	56	64	60
		P_4	40	54	68	32	52	54	62	32	45	36
		P_5	40	52	66	30	50	48	56	28	42	32
		P_6	40	52	66	31	50	48	56	28	42	33
		P_7	48	60	80	40	61	62	72	42	42	42
	Hyperbolic Tangent	P_1	66	44	52	46	50	81	60	46	52	36
		P_2	80	72	80	66	72	68	81	70	79	64
		P_3	84	80	84	79	76	80	86	79	86	82
		P_4	74	66	71	62	74	68	81	66	60	56
		P_5	72	30	64	56	72	68	75	62	60	48
		P_6	72	61	64	56	72	68	76	62	60	48
		P_7	80	70	76	66	72	68	83	70	74	74

From Tables 2 through 5, it can be observed that the overall performance of the MS_p statistic is better than the MC_p statistic. The performance of penalties P_2 through P_7 is better than penalty P_1, with H_1 through H_5, for Models I and II. Based on these simulations, it is recommended that any hidden node selection method be used with penalty P_2 through P_7 and Huber or Hyperbolic Tangent error suppressor function.

Simulation Design B

The experiment was repeated 100 times using the simulation design A. The performance of MS_p and MC_p were compared with Mallows's C_p for Models I and II with sample sizes of 20 and 30. MS_p and MC_p were computed using (P_3,H_1,E_1), and learning parameters (η) = 0.1 and $\hat{\sigma}_1^2$. The results are reported in Table 6.

Table 6. Model selection ability of correct model for 100 repetitions

Error Distribution	Sample sizes	Model I			Model II		
		MS_p	MC_p	C_p	MS_p	MC_p	C_p
Normal	20	94	90	82	83	78	76
	30	92	92	79	86	73	70
Laplace	20	91	84	81	88	78	77
	30	92	87	84	78	74	75

From Table 6, it is clear that the model selection ability of MS_p and MC_p is better than C_p (based on LS estimates) for sample sizes 20 and 30 for both error distributions. The model selection ability of MS_p is uniformly larger than that of MC_p or C_p.

Simulation Design C

Three further models based on MFN are used to evaluate the performance of MS_p and MC_p:

Model III: $$Y = \sqrt{\beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2 + \beta_3 X_3^2 + \beta_4 X_4^2} + \varepsilon,$$

Model IV: $$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2 + \beta_3 X_3^2 + \beta_4 X_4^2 + \varepsilon,$$

Model V: $$Y = e^{\beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2 + \beta_3 X_3^2 + \beta_4 X_4^2} + \varepsilon,$$

where $\boldsymbol{\beta}$ = (1,5,10,0,0).

In this simulation, X_i = (i = 1,2,3,4) were generated from U(0,1) and error was generated from N(0,1) and Laplace(0,1). The response variable Y was generated using Models III, IV and V. MS_p and MC_p were computed using (P_1 –

P_7,H_1,E_1), learning parameters (η) = 0.1 and $\hat{\sigma}_1^2$. The ability of these methods to select the correct model over 100 replications is reported in Table 7.

Table 7. Correct model selection ability over 100 replications

Error distribution	P_n	Model III				Model IV				Model V			
		n = 20		*n = 30*		*n = 20*		*n = 30*		*n = 20*		*n = 30*	
		MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p
Normal	P_1	50	40	78	25	71	57	89	65	04	07	72	76
	P_2	55	35	89	48	78	70	91	73	05	06	90	91
	P_3	55	24	93	58	83	78	88	60	04	07	90	95
	P_4	60	38	80	34	80	76	82	56	05	07	91	85
	P_5	54	37	77	32	79	72	83	56	05	07	83	82
	P_6	55	40	77	35	79	72	85	65	05	06	89	82
	P_7	54	34	90	42	76	69	90	70	05	06	75	90
Laplace	P_1	20	16	60	40	15	16	89	70	07	05	89	19
	P_2	21	14	80	66	12	14	93	80	07	04	99	18
	P_3	25	15	86	80	7	11	82	65	06	04	100	13
	P_4	22	14	75	56	12	15	80	52	05	03	96	10
	P_5	20	14	75	50	13	16	80	52	05	04	90	16
	P_6	20	15	75	50	13	16	90	70	08	05	90	16
	P_7	18	14	80	64	13	14	91	72	04	06	99	14

From Table 7, it is clear that performance of MS_p is better than MC_p for all models and sample size 30. The performance of both criteria MS_p and MC_p is very poor for all models when error distribution is Laplace for small samples: the sample size must be moderate to large for selection of relevant variables when regression model is nonlinear.

Performance of MC_p and MS_p in the presence of multicolinearity and outlier

The performance of MS_p and MC_p is studied using the Hald data (Montgomery et. al, 2006). The variance inflation factors (VIF) corresponding to each term are 38.5, 254.4, 46.9, and 282.5. The VIF values indicate that multicollinearity exists in the data. Consider the following cases:

Case I: Data with multicolinearity (original data)

Case II: Data with multicolinearity and single outlier ($Y_6 = 109.2$ is replaced by 150)

Case III: Data with multicolinearity and two outliers ($Y_2 = 73.4$ and $Y_6 = 109.2$ are replaced by 150 and 200 respectively)

MS_p and MC_p was computed for all possible subset models with different penalty functions and estimators of σ^2. The selected subset model, by various combinations of $(P_l, \hat{\sigma}_s^2)$, $l = 1,2,...,7$, $s = 1,2,3$ is reported in Table 8. For training the network, the simulation employs the Huber error suppressor function, number of hidden neurons H_1, and learning parameter $(\eta) = 0.1$. The results are reported in Table 8.

Table 8. Selected subset by MS_p and MC_p for Cases I – III

		Case I			Case II			Case III		
Statistic	P_n	σ_1^2	σ_2^2	σ_3^2	σ_1^2	σ_2^2	σ_3^2	σ_1^2	σ_2^2	σ_3^2
	P_1	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2
	P_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2
	P_3	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2
MS_p	P_4	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_2	x_1x_2	x_1x_2
	P_5	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_2	x_1x_2	x_1x_2
	P_6	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_2	x_1x_2	x_1x_2
	P_7	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_1x_2	x_2	x_1x_2	x_1x_2
	P_1	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_2	x_1x_4	x_1x_4
	P_2	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_2	x_1x_4	x_1x_4
	P_3	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_2	x_1x_4	x_1x_4
MC_p	P_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_2	x_1x_4	x_1x_4
	P_5	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_2	x_1x_4	x_1x_4
	P_6	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_2	x_1x_4	x_1x_4
	P_7	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_1x_4	x_2	x_1x_4	x_1x_4

This data is analyzed in the connection of multicolinearity and outlier (see Ronchetti and Staudte, 1994; Sommer and Huggins, 1996; and Kashid and Kulkarni, 2002). They have suggested $\{X_1, X_2\}$ is the best subset model for clean data and outlier data. The MS_p statistic selects the same subset model for all combinations of $(P_l, \hat{\sigma}_s^2)$, $l = 1,2,...,7$, $s = 1,2,3$, for Case I and II. In Case III, MS_p fails to select correct model for penalty $P_4 - P_7$ with $\hat{\sigma}_1^2$. Conclusion: the MS_p statistic performs better than MC_p for all cases with all penalty functions and estimators of σ^2, excluding few cases.

Conclusion

The proposed modified methods are model-free. It is clear that the performance of proposed MS_p statistic is better than classical regression methods in the presence of multicollinearity, outlier, or both simultaneously. The MS_p statistic selects the correct model in cases of nonlinear model for moderate to large sample sizes. From the simulation study, it can be observed that MFN is useful when there is no idea about the functional relationship between response and predictor variables. The MS_p statistic is also useful for selection of inputs from a large set of inputs in a network model, in order to find which network output is closest to the desired output.

Acknowledgements

This research was partially funded by the University Grant Commission, New Delhi, India, under the Rajiv Gandhi National Fellowship scheme vide letter number F.14-2(SC)/2010(SA-III).

References

Chang-Xue, J. F., Zhi-Guang, Yu. and Kusiak, A. (2006) Selection and validation of predictive regression and neural network models based on designed experiment. *IIE Transactions*, *38*(1), 13-23. doi: 10.1080/07408170500346378

Cybenko, G. (1989) Approximation by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, *2*(4), 303-314. doi: 10.1007/BF02551274

Castellano G. and Fanelli A. M. (2000) Variable selection using neural network models. *Neurocomputing*, *31*(1-4), 1-13. doi: 10.1016/S0925-2312(99)00146-0

Hecht-Nelson, R. (1987) Kolmogorov's mapping neural network existence theorem. In *Proceedings of the IEEE International Conference on Neural Networks 111*. New York: IEEE Press, pp. 11-14.

Hornik, K., Stinchcombe, M. and White, H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359-366.

Kashid, D. N. and Kulkarni, S. R. (2002) A more general criterion for subset selection in multiple linear regressions. *Communication in Statistics–Theory & Method, 31*(5), 795-811. doi: 10.1081/STA-120003653

Kasko, B. (1992) *Neural networks and fuzzy systems: a dynamic systems approach to machine intelligence*. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Lawrence, J. (1994) *Introduction to neural networks: design theory and applications*, 6th Ed. Nevada City, CA: California Scientific Software.

Lawrence, J. and Fredrickson, J. (1998) *Brain Maker user's guide and reference manual*. Nevada City, CA: California Scientific Software.

Lippmann, R. P. (1987) An introduction to computing with neural nets. *IEEE Acoustics, Speech and Signal Processing Magazine, 4*(2), 4–22. doi: 10.1109/MASSP.1987.1165576

Mallows, C. L. (1973) Some comments on C_p. *Technometrics*, *15*(4), 661-675. doi: 10.1080/00401706.1973.10489103

Marchandani, G. and Cao, W. (1989) On hidden nodes for neural nets. *IEEE Transactions on Circuits and Systems*, *36*(5), 661–664. doi: 10.1109/31.31313

McDonald, G. C. and Galarneau, D. I. (1975) A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, *70*(350), 407-412. doi: 10.1080/01621459.1975.10479882

Miller, A. J. (2002) *Subset selection in regression*. London: Chapman and Hall.

Montgomery, D., Peck, E. and Vining, G. (2006) *Introduction to linear regression analysis*. New York: John Wiley and Sons Inc.

Ronchetti, E. M. and Staudte, R. G. (1994) A robust version of Mallows's C_p. *Journal of the American Statistical Association, 89*(426), 550-559. doi: 10.1080/01621459.1994.10476780

Sakate D. M. and Kashid D. N. (2014). A deviance-based criterion for model selection in GLM. *Statistics: A Journal of Theoretical and Applied Statistics*, *48*(1), 34-48. doi: 10.1080/02331888.2012.708035

Sommer S. and Huggins R. M. (1996). Variable selection using the Wald test and a robust C_p. *Journal of the Royal Statistical Society C: Applied Statistics*, *45*(1), 15-29. doi: 10.2307/2986219

Springer Link

A New Scale-Invariant Nonparametric Test for Two-Sample Bivariate Location Problem with Application

Robust Rank-Based and Nonparametric Methods pp 175-187 | Cite as

- Sunil Mathur (1) Email author (smathur@gru.edu)
- Deepak M. Sakate (1)
- Sujay Datta (2)

1. Department of Biostatistics and Epidemiology, Medical College of Georgia, Georgia Regents University, , Augusta, USA
2. Department of Statistics, Buchtel College of Arts and Sciences, University of Akron, , Akron, USA

Conference paper
First Online: 21 September 2016

Part of the Springer Proceedings in Mathematics & Statistics book series (PROMS, volume 168)

Abstract

Diagnostic testing in medicine is crucial in determining interventions and treatment plans. It is important to analyze diagnostic tests accurately so that the right decision can be made by clinicians. A scale-invariant test is proposed for when treatment and control samples are available and a change in condition between the treatment and control groups is investigated. The proposed test statistic is shown to have an asymptotically normal distribution. The power of the proposed test is compared with that of several existing tests using Monte Carlo simulation techniques under different bivariate population set-ups. The power study shows that the proposed test statistic performs very well as compared to its competitors for almost all the changes in location and for almost all the distributions considered in this study. The computation of proposed test statistic is shown using a real-life data set.

Keywords

Location test Power Bivariate Wilcoxon's rank sum test Mardia's test

Notes

Acknowledgements

Authors would like to thank the anonymous referee for his/her useful comments which enhanced the clarity of the paper and led to significant improvements in the paper.

Appendix

See Tables 10.2 and 10.3.

Table 10.2

Monte Carlo rejection proportion, sample size $m = 15$, $n = 18$

Distribution	δ_1, δ_2	U	M	WRS	T^2
Bivariate normal	0.00, 0.00	0.0545	0.046	0.0445	0.0485
	0.10, 0.07	0.176	0.065	0.098	0.3875
	0.30, 0.10	0.3885	0.0905	0.145	0.7405
	0.70, 0.50	0.8995	0.367	0.81	1
	1.20, 1.00	0.99	0.659	0.9995	1
	2.40, 3.00	1	0.9375	1	1
BVN mixture P = 0.5	0.00, 0.00	0.055	0.0505	0.0475	0.0505
	0.10, 0.07	0.2115	0.068	0.066	0.112
	0.30, 0.10	0.4735	0.152	0.1725	0.213
	0.70, 0.50	0.8975	0.523	0.845	0.8705
	1.20, 1.00	1	0.855	1	1
	2.40, 3.00	1	0.9995	1	1
BVN mixture P = 0.9	0.00, 0.00	0.047	0.0495	0.0475	0.044
	0.10, 0.07	0.295	0.067	0.061	0.0615
	0.30, 0.10	0.546	0.152	0.125	0.135
	0.70, 0.50	0.9575	0.447	0.687	0.682
	1.20, 1.00	0.9985	0.798	0.9975	0.995
	2.40, 3.00	1	0.9995	1	1

Distribution	δ, δ	U	M	WRS	T
Type VII	0.00, 0.00	0.0515	0.0515	0.0445	0.0495
	0.10, 0.07	0.266	0.0725	0.154	0.098
	0.30, 0.10	0.699	0.193	0.49	0.5965
	0.70, 0.50	1	0.8855	0.9985	1
	1.20, 1.00	1	0.9885	1	1
	2.40, 3.00	1	0.999	1	1
Type II	0.00, 0.00	0.0455	0.035	0.051	0.039
	0.10, 0.07	0.393	0.058	0.1505	0.1515
	0.30, 0.10	0.785	0.2105	0.5135	0.716
	0.70, 0.50	1	0.685	1	1
	1.20, 1.00	1	0.7655	1	1
	2.40, 3.00	1	0.8655	1	1
Population 6	0.00, 0.00	0.05	0.0535	0.0585	0.044
	0.10, 0.07	0.5935	0.069	0.519	0.1825
	0.30, 0.10	0.891	0.0705	0.8635	0.714
	0.70, 0.50	1	0.2755	1	1
	1.20, 1.00	1	0.4225	1	1
	2.40, 3.00	1	0.5110	1	1

Distribution	δ , δ	U	M	WRS	T
Population 7	0.00, 0.00	0.053	0.0535	0.049	0.0495
	0.10, 0.07	0.2135	0.071	0.062	0.0575
	0.30, 0.10	0.368	0.1165	0.1115	0.105
	0.70, 0.50	0.7785	0.3495	0.422	0.4645
	1.20, 1.00	0.9735	0.803	0.927	0.9605
	2.40, 3.00	1	1	1	1

Table 10.3

Monte Carlo rejection proportion, sample size $m = 25$, $n = 28$

Distribution	δ_1, δ_2	U	M	WRS	T^2
Bivariate normal	0.00, 0.00	0.0525	0.046	0.0445	0.0485
	0.10, 0.07	0.0905	0.0585	0.058	0.183
	0.30, 0.10	0.3855	0.0905	0.145	0.7405
	0.70, 0.50	0.888	0.367	0.81	1
	1.20, 1.00	1	0.659	0.9995	1
	2.40, 3.00	1	0.9375	1	1
BVN mixture P = 0.5	0.00, 0.00	0.0535	0.0505	0.0445	0.05
	0.10, 0.07	0.1715	0.0725	0.075	0.0815
	0.30, 0.10	0.387	0.2325	0.2395	0.347
	0.70, 050	1	0.736	0.948	0.991
	1.20, 1.00	1	0.993	1	1
	2.40, 3.00	1	1	1	1
BVN mixture P = 0.9	0.00, 0.00	0.045	0.051	0.058	0.0515
	0.10, 0.07	0.1655	0.0785	0.0795	0.0775
	0.30, 0.10	0.381	0.2085	0.1965	0.2375
	0.70, 0.50	0.995	0.5915	0.833	0.9165
	1.20, 1.00	1	0.9105	1	1
	2.40, 3.00	1	1	1	1

Distribution	δ , δ	U	M	WRS	T
Type VII	0.00, 0.00	0.0445	0.0425	0.047	0.0525
	0.10, 0.07	0.4095	0.0725	0.3285	0.2525
	0.30, 0.10	1	0.353	0.6985	0.9445
	0.70, 0.50	1	0.9995	1	1
	1.20, 1.00	1	1	1	1
	2.40, 3.00	1	1	1	1
Type II	0.00, 0.00	0.048	0.046	0.043	0.046
	0.10, 0.07	0.3205	0.0835	0.1925	0.219
	0.30, 0.10	1	0.416	0.611	0.914
	0.70, 0.50	1	0.825	1	1
	1.20, 1.00	1	0.858	1	1
	2.40, 3.00	1	0.9985	1	1
Population 6	0.00, 0.00	0.0506	0.057	0.049	0.051
	0.10, 0.07	0.899	0.0685	0.7225	0.275
	0.30, 0.10	1	0.0975	0.921	0.8975
	0.70, 0.50	1	0.2345	1	1
	1.20, 1.00	1	0.4775	1	1
	2.40, 3.00	1	0.5655	1	1

Distribution	δ , δ	U	M	WRS	T
Population 7	0.00, 0.00	0.051	0.0465	0.047	0.0585
	0.10, 0.07	0.323	0.0715	0.0655	0.0725
	0.30, 0.10	0.536	0.2005	0.125	0.145
	0.70, 0.50	0.941	0.604	0.6025	0.762
	1.20, 1.00	0.999	0.9815	0.99	0.9995
	2.40, 3.00	1	1	1	1

References

Armitage, P., Berry, G., & Matthews, J. N. S. (2008). *Statistical methods in medical research*. Wiley-Blackwell, Massachusetts, USA.
Google Scholar (http://scholar.google.com/scholar_lookup?title=Statistical%20methods%20in%20medical%20research&author=P..%20Armitage&author=G..%20Berry&author=J.%20N.%20S..%20Matthews&publication_year=2008)

Baringhaus, L., & Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis, 88*, 190–206.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=2021870)
CrossRef (https://doi.org/10.1016/S0047-259X(03)00079-4)
zbMATH (http://www.emis.de/MATH-item?1035.62052)
Google Scholar (http://scholar.google.com/scholar_lookup?title=On%20a%20new%20multivariate%20two-sample%20test&author=L..%20Baringhaus&author=C..%20Franz&journal=Journal%20of%20Multivariate%20Analysis&volume=88&pages=190-206&publication_year=2004)

Beghi, E. (2004). Efficacy and tolerability of the new antiepileptic drugs: Comparison of two recent guidelines. *The Lancet Neurology, 3*, 618–621.
CrossRef (https://doi.org/10.1016/S1474-4422(04)00882-8)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Efficacy%20and%20tolerability%20of%20the%20new%20antiepileptic%20drugs%3A%20Comparison%20of%20two%20recent%20guidelines&author=E..%20Beghi&journal=The%20Lancet%20Neurology&volume=3&pages=618-621&publication_year=2004)

Belle, G. V., Fisher, L. D., Heaerty, P. J., & Lumley, T. (2004). *Biostatistics: A methodology for the health sciences* (2nd ed.). New York: Wiley.
CrossRef (https://doi.org/10.1002/0471602396)
zbMATH (http://www.emis.de/MATH-item?1136.62401)

Google Scholar (http://scholar.google.com/scholar_lookup?title=Biostatistics%3A%20A%20methodology%20for%20the%20health%20sciences&author=G.%20V..%20Belle&author=L.%20D..%20Fisher&author=P.%20J..%20Heaerty&author=T..%20Lumley&publication_year=2004)

Blumen, I. (1958). A new bivariate sign test. *Journal of the American Statistical Association, 53*, 448–456.
CrossRef (https://doi.org/10.1080/01621459.1958.10501451)
zbMATH (http://www.emis.de/MATH-item?0087.14702)
Google Scholar (http://scholar.google.com/scholar_lookup?title=A%20new%20bivariate%20sign%20test&author=I..%20Blumen&journal=Journal%20of%20the%20American%20Statistical%20Association&volume=53&pages=448-456&publication_year=1958)

Brown, B., & Hettmansperger, T. (1987). Affine invariant rank methods in the bivariate location model. *Journal of the Royal Statistical Society Series B (Methodological), 49*, 301–310.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=928938)
zbMATH (http://www.emis.de/MATH-item?0653.62039)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Affine%20invariant%20rank%20methods%20in%20the%20bivariate%20location%20model&author=B..%20Brown&author=T..%20Hettmansperger&journal=Journal%20of%20the%20Royal%20Statistical%20Society%20Series%20B%20%28Methodological%29&volume=49&pages=301-310&publication_year=1987)

Chung, E., & Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics, 41*, 484–507.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=3099111)
CrossRef (https://doi.org/10.1214/13-AOS1090)
zbMATH (http://www.emis.de/MATH-item?1267.62064)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Exact%20and%20asymptotically%20robust%20permutation%20tests&author=E..%20Chung&author=J.%20P..%20Romano&journal=The%20Annals%20of%20Statistics&volume=41&pages=484-507&publication_year=2013)

Davis, J. B., & McKean, J. W. (1993). Rank-based methods for multivariate linear models. *Journal of the American Statistical Association, 88*, 245–251.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=1212488)
zbMATH (http://www.emis.de/MATH-item?0779.65093)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Rank-based%20methods%20for%20multivariate%20linear%20models&author=J.%20B..%20Davis&author=J.%20W..%20McKean&journal=Journal%20of%20the%20American%20Statistical%20Association&volume=88&pages=245-251&publication_year=1993)

Deeks, J. J. (2001). Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *British Medical Journal, 323*, 157.
CrossRef (https://doi.org/10.1136/bmj.323.7305.157)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Systematic%20reviews%20in%20health%20care%3A%20Systematic%20reviews%20of%20evaluations%20of%20diagnostic%20and%20screening%20tests&author=J.%20J..%20Deeks&journal=British%20Medical%20Journal&volume=323&pages=157&publication_year=2001)

Dietz, E. J. (1982). Bivariate nonparametric tests for the one-sample location problem. *Journal of the American Statistical Association, 77*, 163–169.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=648040)

CrossRef (https://doi.org/10.1080/01621459.1982.10477781)
zbMATH (http://www.emis.de/MATH-item?0489.62043)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Bivariate%20nonparametric%20tests%20for%20the%20one-sample%20location%20problem&author=E.%20J..%20Dietz&journal=Journal%20of%20the%20American%20Statistical%20Association&volume=77&pages=163-169&publication_year=1982)

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences, 180*, 2044–2064.
CrossRef (https://doi.org/10.1016/j.ins.2009.12.010)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Advanced%20nonparametric%20tests%20for%20multiple%20comparisons%20in%20the%20design%20of%20experiments%20in%20computational%20intelligence%20and%20data%20mining%3A%20Experimental%20analysis%20of%20power&author=S..%20Garc%C3%ADa&author=A..%20Fern%C3%A1ndez&author=J..%20Luengo&author=F..%20Herrera&journal=Information%20Sciences&volume=180&pages=2044-2064&publication_year=2010)

Jurečková, J., & Kalina, J. (2012). Nonparametric multivariate rank tests and their unbiasedness. *Bernoulli, 18*, 229–251.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=2888705)
CrossRef (https://doi.org/10.3150/10-BEJ326)
zbMATH (http://www.emis.de/MATH-item?1291.62095)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Nonparametric%20multivariate%20rank%20tests%20and%20their%20unbiasedness&author=J..%20Jure%C4%8Dkov%C3%A1&author=J..%20Kalina&journal=Bernoulli&volume=18&pages=229-251&publication_year=2012)

Kowalski, J., & Tu, X. M. (2008). *Modern applied U-statistics* (Vol. 714). Wiley.com (http://Wiley.com).

Larocque, D., Tardif, S., & Eeden, C. V. (2003). An affine-invariant generalization of the Wilcoxon signed-rank test for the bivariate location problem. *Australian and New Zealand Journal of Statistics, 45*, 153–165.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=1983348)
CrossRef (https://doi.org/10.1111/1467-842X.00271)
zbMATH (http://www.emis.de/MATH-item?1064.62054)
Google Scholar (http://scholar.google.com/scholar_lookup?title=An%20affine-invariant%20generalization%20of%20the%20Wilcoxon%20signed-rank%20test%20for%20the%20bivariate%20location%20problem&author=D..%20Larocque&author=S..%20Tardif&author=C.%20V..%20Eeden&journal=Australian%20and%20New%20Zealand%20Journal%20of%20Statistics&volume=45&pages=153-165&publication_year=2003)

Lee, A. J. (1990). *U-Statistics: Theory and practice*. Boca Raton, FL: CRC Press.
zbMATH (http://www.emis.de/MATH-item?0771.62001)
Google Scholar (http://scholar.google.com/scholar_lookup?title=U-Statistics%3A%20Theory%20and%20practice&author=A.%20J..%20Lee&publication_year=1990)

Mardia, K. (1967). A non-parametric test for the bivariate two-sample location problem. *Journal of the Royal Statistical Society. Series B (Methodological), 29*, 320–342.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=221690)

zbMATH (http://www.emis.de/MATH-item?0157.48004)
Google Scholar (http://scholar.google.com/scholar_lookup?title=A%20non-parametric%20test%20for%20the%20bivariate%20two-sample%20location%20problem&author=K..%20Mardia&journal=Journal%20of%20the%20Royal%20Statistical%20Society.%20Series%20B%20%28Methodological%29&volume=29&pages=320-342&publication_year=1967)

Mathur, S. K., & Smith, P. F. (2008). An efficient nonparametric test for bivariate two-sample location problem. *Statistical Methodology, 5*, 142–159.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=2424750)
CrossRef (https://doi.org/10.1016/j.stamet.2007.07.001)
zbMATH (http://www.emis.de/MATH-item?1248.62067)
Google Scholar (http://scholar.google.com/scholar_lookup?title=An%20efficient%20nonparametric%20test%20for%20bivariate%20two-sample%20location%20problem&author=S.%20K..%20Mathur&author=P.%20F..%20Smith&journal=Statistical%20Methodology&volume=5&pages=142-159&publication_year=2008)

McDonald, H. P., Garg, A. X., & Haynes, R. B. (2002). Interventions to enhance patient adherence to medication prescriptions. *The Journal of the American Medical Association, 288*, 2868–2879.
CrossRef (https://doi.org/10.1001/jama.288.22.2868)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Interventions%20to%20enhance%20patient%20adherence%20to%20medication%20prescriptions&author=H.%20P..%20McDonald&author=A.%20X..%20Garg&author=R.%20B..%20Haynes&journal=The%20Journal%20of%20the%20American%20Medical%20Association&volume=288&pages=2868-2879&publication_year=2002)

Oja, H. (1999). Affine invariant multivariate sign and rank tests and corresponding estimates: A review. *Scandinavian Journal of Statistics, 26*, 319–343.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=1712063)
CrossRef (https://doi.org/10.1111/1467-9469.00152)
zbMATH (http://www.emis.de/MATH-item?0938.62063)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Affine%20invariant%20multivariate%20sign%20and%20rank%20tests%20and%20corresponding%20estimates%3A%20A%20review&author=H..%20Oja&journal=Scandinavian%20Journal%20of%20Statistics&volume=26&pages=319-343&publication_year=1999)

Peddada, S. D., Haseman, J. K., Tan, X., & Travlos, G. (2006). Tests for a simple tree order restriction with application to dose–response studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 55*, 493–506.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=2242276)
CrossRef (https://doi.org/10.1111/j.1467-9876.2006.00549.x)
zbMATH (http://www.emis.de/MATH-item?05188750)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Tests%20for%20a%20simple%20tree%20order%20restriction%20with%20application%20to%20dose%E2%80%93response%20studies&author=S.%20D..%20Peddada&author=J.%20K..%20Haseman&author=X..%20Tan&author=G..%20Travlos&journal=Journal%20of%20the%20Royal%20Statistical%20Society%3A%20Series%20C%20%28Applied%20Statistics%29&volume=55&pages=493-506&publication_year=2006)

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press.
zbMATH (http://www.emis.de/MATH-item?1039.62105)

Google Scholar (http://scholar.google.com/scholar_lookup?title=The%20statistical%20evaluation%20of%20medical%20tests%20for%20classification%20and%20prediction&author=M.%20S..%20Pepe&publication_year=2003)

Peters, D., & Randles, R. H. (1990). A multivariate signed-rank test for the one-sample location problem. *Journal of the American Statistical Association, 85*, 552–557.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=1141757)
CrossRef (https://doi.org/10.1080/01621459.1990.10476234)
zbMATH (http://www.emis.de/MATH-item?0709.62051)
Google Scholar (http://scholar.google.com/scholar_lookup?title=A%20multivariate%20signed-rank%20test%20for%20the%20one-sample%20location%20problem&author=D..%20Peters&author=R.%20H..%20Randles&journal=Journal%20of%20the%20American%20Statistical%20Association&volume=85&pages=552-557&publication_year=1990)

Peters, D., & Randles, R. H. (1991). A bivariate signed rank test for the two-sample location problem. *Journal of the Royal Statistical Society. Series B (Methodological), 53*, 493–504.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=1108344)
zbMATH (http://www.emis.de/MATH-item?0800.62252)
Google Scholar (http://scholar.google.com/scholar_lookup?title=A%20bivariate%20signed%20rank%20test%20for%20the%20two-sample%20location%20problem&author=D..%20Peters&author=R.%20H..%20Randles&journal=Journal%20of%20the%20Royal%20Statistical%20Society.%20Series%20B%20%28Methodological%29&volume=53&pages=493-504&publication_year=1991)

Randles, R. H., & Peters, D. (1990). Multivariate rank tests for the two-sample location problem. *Communications in Statistics-Theory and Methods, 19*, 4225–4238.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=1103009)
CrossRef (https://doi.org/10.1080/03610929008830439)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Multivariate%20rank%20tests%20for%20the%20two-sample%20location%20problem&author=R.%20H..%20Randles&author=D..%20Peters&journal=Communications%20in%20Statistics-Theory%20and%20Methods&volume=19&pages=4225-4238&publication_year=1990)

Schneeweiss, S., Gagne, J., Glynn, R., Ruhl, M., & Rassen, J. (2011). Assessing the comparative effectiveness of newly marketed medications: Methodological challenges and implications for drug development. *Clinical Pharmacology and Therapeutics, 90*, 777–790.
CrossRef (https://doi.org/10.1038/clpt.2011.235)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Assessing%20the%20comparative%20effectiveness%20of%20newly%20marketed%20medications%3A%20Methodological%20challenges%20and%20implications%20for%20drug%20development&author=S..%20Schneeweiss&author=J..%20Gagne&author=R..%20Glynn&author=M..%20Ruhl&author=J..%20Rassen&journal=Clinical%20Pharmacology%20and%20Therapeutics&volume=90&pages=777-790&publication_year=2011)

Sugiura, N. (1965). *Multisample and multivariate nonparametric tests based on U statistics and their asymptotic efficiencies. Osaka Journal of Mathematics, 2*, 385–426.
MathSciNet (http://www.ams.org/mathscinet-getitem?mr=192607)
Google Scholar (http://scholar.google.com/scholar_lookup?&author=N..%20Sugiura&journal=Multisample%20and%20multivariate%20nonparametric%20tests%20based%20on%20U%20statistics%20and%20their%20asymptotic

%20efficiencies.%20Osaka%20Journal%20of%20Mathematics&volume=2&pages=385-426&publication_year=1965)

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge: Cambridge University Press.
zbMATH (http://www.emis.de/MATH-item?0910.62001)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Asymptotic%20statistics&author=A.%20W..%20Vaart&publication_year=2000)

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic Press, USA.
zbMATH (http://www.emis.de/MATH-item?1270.62051)
Google Scholar (http://scholar.google.com/scholar_lookup?title=Introduction%20to%20robust%20estimation%20and%20hypothesis%20testing&author=R.%20R..%20Wilcox&publication_year=2012)

Woolson, R. F., & Clarke, W. R. (2011). *Statistical methods for the analysis of biomedical data*. (Vol. 371). John Wiley and Sons, New York.
Google Scholar (http://scholar.google.com/scholar_lookup?title=%0AStatistical%20methods%20for%20the%20analysis%20of%20biomedical%20data%20%28Vol&author=R.%20F..%20Woolson&author=W.%20R..%20Clarke&publication_year=2011)

Copyright information

About this paper

Cite this paper as:

Mathur S., Sakate D.M., Datta S. (2016) A New Scale-Invariant Nonparametric Test for Two-Sample Bivariate Location Problem with Application. In: Liu R., McKean J. (eds) Robust Rank-Based and Nonparametric Methods. Springer Proceedings in Mathematics & Statistics, vol 168. Springer, Cham

- DOI https://doi.org/10.1007/978-3-319-39065-9_10
- Publisher Name Springer, Cham
- Print ISBN 978-3-319-39063-5
- Online ISBN 978-3-319-39065-9
- eBook Packages Mathematics and Statistics

Personalised recommendations

Electronic Journal of Applied Statistical Analysis
Vol. 07, Issue 02, 2014, 228-253
DOI: 10.1285/i20705948v7n2p228

Reliability estimation of k-unit series system based on progressively censored data

K. G. Potdar*[a] and D. T. Shirke[b]

[a]*Department of Statistics, Ajara Mahavidyalaya, Ajara, Dist-Kolhapur, Maharashtra, India - 416505.*

[b]*Department of Statistics, Shivaji University, Kolhapur, Dist-Kolhapur, Maharashtra, India - 416004.*

Published: 14 October 2014

In this article, we consider a k-unit series system with component lifetime distribution to be a member of the scale family of distributions. We discuss estimation of the scale parameter and estimation of reliability function of the family based on progressively Type-II censored sample. The maximum likelihood estimator (MLE) of the scale parameter is derived using Expectation-Maximization (EM) algorithm and is used to estimate reliability function. Confidence intervals are constructed using asymptotic distribution of MLE. β -expectation tolerance interval for lifetime of the system is obtained. We consider half-logistic distribution as a member of the scale family and study performance of the MLE, reliability estimate and confidence interval using simulation experiments. Illustration through real data example is provided.

keywords: Progressively Type-II censoring, EM algorithm, MLE, confidence interval, coverage probability, reliability, β-expectation tolerance interval, half-logistic distribution.

1 Introduction

In industrial phenomenon series systems are widely used. Electric, automobile as well as in chemical industry various units are connected in series. Here system is working if all

*Corresponding author: potdarkiran.stat@gmail.com.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

E-Information Literacy Index of University Teachers of Maharashtra, India: A Case Study

Article *in* DESIDOC Journal of Library & Information Technology · November 2017

DOI: 10.14429/djlit.37.10941

1 author:

Prakash Bilawar
Shivaji University, Kolhapur
11 PUBLICATIONS **1** CITATION

SEE PROFILE

DESIDOC Journal of Library & Information Technology, Vol. 37, No. 6, November 2017, pp. 432-436, DOI : 10.14429/djlit.37.10941

E-Information Literacy Index of University Teachers of Maharashtra, India: A Case Study

Prakash Bhairu Bilawar[*], Shamprasad M. Pujar[!] and Somanath Dasharath Pawar[#]

[*]*Balasaheb Khardekar Library, Shivaji University, Kolhapur - 416 004, India*
[!] *Indira Gandhi Institute of Development Research, Mumbai - 400 065, India*
[#]*Department of Statistics, Shivaji University, Kolhapur - 416 004, India*
[*]*E-mail: pbb_lib@unishivaji.ac.in*

ABSTRACT

The purpose of this paper is to propose an e-information literacy index that provides realistic values to distinguish whether university teachers are literate in regard to awareness and use of e-information resources by explaining the characteristics of e-information literate teacher. The present survey attempts to formulate e-information literacy index of university teachers by taking into consideration three components viz. awareness of e-resources, availability of ICT facilities and use of internet services and search techniques to retrieve e-information. The findings shows that 60.52 per cent teachers are e-information literate. Amongst the teachers, the index for Assistant Professors is highest followed by Professors and Associate Professors. It indicates that Assistant Professors are more e-information literate than their superiors. Amongst the universities, the index of Shivaji University, Kolhapur is highest. As far as author's consciousness, there are several indices meant for different purposes but in the higher education sector to define the characteristics of e-information literate university teacher in terms of an index is unique and special.

Keywords: E-information literacy; University teachers; Information literacy; Indicators

1. INTRODUCTION

Today, we leave in an era surrounded by digital sea of information. Owing to the availability of vast array of unfiltered information on a given topic, the process of identifying and selecting peculiar e-information has become complex. In this circumstance E-Information literacy directs the users towards authentic and reliable sources of information available online useful for their informed judgements against the quest for information. E-Information literacy is the ability to properly use and evaluate electronic resources, tools and services and apply it for lifelong learning process. E-information literacy among the university teachers contributes towards their learning process and brings in overall change in the way how they collect and use information.

The present study intends to define the e-information literacy rank amongst the university teachers in tech savvy environment considering their awareness, use and retrieval of e-information from e-resources in the form of an index value. E-information literacy index is a statistical measure used to determine how university teachers are making best use of e-information for their teaching and research purposes. The index values were determined against the responses given by teachers for proposed and defined clusters of components/ indicators mentioned in *Appendix A*. The exercise helped to enlist the qualities of e-information literate teachers in the vast and changing digital sea of information. It has been found that the formulated index values differ amongst teachers and the universities under study depending upon their ability, performance in regard to the use and searching techniques applied for getting e-information.

Received : 1 December 2016, Revised : 14 August 2017
Accepted : 18 September 2017, Online published : 07 November 2017

2. LITERATURE REVIEW

Hargittai[1] recommend for the creation of an index variable as proxies for web-oriented digital literacy measures on Internet use and methodology based on verifying the validity of the measures derived from their relationship with actual skill measures. She again revisited her survey measures with new terms in order to assess the change in digital literacy measures of the respondents and found discrepancy older Internet terms and new web-based concepts thus resulting in change in the index values[2]. Thornbush[3] suggested S-E index that provides a broader classification of weathering processes based on visible surface forms in the field of archaeogeomorphological research. Katz & others[4] conducted a survey to measure the cumulative, holistic impact of discrete ICT (Information and Communication Technologies) and a composite digitisation. An index was developed based on six overarching components, viz. affordability, infrastructure investment, network access, capacity, usage, and human capital. The findings showed that proper ICT infrastructure and attention towards digital technology usage is required for better flow and awareness of digital literacy. Alguliyev & Others[5] explore an index for

evaluating the quality of research output of researchers with the 25 indices which shows that the weighted index may serve as a supplement to h-index and its variants. Sahoo[6] propose the I-index which states that an author's percentage shares in the total citations that his/her papers have attracted. The index is useful to know comprehensive idea of an author's overall research performance.

3. OBJECTIVES

The core objectives of the study :

(i) To know the level of awareness of e-resources and searching techniques applied by the university teachers in retrieving e-information

(ii) To study the availability of ICT facilities for the use of internet services by the university teachers; and

(iii) To formulate an e-information literate index of university teachers.

4. METHODOLOGY

For the present study, descriptive method of research has been used. The data was collected through structured questionnaire distributed to targeted sample of 360 university teachers of 43 different departments working in the 10 state universities of Maharashtra, India in the faculties of sciences, social sciences and humanities (languages). A total of 347 teachers responded (96.38 per cent) to the survey. Their literacy levels were tested based on their self-perceived skills and skills learnt with the help of others.

4.1 Methodology Used

Keeping in mind the search for e-information, access and retrieval techniques applied by a normal user, a common strategy in terms of methodological (measuring) indicators were suggested that defines the qualities of e-information literate user with an index value against suggested cut-off value. These methodological indicators were applied for the targeted group of teachers working in the universities under study. The proposed index is based on analysis of indicators against the clusters which results in certain startling outcomes.

The suggested clusters and their indicators may also be applied to other teachers working in different disciplines / universities by changing the clustered framework in regard to the ICT advancement and its searching techniques. To formulate an e-information literacy index of university teachers a series of questions were designed which comprised of 65 indicators comprising of tick marked and five point scale questions, which were equally weighted (0.33) Table 1 and grouped in 3 clusters of components viz. Awareness of e-resources (23 indicators); Availability of ICT facilities and Use of Internet Services (14 indicators) and the search techniques to retrieve e-information (28 indicators) to measure the e-information literate characteristics of the teachers, enlisted in *Appendix A*. The equal weight is calculated as 1/3 = 0.33 to represent the index value as '0' and '1' receptively. The resulted measures depend on the aspects related to e-information awareness and use, ICT facilities and searching skills which help in assessing their e-information literacy skills.

The proposed measuring indicators were tested with

Table 1. Weightage criteria

Components	Weightage
A. Awareness of e-resources [23 Indicators]	0.33 [0.33/23= @0.0143/ per Question]
B. Availability of ICT facilities and use of internet services [14 Indicators]	0.33 [0.33/14= @0.023/ per Question]
C. Searching techniques to retrieve e-information [28 Indicators]	0.33 [0.33/28= @0.011/ per Question]

responses given by the university teachers. However, before calculating the index except tick marked questions all the five point question response values were converted between 0 and 1 as 0, 0.25, 0.50, 0.75 and 1 in order to show the similarity that will be useful for calculating an index by proposing a cut-off value at 0.5.

Table 2. E-information literate index of the teacher

A	B	C	D	E	F	G (Index)	H (Literate/ Illiterate)
10	10	13.25	0.4348	0.7143	0.4732	0.5408	* 1

*1 = Literate and 0 = Illiterate

As a sample, the index of first teacher was calculated in the following way:

A = Sum of response value of first component

B = Sum of response value of second component

C = Sum of response value of third component

D = A/23, E= B/14, F= C/28

G = Index (Average of D, E and F)

H = The first teacher suppose to be e-information literate considering cut-off value at 0.5 value and the index is above cut-off value.

Accordingly, an index was calculated for all the teachers under study (shown in histogram) to represent whether they are e-information literate or illiterate.

It is clear from Fig. 1 and Table 3 the lowest index observed was 0.0766 and highest was 0.9167. Majority of the teachers are having e-information literacy index between 0.3 and 0.8. The index level was highest between the ranges 0.6 and 0.7. Out of 347 respondents, 23% (79) of university teachers are having e-information literacy index between 0.6 and 0.7. About

Table 3. Summary of an Index

Statistics	Value
Mean	0.5381
Standard error	0.0091
Median	0.5527
Mode	0.7222
Standard deviation	0.1704
Sample variance	0.0290
Kurtosis	-0.4425
Skewness	-0.2648
Range	0.8401
Minimum	0.0766
Maximum	0.9167
Sum	186.7232
Count	347.0000

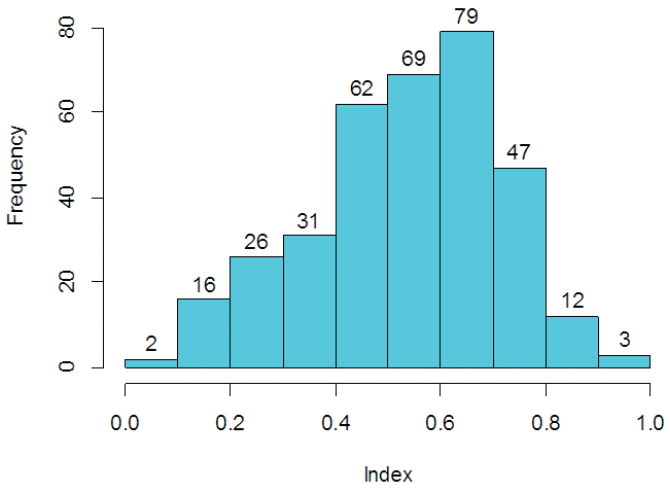


Figure 1. Histogram of an e-information literacy index of teachers.

74% (257) teachers are having e-information literacy index between 0.4 and 0.8. It has been found that the distribution of e-information literacy index is not symmetric owing to differing skill levels of teachers. Further, e-information literacy index has negatively skewed and it shows relatively flat distribution. 210 (60.52%) teachers were found to be e-information literate and remaining 137 (39.48%) were not e-information literate.

5. FINDINGS

From Tables 4 and 5, we may draw following findings;

- It is found that 210 (60.52%) teachers were e-information literate based on index value.
- When looked across the disciplines of sciences, social sciences and arts and humanities, it is proved that Science faculties (0.5835) are more e-information literate than Social Science (0.5427) and Arts and Humanities (0.4616) faculties.
- From the gender based analysis, it was found that the index is high in case of female teachers (0.5516) than the male teachers (0.5309).
- In addition, from the designation wise analysis it was found that index for Assistant Professors was highest (0.5621) followed by Professors (0.5338) and Associate Professors (0.4975).

Table 4. E-Information literate Index ratio

		Index
Faculty	Science	0.5835
	Social Science	0.5427
	Arts and Humanities	0.4616
Gender	Female	0.5516
	Male	0.5309
Designation	Assistant Professor	0.5621
	Associate Professor	0.4975
	Professor	0.5338

- Regarding university wise e-information literacy index, it was observed that the e-information literate index was higher in case of Shivaji University, Kolhapur (0.6093), followed by Sant Gadge Baba Amravati University, Amravati (0.5865), Swami Ramanand Teerth Marathwada University, Nanded (0.5716), North Maharashtra University, Jalgaon (0.5645), University of Mumbai, Mumbai (0.5466), Solapur University, Solapur (0.5338), University of Pune, Pune (0.5274), SNDT (Smt. Nathibai Damodar Thackersey) Women's University, Mumbai (0.5097), Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur (0.4978), Dr Babasaheb Ambedkar Marathwada University, Aurangabad (0.4324).

Table 5. University wise e-information literate Index ratio

University	Index
Sant Gadge Baba Amravati University, Amravati	0.5865
Dr Babasaheb Ambedkar Marathwada Uni, Aurangabad	0.4324
North Maharashtra University, Jalgaon	0.5645
University of Mumbai, Mumbai	0.5466
Rashtrasant Tukadoji Maharaj Nagpur Uni, Nagpur	0.4978
Swami Ramanand Teerth Marathwada Uni, Nanded	0.5716
University of Pune, Pune	0.5274
Shivaji University, Kolhapur	0.6093
SNDT (Smt. Nathibai Damodar Thackersey) Women's University, Mumbai	0.5097
Solapur University, Solapur	0.5338
Grand Total	0.5381

Thus from the above detailed explanations it is revealed that depending upon the ICT/self skills of university teachers in handling e-information, awareness about different e-resources, tools and techniques for searching, accessing and retrieving e-information either from the internet or from subscribed e-resources and availability of sufficient infrastructure at the universities, the e-information literacy index of teachers calculated varies from teacher to teacher amongst the faculties and universities. The awareness and use of Web 2.0 along with the internet services by the university teachers was an additional verifying criteria used to measure the e-information literacy level of the teachers in terms of an index value.

6. CONCLUSIONS

The difference in e-information literacy index among the institutions and groups may be attributed to the efforts taken by each of the universities in building the required ICT infrastructure, training teachers in the effective retrieval and use of e-information and teachers self skills. The poor index value of university teachers needs to be accounted with sufficient awareness campaigns, ICT facilities and online training about searching techniques by the universities/ university libraries. Further academic/learning and research tasks of the university teachers may be strengthened by arranging discipline specific user awareness programmes and also by allocating certain hours per week in the time-table especially for searching and seeking e-information from different sources. This also may be made as part of the continued education programme for faculty members to become independent learners.

REFERENCES

1. Hargittai, E. Survey measures of web-oriented digital literacy. *Social Science Computer Review*, 2005, **23** (3), 371-379.
2. Hargittai, E. An Update on survey measures of web-oriented digital literacy. *Social Sci. Comput. Rev.*, 2009, **27**(1), 130-137.
3. Thornbush, M.J. & Thornbush, S.E. The application of a limestone weathering index at churchyards in central Oxford, UK. *Applied Geography,* 2013, **42**, 157-164.
4. Katz, R. & Others, Using a digitization index to measure the economic and social impact of digital agendas. *Info.*, 2014, **1**(1), 32-44.
5. Alguliyev, R. & Others, An aggregated index for assessment of the scientific output of researchers. *Int. J. Knowledge Manag. Stud.*, 2015, **6** (1), 31-62.
6. Sahoo, S. Analyzing research performance: proposition of a new complementary index. *Scientometrics*, 2016, **108**(2), 489-504.
7. Investopedia. http://www.investopedia.com/terms/i/index.asp. (Accessed on 24 November 2016).

CONTRIBUTORS

Dr Prakash Bhairu Bilawar has completed his BLISc, MLISc and PhD in Library and Information Science from Shivaji University, Kolhapur. Presently working as 'Assistant Librarian' (Senior Scale) at B.B.K. Library (Knowledge Resource Center), Shivaji University, Kolhapur, Maharashtra. He has 20 research publications in journals, conference proceedings/books. His areas of interest are ICT, information sources and services.

Dr Shamprasad M. Pujar has received PhD in Library and Information Science from Karnataka University, Dharwad. Presently working as Deputy Librarian at Indira Gandhi Institute of Development Research, Mumbai. He has contributed more than 35 papers in journals and conferences. His area of interest include : ICT applications for libraries, Web 2.0, OERs, open access journals, altmetrics, MOOCs etc.

Mr Somanath D. Pawar is perusing his PhD (Statistics) from Shivaji University, Kolhapur. Presently working as Assistant Professor in Statistics at Department of Statistics, Shivaji University, Kolhapur. He has one research paper to his credit and has presented more than 5 research papers in conferences. His area of interest includes nonparametric statistical inference, applied statistics.

Appendix A

Measuring indicators for E-Information Literacy of the Teacher

Components	Indicators	Type of Questions	Weightage
A. Awareness of e-resources	1. Citation Indexes: Web of Science [SCI, SSCI, AHCI] SCOPUS etc 2. Digital Libraries/E-Print Archives/Institutional Repositories 3. Discussion forums/ Groups 4. E-Books 5. E-Journals (including Open Access/Free Journals) 6. Electronic Abstracting and Indexing Databases 7. Electronic Theses and Dissertations 8. E-Newspapers 9. General Search Engines 10. Journal contents alert services 11. Scholarly Search Engines 12. Subject Gateways and portals 13. Subject Specific Search Engines 14. E-resources from INFLIBNET consortium 15. Open access online databases/resources 16. Web 2.0 tools- Blogs 17. Chatting 18. Micro-blogs [Twitter] 19. Phone 20. Reference management tools like Zotero, Mendeley etc 21. RSS feeds 22. Social Networking sites 23. Wikis	All Tick [√] Marked questions	0.33 [0.33/23= @0.0143/ per Question]

B. Availability of ICT facilities and use of internet services	1. Computer 2. Internet connection 3. Multimedia Projector 4. Photocopying Machines 5. Printer 6. Scanner/Fax 7. CD-ROM/DVD databases 8. Communication i.e. e-mail, chatting, phone etc 9. Downloading information i.e. articles, reports, forms etc 10. Links to abstract, Full Text, Citation (reference) and other useful e- resources in the field 11. Listening to music and watching videos (Ex: You tube) 12. Reading online newspapers, newsletters, blogs etc 13. Searching information 14. Watching video lectures from academic/research organization	All Tick [√] Marked questions	0.33 [0.33/14= @0.023/ per Question]
C. Searching techniques to retrieve e-information	1. Directly going to source of information [Knowing web address from references] 2. E-Journals/Databases, websites 3. E-Resources linked through library website 4. General Search engines 5. Guided Search/FAQ/Help 6. Meta Search Engines 7. Scholarly Search Engines 8. Subject Directories/Gateways 9. Subject Specific Search engines 10. Use Subject bookmarking sites 11. Just enter keywords in simple search box 12. Just enter title or author in simple search box 13. Make use of Advance search options 14. Make use of Boolean operators [and, or and not] along with keywords 15. Make use of mathematical operators [+, -, *] along with keywords 16. Make use of Phrase search by putting content in “ ” 17. Make use of proximity operators [near, between etc] 18. Make search for content within specific domains [.edu, ac.in, co. in etc] 19. Make search for content within specific languages [English, Hindi, French etc] 20. Make search for content within the files [PDF, HTML, DOC, Xls etc] 21. Browsing Content from E-Print archives/Digital Library/ Institutional Repository 22. By browsing journal articles from Journal homepages 23. Search for articles using Google 24. Search for articles using Google Scholar 25. Search for articles using Journals database Search options 26. Through Abstracting and Indexing Databases 27. Through library OPAC [Article Indexing] 28. Through links provided in e-mail table of contents alerts	All 5 Point Scales questions	0.33 [0.33/28= @0.011/ per Question]

tolerance intervals for the lifetime distribution of K-unit parallel system based on generalized variable

mber 2017 *with* 10 Reads

S.S. Godase

Digambar Shirke
18.83 · Shivaji University, Kolhapur

D.N. Kashid

icle, we consider the problem of setting prediction interval for future sample, when lifetime
ın of a unit in a k-unit parallel system has exponential distribution based on generalized
GV) approach. We also discuss one sided tolerance limits and tolerance intervals based on GV
. Performance of both intervals are studied using simulation and compared them with existing
xhibit superiority of the proposed method. The prediction interval is illustrated through real life

Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v10n1p29

Tolerance intervals and confidence intervals for the scale parameter of Pareto-Rayleigh distribution

By Godase, Shirke, Kashid

Published: 26 April 2017

Electronic Journal of Applied Statistical Analysis
Vol. 10, Issue 01, April 2017, 29-49
DOI: 10.1285/i20705948v10n1p29

Tolerance intervals and confidence intervals for the scale parameter of Pareto-Rayleigh distribution

Godase S.S.*[a], Shirke D.T.[b], and Kashid D.N.[b]

[a]*Department of Statistics, S. G. M. College, Karad, India*
[b]*Department of Statistics, Shivaji university, Kolhapur, India*

Published: 26 April 2017

In this paper we consider Pareto-Rayleigh distribution as an example of a Transformed-Transformer family of distributions defined by Alzaatreh et al. (2013b). We construct confidence intervals (CIs) and tolerance intervals (TIs) using generalized variable approach due to Weerahandi (1993) by using maximum likelihood estimator and modified maximum likelihood estimator of the scale parameter. Performance of both the intervals is studied using simulation and compared with the existing ones to exhibit superiority of the proposed intervals. Proposed confidence intervals and tolerance intervals are illustrated through real life data.

keywords: Transformed-transformer (T-X) family, Pareto-Rayleigh distribution, generalized pivotal quantity, confidence intervals and tolerance intervals.

1 Introduction

Pareto distribution has been widely used in modeling heavy-tailed distributions, such as income distribution. Many applications of the Pareto distribution in economics, biology and physics can be found in the literature. Schroeder et al. (2010) presented an application of the Pareto distribution in modeling disk drive sector errors. Mahmoudi (2011) discusses the beta generalized Pareto distribution with application to life time data. The Pareto distribution has been recognized as a suitable model for many non-negative socio-economic variables. Pareto distribution is useful in individual income,

*Corresponding author: suwarna_godase@rediffmail.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

family income and income before taxes etc. In literature various generalizations of the Pareto distribution have been derived such as Beta-Pareto distribution Akinsete et al. (2008).

Raqab and Kundu (2005) introduced the Rayleigh distribution in connection with a problem in the field of acoustics. An important characteristic of the Rayleigh distribution is that its hazard function is an increasing function of time. It means that when the failure times are distributed according to the Rayleigh law, an intense aging of the item takes place. Estimations, predictions and inferential issues for one parameter Rayleigh distribution have been extensively studied by several authors. Rayleigh distributions are useful in modeling and predicting tools in a wide variety of socio-economic contexts. The Rayleigh distribution has a wide range of applications including life testing experiments, reliability analysis, applied statistics and clinical studies. Potdar and Shirke (2013) have provided reliability estimation for the distribution of a k-unit parallel system with Rayleigh distribution as the component life distribution.

In many applied sciences such as medicine, engineering and finance amongst others, modeling and analyzing lifetime data are crucial. Several life time distributions have been used to model such kinds of data. The quality of the procedures used in a statistical analysis depends heavily on the assumed probability model or distributions. Because of this, considerable effort has been expended in the development of large classes of standard probability distributions along with relevant statistical methodologies. However there still remains many important problems, where the real data does not follow any of the classical or standard probability models. Pareto-Rayleigh is an example of Transformed-Transformer family (T-X family) of distributions, defined by Alzaatreh et al. (2013b). Also Alzaatreh et al. (2012) and Alzaatreh et al. (2013a) derived Gamma-Pareto distribution, Weibull-Pareto distribution and its applications.

In the present work, our focus is to provide confidence intervals and tolerance intervals based on maximum likelihood estimator (MLE) and modified maximum likelihood estimator (MMLE) of the parameter of Pareto-Rayleigh distribution. MLE in the present case is not available in the closed form and is to be obtained by using a suitable iterative method. Tiku (1967) obtained modified maximum likelihood (MML) equations which have explicit solutions by replacing the intractable terms by their linear approximations. Tiku and Suresh (1992) used the Taylor series expansion of the intractable terms in estimating the location and scale parameters in a symmetric family of distributions, which includes a number of well-known distributions such as normal, Students t etc. They also showed that the MML estimators, thus derived are asymptotically fully efficient for small samples. One may refer to Vaughan (1992), Suresh (1997) and Tiku (1967, 1968) for more details. In this article we use MLE and MML estimator to construct CIs and TIs.

A (β,1-γ) TI based on a sample is constructed so that it would include at least a proportion β of the sampled population with confidence 1-γ. Such a TI is usually referred to as β-content-(1-γ) coverage TI or simply (β, 1-γ) TI. A (β, 1-γ) upper tolerance limit (TL) is simply an (1-γ)th upper confidence limit for the (100γ)th percentile of the population and a (β, 1-γ) lower TL is an (1-γ)th lower confidence limit for the (100(1-γ))th percentile of the population. In this article, we are mainly concerned with

one-sided TI using large sample (LS) approach and generalized variable (GV) approach for Pareto-Rayleigh distribution. Kumbhar and Shirke (2004) described TIs for lifetime distribution of k-unit parallel system, when component lifetime distribution is exponential. Liao et al. (2005) have proposed a method for constructing TIs in one-way random model based on the GV approach due to Weerahandi (1993).

Concept of GV has recently become popular in small sample inferences for complex problems such as Behrens-Fisher problem. These techniques have been shown to be efficient in specific distributions by using MLEs. The GV method was motivated by the fact that the small sample optimal CIs in statistical problems involving nuisance parameters may not be available. The method of generalized confidence interval (GCI) based on GV is used whenever standard pivotal quantities either do not exist or are difficult to obtain. Weerahandi (1993) introduced the concept of GCI. As described in the cited papers, GCI is based on the so-called generalized pivotal quantity (GPQ). For some problems, where the classical procedures are not optimal, GCI performs well. Krishnamoorthy and Mathew (2003) developed exact CI and tests for single lognormal mean using ideas of generalized p-values and GCIs. Guo and Krishnamoorthy (2005) explained a problem of interval estimation and testing for the difference between the quantiles of two populations using GV approach. Krishnamoorthy et al. (2006) explained generalized p-values and CIs with a novel approach for analyzing lognormal distributed exposure data. Krishnamoorthy et al. (2007) explained a problem of hypothesis testing and interval estimation of the reliability parameter in a stress-strength model involving two-parameter exponential distribution using GV approach. Verrill and Johnson (2007) considered confidence bounds and hypothesis tests for coefficient of variation of normal distribution. Kurian et al. (2008) have provided GCI for process capability indices in one-way random model. Krishnamoorthy and Lian (2012) derived generalized TIs for some general linear models based on GV approach. The literature survey reveals that during last ten years number of researchers have reported inference for the well known models using GV approach, which motivated us to consider the problem of generalized CI and generalized TI for Pareto-Rayleigh distribution. Rest of the paper is organized as follows.

In Section 2, the Pareto-Rayleigh distribution is considered and MLE and MMLE of the scale parameter are obtained. Section 3, provides CIs based on MLE and MMLE using LS procedure and GV approach. Section 4, provides TIs using LS procedure and GV approach. In section 5, the performance of the CIs and TIs using LS and GV approaches based on MLE and MMLE for small samples is investigated using simulations. Results of the simulation study have been reported in same section. In section 6, a real data set has been analyzed as an illustration.

2 Model and estimation of the scale parameter

Let F(.) be the cumulative distribution function (cdf) of any random variable X defined on $[0,\infty)$ and f (.) be the probability density function (pdf) of a random variable

T, defined on [0,∞). The cdf of the T-X family of distributions defined by Alzaatreh et al. (2013b) is given by

$$G(x) = \int_0^{-log(1-F(x))} f(t)dt \tag{1}$$

Alzaatreh et al. (2013b) named this family of distributions the Transformed-Transformer family (or T-X family) of distributions. If a random variable T follows the Pareto distribution type IV with parameter α then pdf of T is given by,

$$f(t) = \alpha(1+t)^{-(\alpha+1)} \qquad t > 0, \alpha > 1 \tag{2}$$

If a random variable X follows the Rayleigh distribution with parameter σ then cdf of X is given by,

$$F(x) = 1 - \exp(-x^2/2\sigma^2) \qquad \sigma > 0, x > 0 \tag{3}$$

Using (1), (2) and (3), the cdf of Pareto-Rayleigh distribution (as a member of T-X family) is given by,

$$G(x) = \int_0^{x^2/2\sigma^2} \alpha(1+t)^{(-\alpha+1)}dt = 1 - \left(1 + \frac{x^2}{2\sigma^2}\right)^{-\alpha} \qquad x > 0, \alpha > 1, \sigma > 0 \tag{4}$$

The pdf of Pareto-Rayleigh distribution is given by,

$$g(x) = \frac{\alpha}{\sigma^2} x \left(1 + \frac{x^2}{2\sigma^2}\right)^{(-\alpha-1)} \qquad x > 0, \alpha > 1, \sigma > 0, \tag{5}$$

where α is the known shape parameter and σ is the unknown scale parameter. In this article, we are mainly concerned with CIs and TIs of Pareto-Rayleigh distribution using MLE and MMLE of the scale parameter σ .

2.1 Maximum Likelihood Estimation

The pdf of the Pareto-Rayleigh distribution with scale parameter σ and shape parameter α is given by (5).
Let $X_1, X_2, ..., X_n$ be a random sample of size n obtained from Pareto-Rayleigh distribution. By taking the derivative of log likelihood equation, the MLE of the scale parameter σ is the solution of the following equation.

$$\frac{\partial lnL}{\partial \sigma} = 0 = -2n + \frac{\alpha+1}{\sigma^2} \sum_{i=1}^{n} \frac{x_i^2}{(1 + \frac{x_i^2}{2\sigma^2})} = 0.$$

This equation shows that maximum likelihood estimator of $\sigma(\hat{\sigma_n})$ is an iterative solution which can be obtained by suitable iterative method like bisection method. Then Fisher information about σ is given by

$$I = -nE\left[\frac{\partial^2 lnf(x,\alpha,\sigma)}{\partial \sigma^2}\right] = \frac{2n(3\alpha+2)}{\sigma^2(\alpha+2)} - \frac{2n}{\sigma^2}$$

2.2 Modified Maximum Likelihood Estimation

We have seen that MLE of scale parameter σ is not in the closed form as the likelihood equation is intractable. To overcome this difficulty, we use MML method of estimation (Tiku and Suresh (1992)) to find the estimate of scale parameter σ. This can be done by first expressing the maximum likelihood equation in terms of order statistics and then replacing the intractable terms by their linear approximation.
Maximum likelihood equation can be written as

$$\frac{\partial lnL}{\partial \sigma} = 0 = -2n + (\alpha+1)\sum_{i=1}^{n} \frac{z_i^2}{1+\frac{z_i^2}{2}} = 0 \tag{6}$$

where

$$z_i = \frac{x_i}{\sigma}.$$

The maximum likelihood equation (6) does not have explicit solution for scale parameter σ. This is due to the fact that the term

$$g(z_i) = \frac{z_i^2}{1+\frac{z_i^2}{2}}$$

is intractable. To formulate MML equation, which has explicit solution, we express this equation in terms of order statistics that is

$$\frac{\partial lnL}{\partial \sigma} = 0 = -2n + (\alpha+1)\sum_{i=1}^{n} \frac{z_{(i)}^2}{1+\frac{z_{(i)}^2}{2}} = 0 \tag{7}$$

where $z_{(i)}$ is the order statistic of the sample observations x_i,(i=1,2,...,n). The second step is to linearize equation (7) by using Taylor series expansion around the quantile point of G. The linearization is done in such a way that the derived MMLE retains all the desirable asymptotic properties of the MLEs. Thus we have,

$$g(z_{(i)}) = \frac{z_{(i)}^2}{1+\frac{z_{(i)}^2}{2}} = a_i + b_i z_{(i)} \tag{8}$$

The third step is to obtain the modified maximum likelihood equation by incorporating (8) in (7), that is

$$\frac{\partial lnL}{\partial \sigma} = 0 = \frac{\partial lnL^*}{\partial \sigma} = -2n + (\alpha+1)\sum(a_i + b_i z_{(i)}) \tag{9}$$

The solution to equation (9) is the MMLE, which is given by

$$\hat{\sigma} = \frac{\sum b_i x_{(i)}}{\frac{n}{\alpha+1} - \sum a_i} \tag{10}$$

where

$$b_i = g^{'}(z_{(i)}), a_i = g(z_{(i)}) - b_i z_{(i)}$$

One may refer to Tiku and Suresh (1992) and Suresh (2004) for more details.
In the following, we shall see two methods of finding confidence intervals for scale parameter σ using MLE and MMLE.
Lemma 2.1: Distribution of $(\frac{\hat{\sigma}_n}{\sigma})$ and $(\frac{\hat{\sigma}}{\sigma})$, both are free from σ where $\hat{\sigma}_n$ is MLE and $\hat{\sigma}$ is MMLE of σ.
Proof: The proof is similar to the one given by Gulati and Mi (2006). This lemma can be used to find GPQ.

3 Confidence Intervals

3.1 Large sample confidence interval

Theorem: As $n \to \infty$,

$$\sqrt{n}(\hat{\sigma} - \sigma) \longrightarrow N_2(0, I^{-1})$$

where I is the Fisher information given in section (2.1).
Proof: Proof follows from asymptotic properties of MLEs under regularity conditions. Since σ is unknown, I is estimated by replcing σ by its MLE or MMLE and this can be used to obtain the asymptotic CI of σ.
The approximate $100(1-\tau)\%$ asymptotic confidence interval (ACI) for σ is given by

$$\left(\hat{\sigma} \pm z_{1-\tau/2}\sqrt{\frac{I^{-1}}{n}}\right) \tag{11}$$

where $z_{1-\tau/2}$ is the $(1-\tau/2)^{th}$ quantile of the standard normal distribution.

According to Tiku and Suresh (1992) the derived MMLEs retain all the desirable asymptotic properties of the MLEs. Hence simply by replacing MLEs with MMLEs we can obtain confidence interval using large sample approach based on MMLE. We denote this interval by I_1.

3.2 Generalized variable approach

The concept of a generalized confidence interval is due to Weerahandi (1993). One may also refer to Weerahandi (2013) for a detailed discussion along with numerous examples. Consider a random variable X (scalar or vector) whose distribution $g(x, \sigma, \delta)$ depends on a scalar parameter of interest σ and a nuisance parameter (parameter that is not of direct inferential interest)δ, where δ could be a vector. Suppose we are interested in computing a confidence interval for scale parameter σ. Let, x denotes the observed value of X. To construct a GCI for σ, we first define a GPQ, $T(X; x, \sigma, \delta)$ which is a function of random variable X, its observed data x, the parameters σ and δ. A quantity $T(X; x, \sigma, \delta)$ is required to satisfy the following two conditions.
i) For a fixed x, the probability distribution of $T(X; x, \sigma, \delta)$ is free of unknown parameters σ and δ;
ii) The observed value of $T(X; x, \sigma, \delta)$, namely $T(x; x, \sigma, \delta)$ is simply σ.

The percentiles of $T(X;x,\sigma,\delta)$ can then be used to obtain confidence intervals for σ. Such confidence intervals are referred to as generalized confidence intervals. For example, if $T_{1-\tau}$ denotes the $100_{1-\tau}$ th percentile of $T(X;x,\sigma,\delta)$, then $T_{1-\tau}$ is a generalized upper confidence limit for σ. Therefore $100(1-\tau)\%$ two-sided GCI for parameter σ is given by

$$(T_{\tau/2}, T_{1-\tau/2}).$$

Define GPQ as

$$T_1(X;x,\sigma) = \frac{\hat{\sigma}_o}{\frac{\hat{\sigma}}{\sigma}},$$

where $\hat{\sigma}_o$ is the MLE obtained using observed data. We note the following:
i) Distribution of $T_1(X;x,\sigma)$ is free from σ, which follows from Lemma (2.1) and
ii) $T_1(X;x,\sigma) = \sigma$, since for observed data, $\hat{\sigma} = \hat{\sigma}_o$. A GCI based on $T_1(X;x,\sigma)$ is obtained by using the following algorithm. The GCI is denoted by I_2.

I. Algorithm to obtain GCI for σ using GPQ

1. Input n, N, α, σ, τ.
2. Generate independently and identically distributed observations $(U_1, U_2, ..., U_n)$ from U(0,1).
3. For the given value of the parameter σ , set

$$x_i = \sqrt{2\sigma^2((1-U_i)^{-1/\alpha} - 1)} \qquad for \quad i = 1, 2, ..., n.$$

Then $(x_1, x_2, ..., x_n)$ is random sample of size n from Pareto-Rayleigh distribution with parameter σ.
4. Based on observations in step 3, obtain MLE of σ (say $\hat{\sigma}_o$), using bisection method.
5. Generate random sample of size n from Pareto-Rayleigh distribution with parameter σ=1.
6. Based on observations in step 5, obtain MLE of σ (say $\hat{\sigma}$)using bisection method.
7. Compute GPQ, $T_1 = \frac{\hat{\sigma}_o}{\hat{\sigma}}$
8. Repeat steps (5) to (7) N times, so as to get $T_{11}, T_{12}, ..., T_{1N}$.
9. Arrange T_{1i}^s in an ascending order. Denote them by $T_{(11)}, T_{(12)}, ..., T_{(1N)}$.
10. Compute a $100(1-\tau)\%$ GCI for σ as $(T_{(1,([(\tau_2)N]))}, T_{(1,([(1-\tau_2)N]))})$.

Extending above algorithm one can estimate coverage probability of the proposed GCI. In the above algorithm, we can replace MLE by MMLE and obtain GCI based on MMLE.

4 Tolerance Intervals

4.1 Large Sample Tolerance Intervals

There are two types of tolerance intervals namely β-expectation tolerance interval (TI) and β-content-(1-γ) coverage tolerance interval.

4.1.1 β-expectation TI for the distribution function G (.; σ)

Let $X_\beta(\sigma)$ be the lower quantile of order β of the distribution function G (.; σ). Then, we have

$$X_\beta(\sigma) = \sqrt{2\sigma^2\{(1-\beta)^{-1/\alpha} - 1\}}$$

Since σ is unknown, we replace it by its MLE. Hence maximum likelihood estimate of $X_\beta(\sigma)$ is given by

$$X_\beta(\hat{\sigma}) = \sqrt{2\hat{\sigma}^2\{(1-\beta)^{-1/\alpha} - 1\}} \tag{12}$$

having an approximate upper β-expectation TI for G (.; σ) as

$$J_1(X) = (0, X_\beta(\hat{\sigma})) \tag{13}$$

We approximate $E[G(X_\beta(\sigma); \sigma)]$ using Atwood (1984) and is given as

$$E[G(X_\beta(\hat{\sigma}); \sigma)] \approx \beta - 0.5F_{02}Var(\hat{\sigma}) + \frac{F_{01}Var(\hat{\sigma})F_{11}}{F_{10}} \tag{14}$$

where $F_{10} = \frac{\partial G(x;\sigma)}{\partial x}$, $F_{01} = \frac{\partial G(x;\sigma)}{\partial \sigma}$, $F_{11} = \frac{\partial^2 G(x;\sigma)}{\partial x \partial \sigma}$, $F_{02} = \frac{\partial^2 G(x;\sigma)}{\partial \sigma^2}$ with $x = X_\beta(\sigma)$ and all the derivatives are evaluated at X_β and σ. We can replace MLE by MMLE and obtain β-expectation TI for G (.; σ) based on MMLE. Simulated and approximate values of expected coverage of $J_1(X)$ using MLE and MMLE have been reported in section 5 for different values of n, β and α.

4.1.2 β-content-(1-γ) coverage Tolerance Interval

Let $J_2(X) = (0, D\hat{\sigma})$ be an upper β-content-(1-γ) coverage TI for the distribution having distribution function (4). The constant $D(> 0)$ for $\beta\epsilon(0,1)$, $\gamma\epsilon(0,1)$ is to be determined such that

$$P\{G(D\hat{\sigma}; \sigma) \leq \beta\} = 1 - \gamma$$

That is

$$P\left\{\hat{\sigma} \leq \sigma\frac{\sqrt{2\{(1-\beta)^{-1/\alpha} - 1\}}}{D}\right\} = 1 - \gamma \tag{15}$$

Using asymptotic normality of $\hat{\sigma}$ equation (15) can be equivalently written as

$$P\left\{Z \leq (\frac{\sigma}{var(\sigma)})\frac{\sqrt{2\{(1-\beta)^{-1/\alpha} - 1\}}}{D} - 1\right\} = 1 - \gamma,$$

where Z follows N(0,1). This gives

$$D = \frac{\sqrt{2\{(1-\beta)^{-1/\alpha}-1\}}}{1+\frac{var(\sigma)}{\sigma}z_{1-\gamma}}$$

Hence, an upper tolerance limit of β-content-(1-γ) coverage tolerance interval ($J_2(X)$) is given by

$$U(X) = \hat{\sigma}\left\{\frac{\sqrt{2\{(1-\beta)^{-1/\alpha}-1\}}}{1+\frac{var(\sigma)}{\sigma}z_{1-\gamma}}\right\} \tag{16}$$

4.2 Generalized Tolerance Intervals

The problem of computing a one-sided tolerance limit reduces to that of computing a one-sided confidence limit for the percentile of the relevant probability distribution. That is a $(\beta,(1-\gamma))$ upper tolerance limit is simply an (1-γ)th upper confidence limit for the (100β)th percentile of the population. It is easily seen that a $(\beta,(1-\gamma))$ upper tolerance limit for G (.; σ) is simply a 100(1-γ)% upper confidence limit for $\sqrt{2\sigma^2[(1-\beta)^{-1/\alpha}-1]}$. We use the GV approach for obtaining the aforementioned upper confidence limit.
Let $\hat{\sigma}_o$ is the MLE obtained using observed data. The GPQ for constructing a confidence interval for σ is given by $T_1(X;x,\sigma)=\frac{\hat{\sigma}_o}{\hat{\sigma}_i/\sigma}$,i=1,2,...,N. The GPQ for $\sqrt{2\sigma^2[(1-\beta)^{-1/\alpha}-1]}$ is given by

$$T_2 = \frac{\hat{\sigma}_o}{\hat{\sigma}_i/\sigma}\sqrt{2[(1-\beta)^{-1/\alpha}-1]}, \qquad i=1,2,...,N.$$

The $(1-\gamma)$th quantile of T_2 is a $(1-\gamma)$th generalized upper confidence bound for $\sqrt{2\sigma^2[(1-\beta)^{-1/\alpha}-1]}$. Hence $(\beta,(1-\gamma))$ upper tolerance limit for $G(.;\sigma)$ is $(0,T_{2,1-\gamma})$. A generalized tolerance interval based on $T_2(X;x,\sigma)$is obtained by using the following algorithm.

II. Algorithm to obtain Generalized Tolerance Interval for $G(.;\sigma)$ using GPQ

1. Input n, N, $\alpha,\sigma,\beta,\gamma$.
2. Input random sample of size n from Pareto-Rayleigh distribution with an unknown parameter σ.
3. Based on observations in step 2, obtain MLE of σ (say $\hat{\sigma}_o$), using bisection method.
4. Generate random sample of size n from Pareto-Rayleigh distribution with parameter $\sigma=1$.
5. Based on observations in step 4, obtain MLE of σ (say $\hat{\sigma}$), using bisection method.
6. Compute GPQ,
$T_2=\frac{\hat{\sigma}_o}{\hat{\sigma}_i/\sigma}\sqrt{2[(1-\beta)^{-1/\alpha}-1]}, \qquad i=1,2,...,N.$
7. Repeat steps (4) to (6) N times, so as to get $T_{21},T_{22},...,T_{2N}$.
8. Arrange $T'_{2i}s$ in an ascending order. Denote them by $T_{21},T_{22},...,T_{2N}$
9. Compute an upper tolerance limit of generalized TI $J_2(X)=(0,T_{2,1-\gamma})$.

Extending above algorithm one can estimate coverage probability of the proposed generalized TI. In the above algorithm, we can replace MLE by MMLE and obtain generalized TI, based on MMLE.

5 Numerical and simulation study

We conduct extensive simulation experiments to evaluate performance of CIs (LS approach and GV approach) based on MLE and MMLE. We choose different values of σ, β, n and α. Results are tabulated in Tables 1-2. Figures in the 1st row are based on MLE, while figures in the 2nd row are based on MMLE. From Tables 1-2, we observe that simulated coverage of GCI does not differ significantly whether it can be computed from MLE as well as MMLE. However, large sample approach underestimates the coverage probabilities for most of the scenarios, especially when the sample size is small and (or) the parameter σ is large. Also the performance of the proposed GCI does not depend on σ. As the sample size is large, the two estimators (MLE, MMLE) are equally efficient.

We investigate coverage (numerical and simulation) of β-expectation TI for Pareto-Rayleigh distribution with $\alpha = 3$ and β= 0.90, 0.95,0.99 by using MLE and MMLE. Figures in the 1st row are based on MLE, while figures in the 2nd row are based on MMLE. An upper β-expectation tolerance limit is given in equation (12). Results of the simulation study for the β-expectation tolerance interval, which is tabulated in Table 3, indicate that, the estimated expectation and simulation mean for small sample size are marginally lower than the nominal value. As the sample size increases, the performance of tolerance intervals improves. We observe the following from Table 3.
The estimated expectation of the coverage of the approximate β-expectation tolerance intervals shows satisfactory result for large n. Estimated expectation and simulated mean of the coverage increase as sample size n increase. Estimated expectation and simulated mean of the coverage remains same as shape parameter increases. Simulated mean of the coverage for small sample size is below nominal level.
A simulation study of an upper β-content- (1- γ) coverage TI, having an upper limit (16) is also conducted, for σ=1, 2 and for known values of n, β, α and γ. In this simulation study 5000 samples from G (.; σ) were generated and for each of the samples U(X) was computed, for different combinations of β, σ , γ. The proportion of samples for which $\sqrt{2\sigma^2[(1-\beta)^{-1/\alpha}-1]}$ exceeded U(X) was computed 100 times and the mean of these 100 proportions is taken as simulated value of γ. The simulation study for the generalized TI was carried out as algorithm (II). Tables 5-6 give the simulated values of confidence level γ when σ=1, 2 respectively. The proposed confidence interval performs satisfactory for small to moderate sample sizes. These intervals are superior to the asymptotic confidence intervals.

Table 1: Mean coverage of Confidence Intervals (using MLE and MMLE) for transformed transformer (Pareto-Rayleigh) distribution I_1) Large Sample procedure I_2) Generalized variable approach when σ=1.0, α=2.0

coverage	0.90		0.95		0.99	
n	I_1	I_2	I_1	I_2	I_1	I_2
2	0.8604	0.9012	0.8962	0.9445	0.931	0.9887
	0.8652	0.9004	0.8932	0.9434	0.9291	0.9894
3	0.8723	0.9024	0.8931	0.9552	0.9458	0.9947
	0.8651	0.8994	0.9162	0.9558	0.9454	0.9990
4	0.8741	0.9025	0.9041	0.9537	0.9548	0.9889
	0.8735	0.9036	0.9217	0.9534	0.9634	0.9910
5	0.8811	0.9028	0.9147	0.9502	0.9615	0.9963
	0.8879	0.9047	0.9181	0.9532	0.9664	0.9924
6	0.8805	0.9022	0.9251	0.9534	0.9538	0.9917
	0.8898	0.9019	0.9352	0.9564	0.9644	0.9934
7	0.8841	0.9047	0.9284	0.9521	0.9665	0.9937
	0.8897	0.9024	0.9294	0.9588	0.9724	0.9918
8	0.8889	0.9068	0.9281	0.9588	0.9735	0.9919
	0.8962	0.9088	0.9462	0.9531	0.9654	0.9934
9	0.8771	0.9021	0.9354	0.9529	0.9814	0.9935
	0.8981	0.9011	0.9381	0.9574	0.9684	0.9928
10	0.8910	0.9024	0.9474	0.9534	0.9715	0.9915
	0.8907	0.9024	0.9474	0.9538	0.9764	0.9966
15	0.8888	0.9008	0.9364	0.9536	0.9775	0.9919
	0.8946	0.9064	0.9464	0.9587	0.9814	0.9921
30	0.8947	0.9055	0.9484	0.9537	0.9865	0.9926
	0.9014	0.9027	0.9562	0.9564	0.984	0.9987
50	0.8932	0.9064	0.9314	0.9528	0.9845	0.9928
	0.9016	0.9033	0.9414	0.9508	0.9894	0.9980

Table 2: Mean coverage of Confidence Intervals (using MLE and MMLE) for transformed transformer (Pareto-Rayleigh) distribution I_1) Large Sample procedure I_2) Generalized variable approach when σ=2.0, α=2.0

coverage	0.90		0.95		0.99	
n	I_1	I_2	I_1	I_2	I_1	I_2
2	0.8605	0.8992	0.8894	0.9487	0.9312	0.9887
	0.8625	0.8988	0.8905	0.9425	0.9219	0.9805
3	0.8736	0.8989	0.9008	0.9432	0.9448	0.9928
	0.8715	0.9080	0.9020	0.9485	0.9321	0.9865
4	0.8781	0.9030	0.9172	0.9506	0.9504	0.9889
	0.8724	0.9053	0.9251	0.9538	0.9603	0.9932
5	0.8921	0.9021	0.9204	0.9524	0.9614	0.9962
	0.8829	0.9026	0.9148	0.9519	0.9668	0.9937
6	0.8938	0.9062	0.9224	0.9522	0.9534	0.9932
	0.8905	0.9028	0.9321	0.9537	0.9617	0.9919
7	0.8908	0.9081	0.9318	0.9540	0.9624	0.9984
	0.8842	0.9024	0.9304	0.9531	0.9724	0.9941
8	0.8921	0.9061	0.9326	0.9565	0.9735	0.9958
	0.8955	0.9008	0.9428	0.9528	0.9625	0.9935
9	0.8881	0.9073	0.9306	0.9535	0.9814	0.9931
	0.8918	0.9026	0.9325	0.9522	0.9757	0.9984
10	0.8962	0.9083	0.9341	0.9557	0.9795	0.9922
	0.8925	0.9034	0.9487	0.9565	0.9743	0.9957
15	0.8994	0.9043	0.9412	0.9548	0.9724	0.9943
	0.8997	0.9050	0.9427	0.9566	0.9817	0.9980
30	0.8934	0.9018	0.9474	0.9541	0.9887	0.9957
	0.8906	0.9024	0.9438	0.9564	0.9814	0.9972
50	0.8956	0.9028	0.9518	0.9561	0.9822	0.9964
	0.8941	0.9084	0.958	0.9534	0.9878	0.9955

Table 3: Simulated mean and estimated expectation of the coverage of approximate β-expectation TI using MLE and MMLE for transformed transformer (Pareto-Rayleigh) distribution.

				$\alpha = 3$				
			$\beta(\sigma = 1.0)$				$\beta(\sigma$=2.0)	
n	0.90	0.95	0.97	0.99	0.90	0.95	0.97	0.99
2	0.8112	0.8595	0.9065	0.9515	0.8315	0.8712	0.9172	0.9521
	(0.8251)	(0.8459)	(0.8902)	(0.9625)	(0.8451)	(0.8652)	(0.9251)	(0.9534)
	0.0.7921	0.8888	0.9127	0.9318	0.8298	0.8585	0.9275	0.9434
	(0.7912)	(0.8892)	(0.9021)	(0.9425)	(0.8329)	(0.8625)	(0.9265)	(0.9547)
3	0.8568	0.8996	0.9384	0.9592	0.8436	0.9118	0.9418	0.9637
	(0.8495)	(0.8825)	(0.9365)	(0.9469)	(0.8492)	(0.9028)	(0.9356)	(0.9645)
	0.8465	0.9124	0.9386	0.9544	0.8494	0.8917	0.9374	0.9568
	(0.8520)	(0.9062)	(0.9255)	(0.9528)	(0.8574)	(0.9054)	(0.9487)	(0.9534)
4	0.8716	0.9142	0.9499	0.9756	0.8333	0.9014	0.9375	0.9725
	(0.8724)	(0.9028)	(0.9589)	(0.9728)	(0.8365)	(0.9124)	(0.9425)	(0.9824)
	0.0.8588	0.8923	0.9491	0.9693	0.8514	0.9151	0.9438	0.9695
	(0.8459)	(0.8902)	(0.9425)	(0.9714)	(0.8495)	(0.9024)	(0.9457)	(0.9748)
5	0.8632	0.9151	0.9454	0.9737	0.8697	0.9222	0.9537	0.9786
	(0.8794)	(0.9215)	(0.9316)	(0.9722)	(0.8724)	(0.9365)	(0.9633)	(0.9748)
	0.0.8610	0.9244	0.9558	0.9611	0.8712	0.9023	0.9449	0.9659
	(0.8705)	(0.9145)	(0.9420)	(0.9784)	(0.8790)	(0.9124)	(0.9584)	(0.9721)
6	0.8754	0.9359	0.9565	0.9859	0.8725	0.9179	0.9539	0.9791
	(0.8715)	(0.9302)	(0.9536)	(0.9850)	(0.8837)	(0.9274)	(0.9521)	(0.9701)
	0.0.8665	0.9197	0.9523	0.9750	0.8774	0.9178	0.9494	0.9814
	(0.8714)	(0.9028)	(0.9577)	(0.9815)	(0.8791)	(0.9154)	(0.9524)	(0.9825)
7	0.8668	0.9417	0.9647	0.9847	0.8839	0.9346	0.9689	0.9817
	(0.8628)	(0.9459)	(0.9619)	(0.9824)	(0.8829)	(0.9435)	(0.9752)	(0.9932)
	0.0.8577	0.9278	0.9569	0.9794	0.8746	0.9244	0.9516	0.9735
	(0.8459)	(0.9160)	(0.9654)	(0.9728)	(0.8859)	(0.9284)	(0.9654)	(0.9849)
8	0.8889	0.9296	0.9692	0.9859	0.8654	0.9328	0.9614	0.9872
	(0.8749)	(0.9239)	(0.9628)	(0.9822)	(0.8735)	(0.9475)	(0.9672)	(0.9824)
	0.0.8945	0.9344	0.9674	0.9765	0.8747	0.9325	0.9577	0.9840
	(0.8891)	(0.9385)	(0.9587)	(0.9711)	(0.8815)	(0.9425)	(0.9657)	(0.9864)

Table 4: Simulated mean and estimated expectation of the coverage of approximate β-expectation TI using MLE and MMLE for transformed transformer (Pareto-Rayleigh) distribution. Continued

$\alpha = 3$								
	$\beta(\sigma = 1.0)$				$\beta(\sigma$=2.0)			
n	0.90	0.95	0.97	0.99	0.90	0.95	0.97	0.99
9	0.8787	0.9277	0.9616	0.9872	0.8781	0.9342	0.9645	0.9896
	(0.8892)	(0.9258)	(0.9621)	(0.9826)	(0.8724)	(0.9451)	(0.9754)	(0.9833)
	0.0.8914	0.9389	0.9647	0.9813	0.8790	0.9294	0.9592	0.98140
	(0.8928)	(0.9225)	(0.9618)	(0.9837)	(0.8739)	(0.9321)	(0.9625)	(0.9802)
10	0.8856	0.9265	0.9588	0.9885	0.8838	0.9416	0.9671	0.9829
	(0.8821)	(0.9368)	(0.9548)	(0.9814)	(0.8902)	(0.9478)	(0.9784)	(0.9820)
	0.0.8831	0.9314	0.9692	0.9848	0.8765	0.9333	0.9664	0.9817
	(0.8834)	(0.9425)	(0.9664)	(0.9834)	(0.8834)	(0.9401)	(0.9725)	(0.9849)
15	0.8919	0.9346	0.9631	0.9914	0.8769	0.9314	0.9657	0.9885
	(0.8940)	(0.9365)	(0.9748)	(0.9889)	(0.8729)	(0.9365)	(0.9781)	(0.9804)
	0.0.8994	0.9475	0.9715	0.9886	0.8836	0.9379	0.9698	0.9851
	(0.8921)	(0.9428)	(0.9708)	(0.9948)	(0.8924)	(0.9425)	(0.9748)	(0.9834)
30	0.8837	0.9428	0.9779	0.9927	0.8993	0.9517	0.9685	0.9952
	(0.9024)	(0.9458)	(0.9645)	(0.9917)	(0.9028)	(0.9538)	(0.9677)	(0.9889)
	0.0.9016	0.9492	0.9737	0.9879	0.8865	0.9495	0.9769	0.9826
	(0.9099)	(0.9359)	(0.9721)	(0.9950)	(0.8949)	(0.9584)	(0.9780)	(0.9887)
50	0.9028	0.9492	0.9695	0.9987	0.9014	0.9532	0.9746	0.9949
	(0.9082)	(0.9584)	(0.9635)	(0.9980)	(0.9147)	(0.9502)	(0.9722)	(0.9924)
	0.0.9092	0.9514	0.9753	0.9914	0.8916	0.9534	0.9753	0.99140
	(0.9158)	(0.9506)	(0.9748)	(0.9940)	(0.9025)	(0.9524)	(0.9824)	(0.9914)

Table 5: Coverage probabilities of Tolerance Intervals for Pareto-Rayleigh distribution I_1) Large sample procedure I_2) Generalized variable approach σ=1.0, α=2.0

	γ=0.90				γ=0.95			
coverage	β=0.90		β=0.95		β=0.90		β=0.95	
n	I_1	I_2	I_1	I_2	I_1	I_2	I_1	I_2
2	0.6672	0.9021	0.6432	0.8924	0.5549	0.9448	0.5521	0.9449
	0.6451	0.9028	0.6544	0.8920	0.5441	0.9459	0.5549	0.9428
3	0.7984	0.8992	0.7971	0.8935	0.7231	0.9432	0.7461	0.9452
	0.7846	0.9021	0.7869	0.8922	0.7266	0.9458	0.7361	0.9488
4	0.8156	0.9034	0.8194	0.9031	0.8224	0.9538	0.8564	0.9468
	0.8356	0.9038	0.8347	0.8977	0.8319	0.9458	0.8479	0.9585
5	0.8448	0.9049	0.8435	0.9028	0.8815	0.9562	0.8714	0.9562
	0.8544	0.9028	0.8539	0.9024	0.8819	0.9564	0.8854	0.9534
6	0.8639	0.9125	0.8556	0.9034	0.8901	0.9537	0.9032	0.9538
	0.8634	0.9037	0.8619	0.9021	0.9034	0.9538	0.9035	0.9566
7	0.8644	0.9028	0.8598	0.9055	0.8974	0.9539	0.9074	0.9533
	0.8664	0.8997	0.8686	0.9029	0.9096	0.9532	0.9083	0.9580
8	0.8492	0.9064	0.8429	0.9064	0.9087	0.9654	0.9097	0.9582
	0.8706	0.9035	0.8695	0.9068	0.9144	0.9538	0.9157	0.9534
9	0.8493	0.9038	0.8239	0.9031	0.9124	0.9587	0.9015	0.9524
	0.8716	0.9028	0.8714	0.9024	0.9188	0.9458	0.9183	0.9531
10	0.8614	0.9034	0.8497	0.9124	0.9235	0.9482	0.9032	0.9654
	0.8744	0.9046	0.8724	0.8992	0.9203	0.533	0.9240	0.9587
15	0.8718	0.9029	0.8544	0.8997	0.9114	0.9588	0.9225	0.9528
	0.8798	0.9029	0.8792	0.9034	0.9272	0.9526	0.9284	0.9575
30	0.8790	0.9184	0.8831	0.9024	0.9278	0.9537	0.9315	0.9521
	0.8890	0.9088	0.8872	0.9098	0.9352	0.9538	0.9361	0.9648
50	0.9031	0.9028	0.8951	0.9089	0.9445	0.9526	0.9294	0.9588
	0.8924	0.9090	0.8905	0.9044	0.9449	0.9524	0.9482	0.9584

Table 6: Coverage probabilities of Tolerance Intervals for Pareto-Rayleigh distribution I_1) Large sample procedure I_2) Generalized variable approach σ=2.0, α=2.0

	γ=0.90				γ=0.95			
coverage	β=0.90		β=0.95		β=0.90		β=0.95	
n	I_1	I_2	I_1	I_2	I_1	I_2	I_1	I_2
2	0.6431	0.8925	0.6831	0.8902	0.5621	0.9485	0.5331	0.9487
	0.6401	0.8959	0.6598	0.8988	0.5741	0.9458	0.5521	0.9415
3	0.7811	0.8933	0.8032	0.8953	0.7378	0.9428	0.7394	0.9458
	0.7822	0.8954	0.7852	0.8954	0.7451	0.9462	0.7421	0.9402
4	0.8180	0.9024	0.8394	0.8934	0.8584	0.9521	0.8441	0.9567
	0.8370	0.9024	0.8350	0.9028	0.8566	0.9532	0.8504	0.9435
5	0.8334	0.9028	0.8532	0.9028	0.8893	0.9439	0.9012	0.9548
	0.8537	0.9058	0.8569	0.8937	0.8920	0.9531	0.9135	0.9520
6	0.8521	0.9024	0.8421	0.9054	0.9132	0.9511	0.9035	0.9448
	0.8629	0.9034	0.8694	0.9024	0.9230	0.9489	0.9127	0.9537
7	0.8592	0.9022	0.8584	0.9027	0.8894	0.9560	0.9136	0.9580
	0.8645	0.9031	0.8651	0.9037	0.8904	0.9518	0.9198	0.9582
8	0.8754	0.9065	0.8725	0.9013	0.9052	0.9538	0.9158	0.9502
	0.8779	0.9157	0.8633	0.9026	0.9124	0.9588	0.9230	0.9531
9	0.8531	0.9021	0.8649	0.9157	0.9012	0.9528	0.8869	0.9531
	0.8732	0.9055	0.8724	0.9027	0.9124	0.9575	0.8920	0.9565
10	0.8421	0.9128	0.8564	0.9024	0.8954	0.9582	0.9117	0.9521
	0.8724	0.9071	0.8734	0.9147	0.9280	0.9548	0.9228	0.9533
15	0.8621	0.9034	0.8697	0.8948	0.9235	0.9489	0.9174	0.9587
	0.8799	0.9028	0.8788	0.9088	0.9284	0.9521	0.9257	0.9502
30	0.8674	0.9089	0.8587	0.9028	0.9151	0.9568	0.9239	0.9654
	0.8854	0.9021	0.8876	0.9056	0.9329	0.9588	0.9360	0.9536
50	0.8981	0.9080	0.8879	0.9027	0.9294	0.9586	0.9487	0.9537
	0.8952	0.9072	0.8991	0.9076	0.9428	0.9548	0.9510	0.9531

6 Real life Data Analysis

In this section we present a data analysis of the strength data reported by Bader and Priest (1982). It is already observed by Durham and Padgett (1997) that Weibull model does not work well in this case. Surles and Padgett (1998), Surles and Padgett (2001) and Raqab and Kundu (2005) observed that generalized Rayleigh works quite well for this strength data. Also Raqab and Kundu (2005) observed goodness of fit of the three-parameter generalized exponential distribution to this data set based on modified MLEs.

For illustrative purpose we also consider the same transformed data set as considered by Raqab and Kundu (2005), the single fibers of 10 mm in gauge length with sample size 63. Data set is presented below:
0.101,0.332,0.403,0.428,0.457,0.550,0.561,0.596,0.597,0.645,0.654,0.674,0.718,0.722,
0.725,0.732,0.775,0.814,0.816,0.818,0.824,0.859,0.875,0.938,0.940,1.056,1.117,1.128,
1.137,1.137,1.177,1.196,1.230,1.325,1.339,1.345,1.420,1.423,1.435,1.443,1.464,1.472,
1.494,1.532,1.546,1.577,1.608,1.635,1.693,1.701,1.737,1.754,1.762,1.828,2.052,2.071,
2.086,2.171,2.224,2.227,2.425,2.295,3.220.

First we would like to compute the MLEs of the unknown parameters. The MLE of σ is obtained as 2.036426 and the MLE of α becomes 5.036467 with the associated log-likelihood value as -57.67675. We plot the empirical survival function and the fitted survival function. We used the Kolmogorov-Smirnov (K-S) test for this data set. K-S distance between the fitted Pareto-Rayleigh and empirical cumulative distribution function is 0.094377 and the associated p-value is 0.8431. Therefore, it indicates that the Pareto-Rayleigh model provides reasonable fit to this data set.

Based on the estimates of α and σ, the confidence intervals (using LS and GV approach) are given in the Table 7.

Table 7: Confidence intervals (using LS and GV approach) for strength data.

Coverage	Using Estimator	Using LS approach(ACI)	Using GV approach(GCI)
90%	MLE	(1.787754,2.285098) Length=0.4973437	(1.914382,2.197766) Length=0.283384
	MMLE	(1.786463,2.283543) Length=0.4970807	(1.402121,1.805491) Length=0.4033698
95%	MLE	(1.711940,2.360913) Length=0.6489728	(1.893224,2.246205) Length=0.3496253
	MMLE	(1.737967,2.332039) Length=0.594072	(1.366712,1.836193) Length=0.4694808
99%	MLE	(1.645223,2.427629) Length=0.7824065	(1.852753,2.313607) Length=0.4608534
	MMLE	(1.644007,2.425999) Length=0.7819927	(1.884905,2.698532) Length=0.713627

Therefore, in this case it is clear that the GV approach provides confidence interval having shortest length than the LS approach.
We also evaluated (0.90, 0.90) and (0.95, 0.95) upper tolerance limits for this data set using LS and GV approach. They are 2.921123 (2.875510) and 3.694097(3.56269) respectively. Bracketed tolerance limit is using GV approach.

Table 8: The maximum likelihood estimates and Kolmogorov-Smirnov statistics and p-values for strength data.

The model	MLEs of the parameters	Log-likelihood	K-S statistic	p-value
Generalized Rayleigh	$\hat{\beta}$=1.4216,$\hat{\lambda}$=0.8598	−50.22	0.12	0.2845
Three parameter GE	$\hat{\beta}$=4.3586,$\hat{\lambda}$=1.8303,$\hat{\alpha}$=6.5469	−110.01	0.0933	0.643
Pareto- Rayleigh	$\hat{\alpha}$=5.036467,$\hat{\sigma}$=2.036426	−57.67675	0.094377	0.8431

It is clear from the Table 8 that based on the K-S statistic, the proposed Pareto-Rayleigh model provides a better fit than generalized Rayleigh and three parameter generalized Exponential models to this specific data set. Although, it is not guaranteed that the proposed model always provides a better fit than the other models.

7 Conclusions

In this paper we have considered interval estimation (confidence interval and tolerance interval) using maximum likelihood estimator and modified maximum likelihood estimator in Pareto-Rayleigh distribution (Transformed-Transformer family) based on generalized variable approach. We have compared these generalized intervals with asymptotic intervals. The proposed confidence intervals perform satisfactory for small to moderate sample sizes. These intervals are superior to the asymptotic intervals. The performance of the interval estimation using modified maximum likelihood estimators are also quite satisfactory. One real data analysis has been performed and it is observed that the proposed model provides a better fit than some of the existing models.

8 Acknowledgement

We thank referee for the valuable comments and suggestions. The second and third authors would like to acknowledge the support of University Grants Commission, New

Delhi under Special Assistance Programme to carry out the research work.

References

Akinsete, A., Famoye, F., and Lee, C. (2008). The beta-pareto distribution. *Statistics*, 42(6):547–563.

Alzaatreh, A., Famoye, F., and Lee, C. (2012). Gamma-pareto distribution and its applications. *Journal of Modern Applied Statistical Methods*, 11(1):7.

Alzaatreh, A., Famoye, F., and Lee, C. (2013a). Weibull-pareto distribution and its applications. *Communications in Statistics-Theory and Methods*, 42(9):1673–1691.

Alzaatreh, A., Lee, C., and Famoye, F. (2013b). A new method for generating families of continuous distributions. *Metron*, 71(1):63–79.

Atwood, C. L. (1984). Approximate tolerance intervals, based on maximum likelihood estimates. *Journal of the American Statistical Association*, 79(386):459–465.

Bader, M. and Priest, A. (1982). Statistical aspects of fibre and bundle strength in hybrid composites. *Progress in science and engineering of composites*, pages 1129–1136.

Durham, S. and Padgett, W. (1997). Cumulative damage models for system failure with application to carbon fibers and composites. *Technometrics*, 39(1):34–44.

Gulati, S. and Mi, J. (2006). Testing for scale families using total variation distance. *Journal of Statistical Computation and Simulation*, 76(9):773–792.

Guo, H. and Krishnamoorthy, K. (2005). Comparison between two quantiles: The normal and exponential cases. *Communications in StatisticsSimulation and Computation®*, 34(2):243–252.

Krishnamoorthy, K. and Lian, X. (2012). Closed-form approximate tolerance intervals for some general linear models and comparison studies. *Journal of Statistical Computation and Simulation*, 82(4):547–563.

Krishnamoorthy, K. and Mathew, T. (2003). Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. *Journal of statistical planning and inference*, 115(1):103–121.

Krishnamoorthy, K., Mathew, T., and Ramachandran, G. (2006). Generalized p-values and confidence intervals: A novel approach for analyzing lognormally distributed exposure data. *Journal of occupational and environmental hygiene*, 3(11):642–650.

Krishnamoorthy, K., Mukherjee, S., and Guo, H. (2007). Inference on reliability in two-parameter exponential stress–strength model. *Metrika*, 65(3):261–273.

Kumbhar, R. and Shirke, D. (2004). Tolerance limits for lifetime distribution of k-unit parallel system. *Journal of Statistical Computation and Simulation*, 74(3):201–213.

Kurian, K., Mathew, T., and Sebastian, G. (2008). Generalized confidence intervals for process capability indices in the one-way random model. *Metrika*, 67(1):83–92.

Liao, C., Lin, T., and Iyer, H. (2005). One-and two-sided tolerance intervals for general balanced mixed models and unbalanced one-way random models. *Technometrics*, 47(3):323–335.

Mahmoudi, E. (2011). The beta generalized pareto distribution with application to lifetime data. *Mathematics and computers in Simulation*, 81(11):2414–2430.

Potdar, K. and Shirke, D. (2013). Reliability estimation for the distribution of a k-unit parallel system with rayleigh distribution as the component life distribution. In *International Journal of Engineering Research and Technology*, volume 2. ESRSA Publications.

Raqab, M. Z. and Kundu, D. (2005). Comparison of different estimators of p [y¡ x] for a scaled burr type x distribution. *Communications in StatisticsSimulation and Computation®*, 34(2):465–483.

Schroeder, B., Damouras, S., and Gill, P. (2010). Understanding latent sector errors and how to protect against them. *ACM Transactions on storage (TOS)*, 6(3):9.

Suresh, R. (1997). On approximate likelihood estimators in censored normal samples. *Gujarat Statistical Review*, 24:21–28.

Suresh, R. (2004). Estimation of location and scale parameters in a two-parameter exponential distribution from a censored sample.

Surles, J. and Padgett, W. (1998). Inference for p (y¡ x) in the burr type x model. *Journal of Applied Statistical Science*, 7(4):225–238.

Surles, J. and Padgett, W. (2001). Inference for reliability and stress-strength for a scaled burr type x distribution. *Lifetime Data Analysis*, 7(2):187–200.

Tiku, M. (1967). Estimating the mean and standard deviation from a censored normal sample. *Biometrika*, 54(1-2):155–165.

Tiku, M. (1968). Estimating the parameters of normal and logistic distributions from censored samples. *Australian & New Zealand Journal of Statistics*, 10(2):64–74.

Tiku, M. and Suresh, R. (1992). A new method of estimation for location and scale parameters. *Journal of Statistical Planning and Inference*, 30(2):281–292.

Vaughan, D. C. (1992). On the tiku-suresh method of estimation. *Communications in Statistics-theory and Methods*, 21(2):451–469.

Verrill, S. and Johnson, R. A. (2007). Confidence bounds and hypothesis tests for normal distribution coefficients of variation. *Communications in StatisticsTheory and Methods*, 36(12):2187–2206.

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association*, 88(423):899–905.

Weerahandi, S. (2013). *Exact statistical methods for data analysis*. Springer Science & Business Media.

Journal

Communications in Statistics - Theory and Methods ›

Volume 46, 2017 - Issue 24

Original Articles

A new test for two-sample location problem based on empirical distribution function

S. K. Mathur & D. M. Sakate

Pages 12345-12355 | Received 05 Jul 2016, Accepted 09 Feb 2017, Accepted author version posted online: 21 Feb 2017, Published online: 31 Aug 2017

Download citation https://doi.org/10.1080/03610926.2017.1295158

ABSTRACT

We propose a new test for testing the equality of location parameter of two populations based on empirical distribution function (ECDF). The test statistics is obtained as a power divergence between two ECDFs. The test is shown to be distribution free, and its null distribution is obtained. We conducted empirical power comparison of the proposed test with several other available tests in the literature. We found that the proposed test performs better than its competitors considered here under several population structures. We also used two real datasets to illustrate the procedure.