Analysis of Middle Censored Data with Exponential Lifetime Distributions

Srikanth K. Iyer[†]
Debasis Kundu[*]
S. Rao Jammalamadaka[‡]

Abstract

Recently Jammalamadaka and Mangalam (2003) introduced a general censoring scheme called the "middle-censoring" scheme in non-parametric set up. In this paper we consider this middle censoring scheme when the lifetime distribution of the items are exponentially distributed and the censoring mechanism is independent and non-informative. In this set up, we derive the maximum likelihood estimator and study its consistency and asymptotic normality properties. We also derive the Bayes estimate of the exponential parameter under a gamma prior. Since a theoretical construction of the credible interval becomes quite difficult, we propose and implement Gibbs sampling technique to construct the credible intervals. Monte Carlo simulations are performed to evaluate the small sample behavior of the techniques proposed. A real data set is analyzed to illustrate the practical application of the proposed methods.

Keywords and Phrases: Exponential distribution, Middle censoring, Consistency, Asymptotic Normality, Fixed point solution, HPD credible sets, Bayes estimate.

Postal address: [†] Department of Mathematics, Indian Institute of Science, Bangalore, Pin 560012, INDIA.

Corresponding author: [*]Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, Pin 208016, INDIA; Phone: 91-512-2597141, Fax: 91-512-2597500; e-mail: kundu@iitk.ac.in.

Postal address: [‡] Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106-3110, USA.

1 Introduction

In this paper we analyze lifetime data when they are "middle-censored". Middle censoring occurs if a data point is not observable when it falls inside a random interval. The middle censoring scheme can be described as follows. Suppose n identical items are put on test and the life times of these items are $T_1, \ldots, T_n$. For the i^{th} item, there is a random censoring interval (L_i, R_i), which follows some unknown bivariate distribution. For the i^{th} item, T_i is observable only if $T_i \notin [L_i, R_i]$, otherwise it is not observable. Suppose $\delta_i = I(T_i \notin [L_i, R_i])$, where $I(\cdot)$ denotes the indicator function. Therefore, when $\delta_i = 1$, the observation is not censored and we observe the actual value T_i. In this case we do not observe (L_i, R_i). On the other hand, when $\delta_i = 0$, we observe only the censoring interval $[L_i, R_i]$. For the i^{th} item, we observe the following;

$$(Y_i, \delta_i) = \begin{cases} (T_i, 1) & \text{if } T_i \notin [L_i, R_i] \\ ([L_i, R_i], 0) & \text{otherwise.} \end{cases} \tag{1}$$

Thus, the data obtained here is not the same as that obtained in the interval censoring case. Based on the observations, the problem is to estimate the lifetime distribution functions of T_i's and develop necessary inferential procedures.

The middle censoring scheme was first introduced by Jammalamadaka and Mangalam (2003) under a non-parametric set up. It is an important variation and also a generalization of the existing left censoring, right censoring and double censoring schemes. All the above three censoring schemes can be obtained as special cases of this middle censoring scheme by suitably choosing censoring intervals, which can be infinite. At first glance, middle censoring, where a random middle part is missing, appears as complementary to the idea of double censoring in which the middle part is what is actually observed. However, a careful reflection and analysis shows them to be quite different ideas; see Jammalamadaka and Mangalam (2003) for details.

Before getting into technical details, we mention a few situations where middle censoring occurs. In any lifetime study if the subject is temporarily withdrawn from the study (eg. an individual leaves town for a temporary period and returns, if still alive), we obtain this middle censoring situation. Middle-censoring also occurs when the measuring equipment breaks down for a temporary period or if the clinic where the observations are being taken, is closed for a period, due to an external emergency such as the outbreak of war or a strike. In such cases the event of interest (or failure) could take place during the period when an observation is not possible or is not being made.

In Jammalamadaka and Mangalam (2003), $T_1, \ldots, T_n$ are taken to be independent and identically distributed (*i.i.d.*) random variables with some unknown distribution function $F(\cdot)$. Also, $(L_1, R_1), \ldots, (L_n, R_n)$ are *i.i.d.* with some unknown bivariate distribution function $G(\cdot, \cdot)$ and they are independent of T_i. Based on this, they obtain the non-parametric maximum likelihood estimator of the unknown distribution function $F(\cdot)$ and show that it is a *self-consistent* estimator under the condition that one of the ends is non-random (see the review article of Tarpey and Flury (1996) for a nice account of the *self-consistent* estimators).

In this paper we consider a parametric formulation of the problem. It is assumed that $T_1, \ldots, T_n$ are *i.i.d.* exponential random variables with mean $\frac{1}{\theta_0}$ *i.e.* with the probability density function (PDF) given below;

$$f(x; \theta_0) = \begin{cases} \theta_0 e^{-\theta_0 x}, & x > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Moreover, $(L_1, Z_1), \ldots, (L_n, Z_n)$ are *i.i.d.* where L_i and $Z_i = R_i - L_i$ are independent exponential random variables and they are independent of T_i. It is also assumed that L_i and Z_i have means $\frac{1}{\alpha}$ and $\frac{1}{\beta}$ respectively and they do not depend on θ_0. It implies that the censoring mechanism is independent of the lifetime of the population of interest and has no information on this lifetime.

Such assumptions as the independence, are very standard in the lifetime data analysis. See for example Kaplan and Meier (1958), Turnbull (1974), Babu, Rao and Rao (1992), Jammalamadaka and Mangalam (2003), Jammalamadaka and Iyer (2004) and the references cited there, who make this assumption for a variety of censoring schemes. There are several reasons for that and we mention a couple. First of all, in most of the real life situations it is unlikely that the censoring mechanism depends on the lifetime of the population and that it provides any information on the population distribution function. In all the examples we gave earlier, middle censoring occurs because of an external cause that does not have anything to do with the "life-times". There are a few papers in the literature (see eg. Robertson and Uppuluri (1984)) which discuss non-parametric estimation of the lifetime distribution in the case when the lifetimes and the censoring intervals are dependent. Recently Hongyu *et. al.* (2005) consider the problem of right censoring in a semi-parametric model in which the dependence between the censoring mechanism and the lifetimes is modeled via a gamma frailty copula. In the parametric set-up that we consider, an analytically tractable model for dependence between the lifetimes and the censoring intervals has to be formulated before the estimation questions are tackled. The authors hope to address this question in a future paper.

Based on the above assumptions we obtain different estimators of θ_0 and study their properties. We provide the maximum likelihood estimator (MLE) of θ_0. It is observed that the MLE can not be obtained in a closed form. We propose a simple iterative procedure for finding the MLE and the sufficient condition for the convergence of the iterative method is also provided. We also suggest the EM algorithm which can be used to compute the MLE and provide sufficient condition for its convergence. It is shown that the MLE of θ_0 is consistent and asymptotically normal. As might be expected, the asymptotic variance of MLE of θ_0 depends on the censoring parameters α and β. Thus for constructing asymptotic confidence intervals for θ_0 we use the empirical Fisher information matrix.

We also compute the Bayes estimate of θ_0 under the assumption of Gamma prior distribution on θ_0. No prior distributions on the censoring parameters are assumed. Moreover, the censoring is assumed to be non-informative. After noting that the exact Bayes estimate is difficult to compute in this case, we propose to use the Gibbs sampling procedure to compute the Bayes estimate as well as the highest posterior density (HPD) credible interval of θ_0.

The rest of the paper is organized as follows. In Section 2, we provide the MLE and the proposed EM algorithm followed by the theoretical results in Section 3. The Bayesian formulation and the simulation results are presented in Sections 4 and 5 respectively. An illustrative data analysis and results are given in Section 6 and conclusions in Section 7.

2 Maximum Likelihood Estimator

After re-ordering the data as necessary, we can assume without loss of generality, that the first n_1 and the rest n_2 are the uncensored and censored observations respectively. Therefore, we have the following observed data:

$$\{(T_1, 1), \ldots, (T_{n_1}, 1), (L_{n_1+1}, R_{n_1+1}), \ldots, (L_{n_1+n_2}, R_{n_1+n_2})\}, \tag{3}$$

where $n_1 + n_2 = n$. Thus, $T_i \notin (L_i, R_i)$ for the first n_1 observations, while $T_i \in (L_i, R_i)$ for the last n_2 observations. Based on the above information the likelihood function of the observed data is given by

$$l(\theta) = c\theta^{n_1} e^{-\theta \sum_{i=1}^{n_1} t_i} \prod_{i=n_1+1}^{n_1+n_2} \left(e^{-\theta l_i} - e^{-\theta r_i}\right), \tag{4}$$

where c is the normalizing constant which depends on α and β. Since we are not interested in estimating α and β, we are not making it explicit. Based on (4), the log-likelihood becomes

$$\ln l(\theta) = L(\theta) = \ln c + n_1 \ln \theta - \theta \sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} \ln \left(e^{-\theta l_i} - e^{-\theta r_i}\right). \tag{5}$$

Taking the derivative of $L(\theta)$ and setting it equal to 0, we obtain

$$\frac{\partial L}{\partial \theta} = \frac{n_1}{\theta} - \sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} \frac{(r_i - l_i)}{e^{\theta(r_i-l_i)} - 1} - \sum_{i=n_1+1}^{n_1+n_2} l_i = 0. \tag{6}$$

Therefore, $\hat{\theta}$, the MLE of θ, can be obtained by solving equation (6). Since (6) does not admit an explicit solution, we provide an iterative procedure to solve for the MLE. Note that (6) can be written as

$$h(\theta) = \theta, \tag{7}$$

where

$$h(\theta) = \frac{1}{\sum_{i=n_1+1}^{n_1+n_2} l_i + \sum_{i=1}^{n_1} t_i} \left[n_1 + \theta \sum_{i=n_1+1}^{n_1+n_2} \frac{z_i e^{-\theta z_i}}{1 - e^{-\theta z_i}} \right]. \tag{8}$$

Therefore, a simple iterative procedure can be used to solve (7). For example. we can start with an initial guess $\theta^{(1)}$, then obtain $\theta^{(2)} = h(\theta^{(1)})$ and so on. The iterative procedure may be stopped if $|\theta^{(i)} - \theta^{(i+1)}| < \epsilon$, where ϵ is some preassigned small positive number. For an initial choice of θ, we can use $\theta^{(1)} = n_1 / \sum_{i=1}^{n_1} t_i$.

Alternatively, the EM algorithm also can be used to find the MLE of θ. First let us obtain $E(T|L < T < R)$, where L and R are fixed quantities and T follows an exponential distribution with mean $\frac{1}{\theta}$. Now

$$E(T|L < T < R) = \frac{e^{-\theta L}\left(L + \frac{1}{\theta}\right) - e^{-\theta R}\left(R + \frac{1}{\theta}\right)}{e^{-\theta L} - e^{-\theta R}}. \tag{9}$$

Note that (9) can be used to compute the EM algorithm. The pseudo likelihood function will take the following form:

$$l(\theta) = \theta^{n_1+n_2} e^{-\theta\left(\sum_{i=1}^{n_1} T_i + \sum_{i=n_1+1}^{n_1+n_2} T_i^{(s)}\right)}, \tag{10}$$

where

$$T_i^{(s)} = \frac{e^{-\theta L_i}\left(L_i + \frac{1}{\theta}\right) - e^{-\theta R_i}\left(R_i + \frac{1}{\theta}\right)}{e^{-\theta L_i} - e^{-\theta R_i}}. \tag{11}$$

Therefore, we use (9) for the 'E' step and then the 'M' step becomes quite trivial. The details are given below.

EM Algorithm:

- Step 1: Suppose $\theta_{(j)}$ is the j^{th} iterate of $\hat{\theta}$.
- Step 2: Compute $T_{i(j)}^{(s)}$ by using (11) replacing θ by $\theta_{(j)}$.
- Step 3: $\theta_{(j+1)} = \frac{n_1+n_2}{\sum_{i=1}^{n_1} T_i + \sum_{i=n_1+1}^{n_1+n_2} T_{i(j)}^{(s)}}$

3 Theoretical Results

Theorem 1: The iterative process provided in (7) will converge if

$$\sum_{i=n_1+1}^{n_1+n_2} r_i \leq 2\sum_{i=1}^{n_1} t_i + 3\sum_{i=n_1+1}^{n_1+n_2} l_i. \tag{12}$$

Proof of Theorem 1: Consider

$$|h'(\theta)| = \frac{1}{\sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} l_i} \left| \sum_{i=n_1+1}^{n_1+n_2} \frac{z_i e^{-\theta z_i}\left(1 - e^{-\theta z_i} - \theta z_i\right)}{\left(1 - e^{-\theta z_i}\right)^2} \right|.$$

Note that

$$\frac{|e^{-x}||(1 - e^{-x} - x)|}{|1 - e^{-x}|^2} \leq \frac{1}{2} \quad \text{for all } x \geq 0,$$

therefore,

$$|h'(\theta)| \leq \frac{1}{2} \frac{\sum_{i=n_1+1}^{n_1+n_2} z_i}{\sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} l_i}$$

We know that the iterative process converges if $|h'(\theta)| < 1$, therefore, the result follows.

Now we need the following lemma to prove the consistency of the MLE.

Lemma 1:

$$\frac{1}{n} L(\theta) \longrightarrow g(\theta) \quad a.s.,$$

where

$$g(\theta) = c' + p(\theta_0) \ln \theta - \theta \left\{ \frac{1}{\theta_0} - \frac{(1 - p(\theta_0))(\alpha + \beta + 2\theta_0)}{(\alpha + \theta_0)(\beta + \theta_0)} \right\} - \theta \frac{(1 - p(\theta_0))}{(\alpha + \theta_0)}$$

$$-\frac{\alpha\beta}{\alpha+\theta_0}\left[\sum_{i=1}^{\infty}\frac{1}{i(\beta+i\theta)}-\sum_{i=1}^{\infty}\frac{1}{i(\beta+i\theta+\theta_0)}\right],$$

$$p(\theta)=\frac{\alpha\beta+\beta\theta+\theta^2}{(\alpha+\theta)(\beta+\theta)},\quad \text{and} c'=\frac{1}{n}\ln c. \tag{13}$$

Proof of Lemma 1: Note that

$$\frac{1}{n}L(\theta)=c'+\frac{n_1}{n}\ln\theta-\frac{\theta}{n}\sum_{i=1}^{n_1}T_i-\frac{\theta}{n}\sum_{i=n_1+1}^{n_1+n_2}L_i+\frac{1}{n}\sum_{i=n_1+1}^{n_1+n_2}\ln\left(1-e^{-\theta Z_i}\right).$$

The density function of T, conditional on the event that $T\notin(L,R)$ can be written as

$$f_{T|T\notin(L,R)}(t)=\frac{1}{p(\theta_0)}\left\{\theta_0 e^{-\theta_0 t}\left(1-\frac{\alpha e^{-\beta t}}{\alpha-\beta}\left(1-e^{-(\alpha-\beta)t}\right)\right)\right\}\quad \text{if } \alpha\neq\beta \tag{14}$$

and

$$f_{T|T\notin(L,R)}(t)=\frac{1}{p(\theta_0)}\left\{\theta_0 e^{-\theta_0 t}\left(1-\alpha t e^{-\alpha t}\right)\right\}\quad \text{if } \alpha=\beta. \tag{15}$$

Note that

$$p(\theta)=P_\theta\left(T\notin(L,R)\right),$$

is as defined in (13). Now using (14) and (15)

$$\begin{aligned}E(T|T\notin(L,R)) &= \frac{1}{p(\theta_0)}\left[\frac{1}{\theta_0}-\frac{\theta_0}{(\alpha+\theta_0)^2}\right]\quad \text{if } \alpha=\beta,\\ &= \frac{1}{p(\theta_0)}\left[\frac{\theta_0}{(\alpha+\theta_0)^2}+\frac{1}{\theta_0}-\frac{\theta_0}{(\alpha+\theta_0)^2}-\frac{\alpha\theta_0}{\alpha-\beta}\left(\frac{1}{(\beta+\theta_0)^2}-\frac{1}{(\alpha+\theta_0)^2}\right)\right]\\ &\quad \text{if } \alpha\neq\beta.\end{aligned}$$

Using the fact that the density function of L conditional on the event $T\in(L,R)$ is

$$f_{L|T\in(L,R)}(x)=\frac{1}{1-p(\theta_0)}\times\frac{\alpha\theta_0}{(\beta+\theta_0)}e^{-(\alpha+\theta_0)x},\quad \text{for } x>0,$$

we have,

$$E(L|T\in(L,R))=\frac{1}{1-p(\theta_0)}\times\frac{\alpha\theta_0}{(\beta+\theta_0)(\alpha+\theta_0)^2}.$$

Similarly, since the density function of $Z=R-L$ conditioned on $T\in(L,R)$ is

$$f_{Z|T\in(L,R)}(z)=\frac{1}{1-p(\theta_0)}\times\frac{\alpha\beta e^{-\beta z}}{(\alpha+\theta_0)}\left(1-e^{-\theta_0 z}\right),\quad \text{for } z>0,$$

therefore,

$$E\left(\ln\left(1-e^{-\theta_0 z}\right)\right) = -\frac{1}{1-p(\theta_0)} \times \frac{\alpha\beta}{(\alpha+\theta_0)}\left[\sum_{i=1}^{\infty}\frac{1}{i(\beta+i\theta)} - \sum_{i=1}^{\infty}\frac{1}{i(\beta+i\theta+\theta_0)}\right].$$

Now the result follows using $\frac{n_1}{n} \to p(\theta_0)$ *a.s.*, and the strong law of large numbers.

LEMMA 2: $g(\theta)$ is a unimodal function, with a unique maximum.

PROOF OF LEMMA 2: It follows from the fact that $g'(0) = \infty$, $g'(\infty) < 0$ and $g''(\theta) < 0$.

LEMMA 3: The MLE of θ_0, say $\hat{\theta}$, will converge to θ^*, where θ^* is the unique solution of the non-linear equation

$$\begin{aligned} g'(\theta) &= \frac{p(\theta_0)}{\theta} - \frac{1}{\theta_0} + \frac{(1-p(\theta_0))(\alpha+\beta+2\theta_0)}{(\alpha+\theta_0)(\beta+\theta_0)} - \frac{1-p(\theta_0)}{(\alpha+\theta_0)} \\ &\quad - \frac{\alpha\beta}{(\alpha+\theta_0)}\left[\sum_{i=1}^{\infty}\frac{1}{(\beta+\theta_0+i\theta)^2} - \sum_{i=1}^{\infty}\frac{1}{(\beta+i\theta)^2}\right] = 0, \end{aligned} \tag{16}$$

where $p(\theta)$ is as defined in (13).

PROOF OF LEMMA 3: In this particular proof we denote $\hat{\theta}$ by $\hat{\theta}_n$

Case 1: $\hat{\theta}_n$ is bounded for all n.

Suppose $\hat{\theta}_n$ does not converge to θ^*. Therefore, there exists a subsequence $\{n_k\}$ of $\{n\}$ and $\tilde{\theta} \neq \theta^*$, such that $\hat{\theta}_{n_k} \to \tilde{\theta}$. Since $\hat{\theta}_{n_k}$ is the MLE,

$$\frac{1}{n_k}L(\hat{\theta}_{n_k}) \geq \frac{1}{n_k}L(\theta^*)$$

Taking limits on both sides of (3) we get

$$g(\tilde{\theta}) \geq g(\theta^*),$$

which leads to a contradiction because θ^* is the unique maximum of $g(\theta)$.

Case 2: $\hat{\theta}_n$ is not bounded.

In this case there exists a subsequence $\{n_k\}$ of $\{n\}$ such that $\hat{\theta}_{n_k} \to \infty$. Note that

$$\frac{1}{n_k} L(\hat{\theta}_{n_k}) \geq \frac{1}{n_k} L(\theta^*),$$

and as $\hat{\theta}_{n_k} \to \infty$, $\frac{1}{n_k} L(\hat{\theta}_{n_k}) \to -\infty$. Since $\frac{1}{n_k} L(\theta^*)$ converges to a fixed number, it leads to a contradiction.

Now since θ_0 is a solution of (16), we have

THEOREM 2: The MLE of θ is a consistent estimator of θ_0.

Now we provide the asymptotic distribution of the MLE.

THEOREM 3: The maximum likelihood estimator has the following asymptotic distribution

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \frac{\sigma^2}{c^2}),$$

where

$$\begin{aligned}
\sigma^2 &= \left[E\left\{ \left(T - \frac{1}{\theta_0} \right)^2 \middle| T \notin (L, R) \right\} - \left(\left\{ E\left(T - \frac{1}{\theta_0} \right) \middle| T \notin (L, R) \right\}^2 \right) \right] \\
&\quad + \left[E(L^2 | T \in (L, R)) - (E(L | T \in (L, R)))^2 \right] \\
&\quad + \left[E\left\{ \left(\frac{Ze^{-\theta_0 Z}}{1 - e^{-\theta_0 Z}} \right)^2 \middle| T \in (L, R) \right\} - \left[E\left\{ \left(\frac{Ze^{-\theta_0 Z}}{1 - e^{-\theta_0 Z}} \right) \middle| T \in (L, R) \right\} \right]^2 \right]
\end{aligned}$$

and

$$c = \frac{p(\theta_0)}{\theta_0^2} + (1 - p(\theta_0)) \left\{ E\left(\frac{Z^2 e^{-\theta_0 Z}}{(1 - e^{-\theta_0 Z})^2} \right) \middle| T \in (L, R) \right\}.$$

To prove Theorem 3, we need the following lemma;

LEMMA 4: Suppose U_i's are a sequence of independent and identically distributed random variables with $E(U_1) = 0$, $V(U_1) = 1$ and $\{N(n)\}$ follows Binomial(n, p), *i.e.*, the probability mass function of $N(n)$ is:

$$P(N(n) = i) = \binom{n}{i} p^i (1 - p)^{n-i}; \quad i = 0, \ldots, n,$$

where $0 < p < 1$. Then as $n \to \infty$,

$$\frac{1}{\sqrt{N(n)}} \sum_{i=1}^{N(n)} U_i \xrightarrow{d} N(0,1).$$

Proof of Lemma 4: Suppose

$$Y_{N(n)} = \frac{1}{\sqrt{N(n)}} \sum_{i=1}^{N(n)} U_i$$

and the characteristic function of $Y_{N(n)}$ is $\phi_{N(n)}(t)$. Then,

$$\begin{aligned} \phi_{N(n)}(t) &= E\left(e^{itY_{N(n)}}\right) = E\left(e^{it\frac{1}{\sqrt{N(n)}}\sum_{i=1}^{N(n)} U_i}\right) \\ &= \sum_{k=0}^{n} E\left(e^{it\frac{1}{\sqrt{k}}\sum_{i=1}^{k} U_i}\middle| N(n)=k\right)\binom{n}{k}p^k(1-p)^{n-k}. \end{aligned}$$

Now if $\phi_U(.)$ denotes the characteristic function of U_1, then for fixed t,

$$\begin{aligned} \left|\phi_{N(n)}(t) - e^{-\frac{t^2}{2}}\right| &\leq \sum_{k=0}^{n}\left|E\left(e^{it\frac{1}{\sqrt{k}}\sum_{i=1}^{k} U_i}\middle| N(n)=k\right) - e^{-\frac{t^2}{2}}\right|\binom{n}{k}p^k(1-p)^{n-k} \\ &= \sum_{k=0}^{n}\left|\left(\phi_U\left(\frac{t}{\sqrt{k}}\right)\right)^k - e^{-\frac{t^2}{2}}\right|\binom{n}{k}p^k(1-p)^{n-k}. \end{aligned}$$

Since by Central Limit Theorem

$$\lim_{k\to\infty} \quad \phi_U\left(\frac{t}{\sqrt{k}}\right)^k = e^{-\frac{t^2}{2}},$$

therefore, for a given $\epsilon > 0$, choose $N_1(t)$ large enough so that for $k \geq N_1(t)$

$$\left|\phi_U\left(\frac{t}{\sqrt{k}}\right)^k - e^{-\frac{t^2}{2}}\right| \leq \epsilon.$$

Moreover, for fixed $N_1(t)$, choose n large enough so that

$$\sum_{i=0}^{N_1(t)}\binom{n}{i}p^i(1-p)^{n-i} \leq \epsilon.$$

Therefore,

$$\begin{aligned}\left|\phi_{N(n)}(t) - e^{-\frac{t^2}{2}}\right| &\leq \sum_{k=0}^{N_1(t)} \left|\left(\phi_U\left(\frac{t}{\sqrt{k}}\right)\right)^k - e^{-\frac{t^2}{2}}\right| \binom{n}{k} p^k (1-p)^{n-k} \\ &\quad + \sum_{k=N_1(t)+1}^{n} \left|\left(\phi_U\left(\frac{t}{\sqrt{k}}\right)\right)^k - e^{-\frac{t^2}{2}}\right| \binom{n}{k} p^k (1-p)^{n-k} \\ &\leq 2\epsilon + \epsilon = 3\epsilon.\end{aligned}$$

Since ϵ is arbitrary, the result follows from the fact that $e^{-\frac{t^2}{2}}$ is the characteristic function of $N(0,1)$ random variable.

Proof of Theorem 3. Note that

$$L(\theta) = n_1 \ln\theta - \theta \sum_{i=1}^{n_1} T_i - \theta \sum_{i=n_1+1}^{n_1+n_2} L_i + \sum_{i=n_1+1}^{n_1+n_2} \ln\left(1 - e^{-\theta Z_i}\right),$$

$$L'(\theta) = \frac{n_1}{\theta} - \sum_{i=1}^{n_1} T_i - \sum_{i=n_1+1}^{n_1+n_2} L_i + \sum_{i=n_1+1}^{n_1+n_2} \frac{Z_i e^{-\theta Z_i}}{(1 - e^{-\theta Z_i})},$$

and

$$L''(\theta) = -\frac{n_1}{\theta^2} - \sum_{i=n_1+1}^{n_1+n_2} \frac{Z_i^2 e^{-\theta Z_i}}{(1 - e^{-\theta Z_i})^2}.$$

Using mean value theorem,

$$L'(\hat{\theta}) - L'(\theta_0) = (\hat{\theta} - \theta_0) L''(\bar{\theta}),$$

where $\bar{\theta}$ is a point between $\hat{\theta}$ and θ_0. Therefore,

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\frac{1}{\sqrt{n}} L'(\theta_0)}{\left(\frac{1}{n} L''(\bar{\theta})\right)}.$$

Now the proof will be complete once we show that:

$$\frac{1}{\sqrt{n}} L'(\theta_0) \longrightarrow N(0, \sigma^2) \quad \text{in distribution} \tag{17}$$

and

$$\frac{1}{n} L''(\bar{\theta}) \longrightarrow c. \quad a.s. \tag{18}$$

Now note that (17) follows from Lemma 4. The proof of (18) follows from the fact that $\bar{\theta}$ converges to θ_0 $a.s$ and from the strong law of large numbers.

4 Bayesian Analysis

In this section we consider a Bayesian formulation of the problem of estimating the parameter θ. We will assume that the parameter θ has a gamma prior distribution with the shape parameter a and scale parameter b, denoted by Gamma(a, b). The density function of the prior density of θ for $a, b > 0$, is

$$\pi(\theta) = \pi(\theta|a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}. \tag{19}$$

No prior distribution on the censoring parameters are assumed. Based on the above assumption, the likelihood function of the observed data is

$$l(data|\theta) = c \;\; \theta^{n_1}e^{-\theta\sum_{i=1}^{n_1} t_i} \prod_{i=n_1+1}^{n_1+n_2} \left(1 - e^{-\theta z_i}\right) e^{-\theta\sum_{i=n_1+1}^{n_1+n_2} l_i}. \tag{20}$$

By a slight abuse of the notation, writing $z_i = z_{n_1+i}$ and $l_i = l_{n_1+i}$ we can rewrite (20) as

$$l(data|\theta) = c \;\; \theta^{n_1}e^{-\theta\sum_{i=1}^{n_1} t_i} \prod_{i=1}^{n_2} \left(1 - e^{-\theta z_i}\right) e^{-\theta\sum_{i=1}^{n_2} l_i}. \tag{21}$$

Based on (19), the joint density of the *data* and θ is

$$l(data|\theta)\pi(\theta). \tag{22}$$

Based on (22), we obtain the posterior density of θ given the *data* as

$$\pi(\theta|data) = \frac{l(data|\theta)\pi(\theta)}{\int_0^\infty l(data|\theta)\pi(\theta)d\theta}. \tag{23}$$

We can write the numerator of the right hand side of (23) as;

$$l(data|\theta)\pi(\theta) = c \;\; \theta^{a+n_1-1}e^{-\theta(b+\sum_{i=1}^{n_1} t_i+\sum_{i=1}^{n_2} l_i)} \prod_{i=1}^{n_2} \left(1 - e^{-\theta z_i}\right). \tag{24}$$

Note that

$$\prod_{i=1}^{n_2} \left(1 - e^{-\theta z_i}\right) = \sum_{P_j} (-1)^{|P_j|} e^{-\theta(z.P_j)}, \tag{25}$$

where P_j is a vector length n_2 and each entry of P_j is either a 0 or a 1. $|P_j|$ denotes the sum of elements of P_j and $z = (z_1, \ldots, z_{n_2})$. The summation on the right hand side of (25) is over 2^{n_2} elements and $(z.P_j)$ denotes the usual dot product between the two vectors of equal lengths. Using (25), the numerator of (23) can be written as

$$l(data|\theta)\pi(\theta) = c \quad \sum_{P_j}(-1)^{|P_j|} \quad \theta^{a+n_1-1} \quad e^{-\theta\left(b+\sum_{i=1}^{n_1} t_i+\sum_{i=1}^{n_2} l_i+(z.P_j)\right)}. \tag{26}$$

So we obtain

$$\int_0^\infty l(data|\theta)\pi(\theta)d\theta = c \quad \sum_{P_j}(-1)^{|P_j|} \quad \frac{\Gamma(a+n_1)}{(b+\sum_{i=1}^{n_1} t_i+\sum_{i=1}^{n_2} l_i+(z.P_j))^{a+n_1}}. \tag{27}$$

Therefore, the posterior density of θ given the data for $\theta > 0$, is

$$\pi(\theta|data) = \frac{\sum_{P_j}(-1)^{|P_j|} \quad \theta^{a+n_1-1} \quad e^{-\theta\left(b+\sum_{i=1}^{n_1} t_i+\sum_{i=1}^{n_2} l_i+(z.P_j)\right)}}{\sum_{P_j} \frac{(-1)^{|P_j|}\Gamma(a+N_1)}{\left(b+\sum_{i=1}^{n_1} t_i+\sum_{i=1}^{n_2} l_i+(z.P_j)\right)^{a+n_1}}}. \tag{28}$$

Therefore, the Bayes estimate of θ under squared error loss function is

$$E(\theta|data) = \frac{\sum_{P_j} \frac{(-1)^{|P_j|}}{\left(b+\sum_{i=1}^{n_1} t_i+\sum_{i=1}^{n_2} l_i+(z.P_j)\right)^{a+n_1+1}}}{\sum_{P_j} \frac{(-1)^{|P_j|}}{\left(b+\sum_{i=1}^{n_1} t_i+\sum_{i=1}^{n_2} l_i+(z.P_j)\right)^{a+n_1}}}. \tag{29}$$

When n_2 is small, the evaluation of $E(\theta|data)$ is not difficult, but for large n_2 it is difficult to compute numerically. We propose a simple Gibbs sampling technique to compute $E(\theta|data)$ and for constructing the corresponding credible interval. Note that when $n_2 = 0$, then,

$$\pi(\theta|data) \sim Gamma(a+n_1, b+\sum_{i=1}^{n_1} t_i), \tag{30}$$

as should be expected. Moreover, the conditional density of T, given $T \in (L, R)$, is

$$f_{T|T\in(L,R)}(x|\theta) = \frac{\theta e^{-\theta x}}{e^{-\theta L} - e^{-\theta R}} \qquad \text{if} \quad L < x < R. \tag{31}$$

Using (30) and (31) we propose the following Gibbs sampling scheme to generate θ from its posterior distribution.

Gibbs Sampling Scheme:

Step 1: Generate $\theta_{1,1}$ from Gamma$(a+n_1, b+\sum_{i=1}^{n_1} t_i)$.

Step 2: Generate $t^{(n_1+i)}$ for $i=1,\ldots,n_2$ from $f_{T|T\in(l_{n_1+i},r_{n_1+i})}(.|\theta_{1,1})$.

Step 3: Generate $\theta_{2,1}$ from Gamma$(a+n_1+n_2, b+\sum_{i=1}^{n_1} t_i+\sum_{i=n_1+1}^{n_1+n_2} t^{(i)})$.

Step 4: Go back to Step 2, and replace $\theta_{1,1}$ by $\theta_{2,1}$ and repeat Steps 2 and 3 for N times.

From the generated N $\theta_{2,j}$, the Bayes estimate of θ_0, under squared error loss function can be be computed as

$$\frac{1}{N-M}\sum_{j=M+1}^{N}\theta_{2,j}, \tag{32}$$

where M is the burn-in sample. Similarly, using the method of Chen and Shao (1999), the highest posterior density (HPD) credible interval of θ_0 also can be constructed.

5 Numerical Results

In this section we mainly compare how the different methods work for small sample sizes and for different censoring schemes. Simulations were carried out using the random number generator RAN2 of Press *et al.* (1992), and based on 1000 replications each. The program written in FORTRAN-77, can be obtained on request from the authors.

We considered different sample sizes namely $n=10$, 20, 30, 40, 50 and different censoring schemes. For the censoring scheme we considered the following combinations of $(1/\alpha, 1/\beta)$ = (0.5, 0.25), (0.5, 0.5), (0.5, 0.75), (1.25, 0.25), (1.25, 0.50) and (1.25, 0.75). In all cases without loss of generality, we have kept $\theta_0=1$. Note that the censoring percentages vary between 10% to 30%. From the given sample we compute maximum likelihood estimator of θ_0 using the EM algorithm and also using the iterative method proposed in section 2. It is observed that in both cases they converge to the same value. We also compute the 95% confidence intervals based on the asymptotic distribution of the maximum likelihood

estimator and replacing the expected Fisher information by the empirical Fisher information. Meeker and Escobar (1998) reported that the confidence interval based on the asymptotic distribution of ln $\hat{\theta}$ is usually superior to one of $\hat{\theta}$. We computed the confidence interval based on the asymptotic distribution of ln $\hat{\theta}$. For comparison purposes, the Bayes estimates under squared error loss function and the corresponding 95% Monte Carlo HPD credible interval as suggested by Chen and Shao (1999) are also reported in Tables 1 and 2. All the Bayes estimates are computed using the prior $a = 0$ and $b = 0$. Note that the above prior is non-informative and non-proper prior. Although, the prior is non-proper but the corresponding posterior has a proper density function. As suggested by Congdon (2001), we tried the prior $a = 0.0001$ and $b = 0.0001$, which is a proper prior but which is almost non-informative, the results are not significantly different and they are not reported here.

From Table 1 one can see that as the sample size increases, the average biases and mean squared errors decrease for both the maximum likelihood estimator and Bayes estimator for all the censoring schemes. It verifies the consistency properties of both the estimators. For fixed sample size and for fixed α, as $1/\beta$ increases (severe censoring), the biases and the mean squared errors both increase for the maximum likelihood estimates. In case of Bayes estimates although the mean squared errors decrease, the same can not be said about the biases. Apart from this, they behave quite similarly both in terms of biases and mean squared errors.

From Table 2 it is clear that as the sample size increases, the average lengths of the confidence/ credible intervals decrease for all the 3 suggested methods. Similarly, for fixed sample size and for fixed α as $1/\beta$ increases, the average lengths increase as expected. For all the three cases, the coverage percentages are quite close to the nominal level (95%) even when the sample size is as small as 10. The performances of all the methods are quite similar in nature. The Bayes credible intervals are slightly larger than the asymptotic confidence

intervals, for moderate sample sizes (namely 20, 30 and 40). The average confidence intervals based on the transformed maximum likelihood estimators (MEE)are slightly longer compared to the other two.

6 Data Analysis

For illustrative purposes, we present a real data analysis results using our proposed method. The data set is taken from Lawless (1982, p. 491) and consists of failure times for 36 appliances subject to an automatic life tests. Although the original data has also the cause of failure with each failure time, but here we are interested in the overall failure distribution and we do not consider the cause of failure in this case. This data set was analyzed using exponential and Weibull models by Kundu and Basu (2000) and it was observed that the exponential model can be used instead of Weibull model. For the complete data set it is observed that the maximum likelihood estimate of θ_0 is 0.00036. The Kolmogorov-Smirnov distance between the empirical distribution function and the fitted exponential distribution function is 0.1944 and the corresponding p value is 0.1317. Therefore, exponential model can not be rejected.

Now we created an artificial data by middle censoring, whose left end was an exponential random variable with mean 500 and the width was exponential with mean 1000. The data after rearranging are presented below:
Data Set: 11, 35, 49, 170, 958, 1062, 1167, 1594, 1925, 1990, 2223, 2327, 2400, 2451, 2471, 2551, 2565, 2568, 2694, 2761, 2831, 3034, 3059, 3112, 3214, 3478, 3504, 4329, 6367, 6976, 7846, 13403, (118.66, 1224.04), (377.76, 2011.51), (351.65, 720.48), (125.96, 4226.08).

The summary statistics of the data are as follows: $n = 36$, $n_1 = 32$, $n_2 = 4$, $\sum_{i=1}^{n_1} t_i =$

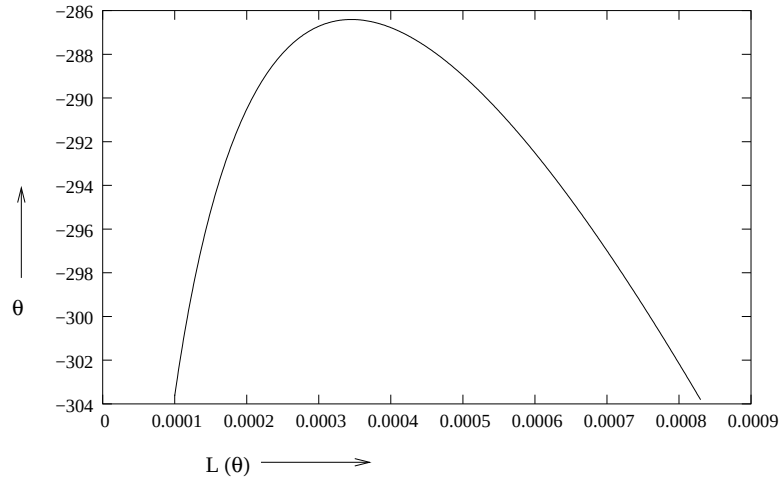

Figure 1: Log-likelihood surface of the given data set.

95125, $\sum_{i=n_1+1}^{n_2} r_i = 8182.11$, $\sum_{i=n_1+1}^{n_2} l_i = 974.03$. Therefore, the iterative process starts with the initial guess $\theta^{(1)} = 32/95125 = 0.000336$. Since r_i, t_i and l_i satisfy the condition (12) of Theorem 1, therefore, the proposed iterative process will converge. The log-likelihood surface with out the additive constant is provided in Figure 1. It clearly shows that the log-likelihood surface is a unimodal function, and therefore the EM algorithm should not have any problem of convergence. The iterative process (7) stops after three iterations and the solution is 0.000364. The 95% confidence intervals based on the asymptotic distribution of $\hat{\theta}$ and $\ln \hat{\theta}$ are (0.00024, 0.00048) and (0.00026, 0.00051) respectively. The Bayes estimate (the posterior mean) under the non-informative and non-proper prior becomes 0.000362 and the corresponding 95% HPD credible interval is (0.00025, 0.00049). The histogram of the generated posterior sample and the fitted gamma distribution are presented in Figure 2. In the same figure we have also plotted the fitted posterior density function assuming $n_2 = 0$. It shows the posterior information of the censored observations.

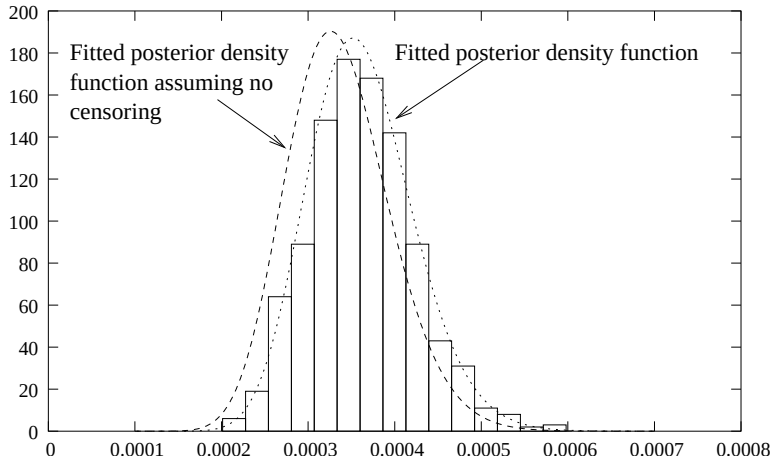

Figure 2: Histogram of the generated 1000 posterior sample and the fitted posterior density functions.

7 Conclusions

In this paper we have considered inference for the exponential distribution when the data is middle censored. Both the classical and Bayesian frameworks are developed. Although this paper focuses on exponential lifetime distributions, similar inferential procedures can be developed for other lifetime distributions such as the Weibull, gamma, log-normal distributions etc. Moreover, in this paper it is assumed that the censoring mechanism is independent and non-informative of the lifetime distribution of the population. Although, it will be difficult, but it might be interesting to consider the case when these assumptions are not valid. We believe, more work is needed along these directions.

Acknowledgements

The authors would like to thank the referees for their very helpful comments and the Guest Editor Professor Takesi Hayakawa for his encouragement.

References

[1] Babu, G.J., Rao, C.R. and Rao, M.B. (1992), "Nonparametric estimation of specific exposure rate in risk and survival analysis", *Journal of the American Statistical Association*, vol. 87, 84-89.

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

[3] Congdon, P. (2001), *Bayesian Statistical Modeling*, John Wiley, New York.

[4] Jammalamadaka, S. Rao and Mangalam, V. (2003), "Non-parametric estimation for middle censored data", *Journal of Nonparametric Statistics*, vol. 15, 253 - 265.

[5] Jammalamadaka, S. Rao and Iyer, S.K. (2004), "Approximate self consistency for middle censored data", *Journal of Statistical Planning and Inference*, vol. 124, 75 - 86.

[6] Jiang, Hongyu J., Jason P. F., Michael R. K., and Rick C. (2005), "Pseudo self-consistent estimation of a copula model with informative censoring", *Scand. J. Statist*, vol. 32, no. 1, 1 - 20.

[7] Kaplan, F.L. and Meier, P. (1958), "Nonparametric estimation from incomplete observations", *Journal of the American Statistical Association*, vol. 63, 457-481.

[8] Kundu, D. and Basu, S.S. (2000), " Analysis of incomplete data in presence of competing risks", *Journal of Statistical Planning and Inference*, vol. 87, 221 - 229.

[9] Lawless, J.F. (1982), *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

[10] Meeker, W.Q. and Escobar, L.A. (1998), *Statistical Methods for Reliability Data*, Wiley, New York.

[11] Press, W.H., Teukolsky, S.A., Vellerling, W.T. and Flannery, B.P. (1992), *Numerical Recipes in FORTRAN, The Art of Scientific Computing*, 2nd. ed., Cambridge University, Cambridge, U.K.

[12] Robertson, J. B., and Uppuluri, V. R. R. (1984), "A generalized kaplan Meier Estimator", *Annals of Statistics*, vol. 12, No. 1, 366 - 371.

[13] Tarpey, T. and Flury, B. (1996), "Self consistency: a fundamental concept in statistics", *Statistical Science*, vol. 11, 229 - 243.

[14] Turnbull, B.W. (1974), "Nonparametric estimation of a survivorship function with doubly censored data", *Journal of the American Statistical Association*, vol. 69, 169-173.

Table 1: The average estimates and the corresponding mean squared errors (within brackets) are reported for the different estimators. Here true value of $\theta_0 = 1$.

n	Methods	(0.5,0.25)	(0.5, 0.5)	(0.5, 0.75)	(1.25, 0.25)	(1.25, 0.5)	(1.25, 0.75)
10	MLE	1.1114	1.1167	1.1295	1.1130	1.1161	1.1237
		(0.1506)	(0.1618)	(0.1836)	(0.1547)	(0.1609)	(0.1738)
	Bayes	1.1043	1.1189	1.1157	1.1220	1.1075	1.1389
		(0.1492)	(0.1629)	(0.1796)	(0.1799)	(0.1589)	(0.2022)
20	MLE	1.0422	1.0446	1.0492	1.0416	1.0436	1.0440
		(0.0631)	(0.0654)	(0.0707)	(0.0633)	(0.0639)	(0.0656)
	Bayes	1.0479	1.0567	1.0471	1.0744	1.0603	1.0485
		(0.0694)	(0.0689)	(0.0739)	(0.0704)	(0.0693)	(0.0613)
30	MLE	1.0352	1.0366	1.0373	1.0350	1.0361	1.0363
		(0.0398)	(0.0407)	(0.0419)	(0.0393)	(0.0400)	(0.0405)
	Bayes	1.0361	1.0349	1.0297	1.0370	1.0335	1.0430
		(0.0409)	(0.0408)	(0.0415)	(0.0395)	(0.0420)	(0.0398)
40	MLE	1.0232	1.0248	1.0254	1.0226	1.0228	1.0239
		(0.0277)	(0.0283)	(0.0287)	(0.0276)	(0.0277)	(0.0286)
	Bayes	1.0308	1.0182	1.0313	1.0286	1.0327	1.0282
		(0.0321)	(0.0265)	(0.0314)	(0.0299)	(0.0303)	(0.0301)
50	MLE	1.0178	1.0189	1.0195	1.0176	1.0182	1.0182
		(0.0229)	(0.0235)	(0.0242)	(0.0276)	(0.0229)	(0.0233)
	Bayes	1.0133	1.0221	1.0136	1.0191	1.0131	1.0224
		(0.0198)	(0.0225)	(0.0211)	(0.0226)	0.0209	(0.0227)

Table 2: The average lengths of the confidence/ credible intervals and the corresponding coverage percentages (within brackets) are reported. Here true value of $\theta_0 = 1$.

n	Methods	(0.5,0.25)	(0.5, 0.5)	(0.5, 0.75)	(1.25, 0.25)	(1.25, 0.5)	(1.25, 0.75)
10	MLE	1.3815	1.4028	1.4440	1.3820	1.3953	1.4194
		(0.97)	(0.97)	(0.97)	(0.96)	(0.96)	(0.96)
	Bayes	1.3531	1.3874	1.4043	1.3706	1.3642	1.4160
		(0.96)	(0.95)	(0.95)	(0.94)	(0.95)	(0.94)
	MEE	1.4722	1.4969	1.5447	1.4726	1.4879	1.5158
		(0.95)	(0.95)	(0.94)	(0.95)	(0.95)	(0.95)
20	MLE	0.9154	0.9253	0.9418	0.9143	0.9210	0.9289
		(0.94)	(0.95)	(0.94)	(0.95)	(0.95)	(0.94)
	Bayes	0.9348	0.9459	0.9480	0.9612	0.9480	0.9402
		(0.94)	(0.94)	(0.94)	(0.95)	(0.94)	(0.96)
	MEE	0.9452	0.9559	0.9738	0.9439	0.9512	0.9598
		(0.95)	(0.95)	(0.94)	(0.95)	(0.94)	(0.95)
30	MLE	0.7423	0.7489	0.7584	0.7416	0.7461	0.7519
		(0.95)	(0.95)	(0.94)	(0.95)	(0.96)	(0.95)
	Bayes	0.7473	0.7554	0.7646	0.7486	0.7488	0.7644
		(0.95)	(0.95)	(0.95)	(0.96)	(0.95)	(0.95)
	MEE	0.7583	0.7653	0.7754	0.7576	0.7624	0.7685
		(0.96)	(0.95)	(0.95)	(0.96)	(0.96)	0.95
40	MLE	0.6354	0.6411	0.6491	0.6346	0.6379	0.6436
		(0.96)	(0.96)	(0.97)	(0.96)	(0.96)	(0.96)
	Bayes	0.6499	0.6433	0.6536	0.6437	0.6534	0.6550
		(0.95)	(0.96)	(0.95)	(0.96)	(0.96)	(0.96)
	MEE	0.6456	0.6517	0.6600	0.6449	0.6483	0.6542
		(0.96)	(0.96)	(0.95)	(0.96)	(0.96)	0.95
50	MLE	0.5652	0.5701	0.5772	0.5648	0.5679	0.5722
		(0.95)	(0.95)	(0.95)	(0.96)	(0.95)	(0.96)
	Bayes	0.5642	0.5699	0.5699	0.5631	0.5665	0.5738
		(0.95)	(0.95)	(0.95)	(0.95)	(0.95)	(0.95)
	MEE	0.5725	0.5776	0.5850	0.5721	0.5753	0.5798
		(0.94)	(0.94)	(0.95)	(0.94)	(0.95)	(0.95)