UC Santa Cruz
UC Santa Cruz Previously Published Works

Title

SEMIPARAMETRIC MODELS WITH COVARIATES FOR LIFETIME DATA UNDER A GENERAL CENSORING SCHEME WITH AN APPLICATION TO CONTINGENT VALUATION

Permalink

https://escholarship.org/uc/item/1m2336k7

Journal
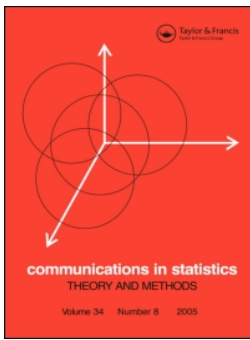
STATISTICA, 76(4)

ISSN

0390-590X

Authors

Bennett, Nathan
Iyer, Srikanth K
Jammalamadaka, Sreenivasa Rao

Publication Date

2016

Peer reviewed

Analysis of Gamma and Weibull lifetime data under a general censoring scheme and in the presence of covariates

Nathan Bennett, Srikanth K. Iyer & S. Rao Jammalamadaka

Taylor & Francis
Taylor & Francis Group

Analysis of Gamma and Weibull lifetime data under a general censoring scheme and in the presence of covariates

Nathan Bennett[a], Srikanth K. Iyer[b], and S. Rao Jammalamadaka[a]

[a]Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA; [b]Department of Mathematics, Indian Institute of Science, Bangalore, India

ABSTRACT

We consider the problem of estimating the lifetime distributions of survival times subject to a general censoring scheme called "middle censoring". The lifetimes are assumed to follow a parametric family of distributions, such as the Gamma or Weibull distributions, and is applied to cases when the lifetimes come with covariates affecting them. For any individual in the sample, there is an independent, random, censoring interval. We will observe the actual lifetime if the lifetime falls outside of this censoring interval, otherwise we only observe the interval of censoring. This censoring mechanism, which includes both right- and left-censoring, has been called "middle censoring" (see Jammalamadaka and Mangalam, 2003). Maximum-likelihood estimation of the parameters as well as their large-sample properties are studied under this censoring scheme, including the case when covariates are available. We conclude with an application to a dataset from Environmental Economics dealing with ContingentValuation of natural resources.

1. Introduction

Our aim in this paper is to estimate the lifetime distribution or its complement, the survival function, for data that is subject to middle censoring. Middle censoring occurs when a data point becomes unobservable if it falls inside a random interval. This is a generalization of left- and right-censored data and is quite distinct from the case of doubly censored data. We consider two families of distributions that are common to many applications, namely the Gamma distribution and the Weibull distribution.

Middle censoring was first introduced by Jammalamadaka and Mangalam (2003) for non parametric estimation of the lifetime distributions, and was studied further in Jammalamadaka and Iyer (2004). Middle-censored data were analyzed in Iyer et al. (2008) when the lifetimes are exponentially distributed, whereas Jammalamadaka and Mangalam (2009) study such censoring in the context of circular data. Gamma and Weibull distributions are natural and the most widely used choices for modeling lifetimes in many applications. Not only does the consideration of these more general models extend the earlier results for the exponential distribution in Iyer, Jammalamadaka, and Kundu (2008), but the current work also discusses how the presence of covariates can be handled nicely in the form of accelerated failure time

---

(AFT) modeling (see Section 3). We derive the maximum-likelihood estimators (MLEs) for the parameters and show how the computation of the MLEs can be done via the EM algorithm. We then establish their large-sample properties.

Let us denote the "actual" lifetimes of n individuals by $t_1, \ldots, t_n$, and not all of them are observable. For each individual there is a random period of time $[\ell_i, r_i]$ for which the lifetime of the ith individual is unobservable. Thus, the actual lifetime is observed if $t_i \notin [\ell_i, r_i]$ and if $t_i \in [\ell_i, r_i]$ then only the interval is observed. Hence, the observed data are given by

$$(x_i, \delta_i) = \begin{cases} (t_i, 1) & if\ t_i \notin [\ell_i,\ r_i] \\ ([\ell_i,\ r_i],\ 0) & otherwise \end{cases} \tag{1}$$

Based on observed data of this type, the goal is to estimate the lifetime distribution function. The lifetimes are assumed to be independent and identically distributed (*i.i.d.*) from an unknown distribution function $F(\cdot)$. Additionally, the censoring intervals, $[L_1, R_1], \ldots, [L_n, R_n]$, are assumed to be *i.i.d.* from an unknown bivariate distribution function $G(\cdot, \cdot)$. Finally, the lifetimes and the censoring intervals are taken to be independent of each other, as is common in survival analysis.

In Section 2, we consider the MLE of the parameters for these two models under middle censoring, discuss the EM algorithm needed for their computation, and establish asymptotic properties like consistency and asymptotic normality of these estimators. In Section 3, we consider estimation under middle-censoring in the presence of covariates employing AFT modeling. Extensive simulations illustrate the robustness of these MLEs even under heavy censoring.

2. Maximum-likelihood estimation

The lifetimes are assumed to follow either a Gamma or a Weibull distribution whose respective probability density functions are given by

$$f_1(t\,|\,a,\ b) = \frac{b^a}{\Gamma(a)} t^{a-1} e^{-bt} \quad for\ t > 0, \tag{2}$$

$$f_2(t\,|\,a,\ b) = ab^a\,t^{a-1} \exp\left[-b^a\,t^a\right] \quad for\ t > 0, \tag{3}$$

and $(a, b) \in \Theta = [0, \infty)^2$. The unknown censoring distribution G is then assumed to be supported on $[0, \infty)^2$. The data can be re-arranged so that the first n_1 observations are uncensored and the last n_2 observations are censored. Then the respective log-likelihood functions for these two models are given by

$$l_n^1(a,\ b) = an_1 \ln b - n_1 \ln(\Gamma(a)) + (a-1) \sum_{i=1}^{n_1} \ln(t_i) - b \sum_{i=1}^{n_1} t_i$$

$$+ \sum_{i=n_1+1}^{n_1+n_2} \ln\left[F_1(r_i\,|\,a,\ b) - F_1(l_i\,|\,a,\ b)\right]. \tag{4}$$

where $F_1(t\,|\,a,\ b)$ is the cumulative distribution function (CDF) of a Gamma (a, b) distribution.

$$l_n^2(a,\ b) = n_1 \ln a + n_1 a \ln b + (a-1) \sum_{i=1}^{n_1} \ln(t_i) - b^a \sum_{i=1}^{n_1} t_i^a$$

$$+ \sum_{i=n_1+1}^{n_1+n_2} \ln\left(\exp\left[-b^a l_i^a\right] - \exp\left[-b^a r_i^a\right]\right) \tag{5}$$

Let $\theta = (a, b)$ denote the unknown parameter vector. The MLE $\hat{\theta}$ of θ is the value of the parameter which maximizes the function in (4) and (5), respectively, for the case of the Gamma and Weibull distributions. We first discuss the large-sample properties of these estimators followed by computational aspects.

2.1. Large-sample properties of the MLEs

Our approach in this section is similar to that of Jammalamadaka and Mangalam (2009). Recall that the censoring mechanism is independent of the lifetime distributions. Conditional on the censoring interval (ℓ, r), define the censoring probability by

$$p_i(\theta, \ell, r) = P(T \in (\ell, r)) = \int_{\ell}^{r} f_i(t|\theta)dt, \qquad i = 1, 2. \tag{6}$$

Let θ_0 denote the true value of the parameter. For convenience, we will work with the parameter b replaced by c^{-1}. Define the functions

$$g_1(\theta, \ell, r) = -a\ln(c) - \ln\left(\Gamma(a)\right) + \int_{t \notin (\ell, r)} \ln\left(t^{a-1}e^{-t/c}\right) f_1(t|\theta_0)dt$$

$$+ p_1(\theta_0, \ell, r)\ln\left(\int_{\ell}^{r} t^{a-1}e^{-t/c}dt\right), \tag{7}$$

$$g_2(\theta, \ell, r) = \ln(a) - a\ln(c) + \int_{t \notin (\ell, r)} \ln\left(t^{a-1}e^{-t^a/c^a}\right) f_2(t|\theta_0)dt$$

$$+ p_2(\theta_0, \ell, r)\ln\left(\int_{\ell}^{r} t^{a-1}e^{-t^a/c^a}dt\right) \tag{8}$$

Define the function

$$h_i(\theta) = \int g_i(\theta, \ell, r)dG(\ell, r), \qquad i = 1, 2. \tag{9}$$

Lemma 2.1. *For $i = 1, 2$, we have $\frac{1}{n}l_n^i(\theta) \to h_i(\theta)$, $P_{\theta_0} - a.s.$*

Proof. For $k = 1, 2, \ldots$, define the sequences of random variables

$$X_k^1 = -a\ln(c) - \ln\left(\Gamma(a)\right) + \delta_k \ln\left(t_k^{a-1}e^{-t_k/c}\right) + (1 - \delta_k)\ln\left(\int_{\ell_k}^{r_k} t^{a-1}e^{-t/c}dt\right), \tag{10}$$

$$X_k^2 = \ln(a) - a\ln(c) + \delta_k \ln\left(t_k^{a-1}e^{-t_k^a/c^a}\right) + (1 - \delta_k)\ln\left(\int_{\ell_k}^{r_k} t^{a-1}e^{-t^a/c^a}dt\right). \tag{11}$$

The above two sequences are i.i.d. with mean $h_i(\theta)$ $i = 1, 2$, respectively, under P_{θ_0}. The result thus follows from the law of large numbers.

The following lemma is a restatement of Lemma 3.3 in Jammalamadaka and Mangalam (2009). $\qquad\square$

Lemma 2.2. *If ℓ and r are two distinct arbitrary points in $(0, \infty)$, then $g_i(\theta, \ell, r) \leq g_i(\theta_0, \ell, r)$ $i = 1, 2$, for all $\theta \in \Theta$ with equality holding only when $\theta = \theta_0$.*

Theorem 2.3. *If the identifiability condition*

$$p(\theta_0) = P_{\theta_0}(T \in (L, R)) < 1,$$

holds, then $\hat{\theta} \to \theta_0$, $P_{\theta_0}-a.s.$

Proof. From Lemma 2.2, it follows that $h_i(\theta) \leq h_i(\theta_0)$ for all $\theta \in \Theta$ with equality holding only when $\theta = \theta_0$.

Fix $\epsilon > 0$ sufficiently small such that $\epsilon < c_0$ and restrict the range of c to (ϵ, ∞). By integrating over the full range of the second integrals in (7) and 8), we obtain

$$g_1(\theta, \ell, r) \leq (-a\ln(c) - \ln(\Gamma(a)))(1 - p_1(\theta_0, \ell, r)) + u(\theta, \ell, r),$$
$$g_2(\theta, \ell, r) \leq (\ln(a) - a\ln(c))(1 - p_2(\theta_0, \ell, r)) + v(\theta, \ell, r),$$

where u, v are the first integrals on the right in (7) and (8), respectively. Under the identifiability condition, $p_i(\theta_0, \ell, r) < 1$ on a set of positive G measure. Hence, $g_i(\theta, \ell, r)$ and hence $h_i(\theta) \to \infty$ as $|\theta| \to \infty$ in $[0, \infty) \times [\epsilon, \infty)$, for $i = 1, 2$. Let Ω_0^i be the set of P_{θ_0} measure 1 where $\frac{1}{n}l_n^i(\theta) \to h_i(\theta)$, $i = 1, 2$. The argument below holds for both $i = 1, 2$ and hence we suppress the index i. Fix any $\omega \in \Omega_0$. If $\hat{\theta}_n \nrightarrow \theta_0$, then there is a subsequence n_k through which $\hat{\theta}_n \to \theta_1 = (a_1, b_1)$, where $(a_1, b_1) \in [0, \infty] \times [\epsilon, \infty]$.

If $|\theta_1| < \infty$, then from Lemma 2.1, $\frac{1}{n_k}l_{n_k}(\hat{\theta}_{n_k}) \to h(\theta_1)$. However, $\frac{1}{n_k}l_{n_k}(\hat{\theta}_{n_k}) \geq \frac{1}{n_k}l_{n_k}(\theta_0) \to h(\theta_0)$ leading to the conclusion that $h(\theta_1) \geq h(\theta_0)$, thus contradicting Lemma 2.2.

If $|\theta_1| = \infty$, then $\frac{1}{n_k}l_{n_k}(\hat{\theta}_{n_k}) = \lim_{\theta \to \theta_1} h(\theta) = -\infty$. Again, $\frac{1}{n_k}l_{n_k}(\hat{\theta}_{n_k}) \geq \frac{1}{n_k}l_{n_k}(\theta_0) \to h(\theta_0)$ leading to a contradiction. □

Theorem 2.4. *Let Σ_1 be the dispersion of $(\frac{\partial X_1^i}{\partial a}, \frac{\partial X_1^i}{\partial b})$, $i = 1, 2$, where X_1^i is defined in (10), (11). Under the identifiability condition given in Theorem 2.3, we have $\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow N_2(0, \Sigma(\theta_0))$, where $\Sigma(\theta) = [h''(\theta)]^{-1}\Sigma_1(\theta)[h''(\theta)]^{-1}$.*

Proof. The proof is fairly straightforward (see for example the proof of Theorem 3.2 in Jammalamadaka and Mangalam, 2009) and so we omit it. □

2.2. Computation of the MLEs

To compute the MLEs, we need to maximize Equations (4) in the case of the Gamma distribution and (5) if the underlying distribution is Weibull. We first describe the EM algorithm and address the issue of convergence of the algorithm for a much wider class of distributions than the ones considered in this paper. Computation of the MLE when the lifetimes are distributed according to the Weibull distribution using the EM algorithm has been considered in Kundu and Pradhan (2014).

Suppose that $x_1, \ldots, x_{n_1}, (l_{n_1+1}, r_{n_1+1}), \ldots, (l_{n_1+n_2}, r_{n_1+n_2})$ is the observed middle-censored data from a continuous exponential family distribution with $k$parameters, namely they have probability density function

$$f(x|\phi) = h(x) \, c(\phi) \exp\left[\sum_{j=1}^{k} w_j(\phi) \, v_j(x)\right], \tag{12}$$

where $h(x)$, $v_j(x)$, $c(\phi)$, and $w_j(\phi)$ are continuous functions. Note that $\phi = (\phi_1, \ldots, \phi_k)$ is a k-dimensional vector of parameters. This results in the following complete log-likelihood:

$$l(\phi) = n \log[c(\phi)] + \sum_{i=1}^{n_1} \left\{ \log[h(t_i)] + \sum_{j=1}^{k} w_j(\phi) v_j(t_i) \right\}$$

$$+ \sum_{i=n_1+1}^{n_1+n_2} \left\{ \log[h(t_i)] + \sum_{j=1}^{k} w_j(\phi) v_j(t_i) \right\} \tag{13}$$

We wish to solve for the MLE of ϕ, which will be done by implementing the EM algorithm. More specifically, we will find the initial estimates for $\phi = (\phi_1, \ldots, \phi_k)$ from the uncensored data, and the estimates of ϕ will be updated using the following procedure:

- Step 1: Suppose that $\phi_{(j)} = (\phi_1, \ldots, \phi_k)_{(j)}$ is the jth estimate
- Step 2: Compute T_i^* by calculating $E\left[T_i | a_i < T_i < b_i, \; \phi = \phi_{(j)}\right]$
- Step 3: Solve Equation (13) with the T_i^*'s imputed for the censored observations for its maximum and set $\phi_{(j+1)}$ as the values that maximizes that equation.
- Step 4: Repeat until convergence criteria is met

We are now ready to prove that this algorithm does indeed converge.

Theorem 2.5. *Let* $x_1, \ldots, x_{n_1}, (l_{n_1+1}, r_{n_1+1}), \ldots, (l_{n_1+n_2}, r_{n_1+n_2})$, *be the observed middle-censored data from a continuous exponential family distribution*

$$f(x|\phi) = h(x)\, c(\phi)\, \exp\left[\sum_{j=1}^{k} w_j(\phi)\, t_j(x)\right]$$

such that $h(x)$, $t_j(x)$, $c(\phi)$, *and* $w_j(\phi)$ *are all continuous functions. Then the EM algorithm will converge for these data.*

Proof. The result follows by an application of the second theorem in Wu (1983) on the EM algorithm. Note that the complete log-likelihood is proportional to

$$l(\phi) \propto n \log[c(\phi)] + \sum_{i=1}^{n_1} \sum_{j=1}^{k} w_j(\phi)\, t_j(x_i) + \sum_{i=n_1+1}^{n_1+n_2} \sum_{j=1}^{k} w_j(\phi)\, t_j(x_i)$$

Also, observe that

$$E\left[t_j(x_i)\,|\phi^*,\, a_i < x_i < b_i\right] = \int_{a_i}^{b_i} t_j(x_i)\, h(x_i)\, c(\phi^*) \exp\left[\sum_{j=1}^{k} w_j(\phi^*)\, t_j(x_i)\right] dx_i$$

is a continuous function. Thus,

$$E\left[l\left(\phi|complete\, data\right)|\phi^*,\, censored\, data\right] \propto n \log[c(\phi)] + \sum_{i=1}^{n_1} \sum_{j=1}^{k} w_j(\phi)\, t_j(x_i)$$

$$+ \sum_{i=n_1+1}^{n_1+n_2} \sum_{j=1}^{k} w_j(\phi)\, E\left[t_j(x_i)\,|\phi^*,\, a_i < x_i < b_i\right]$$

is a continuous function in both ϕ and ϕ^*. Thus by Theorem 2 of Wu (1983), it follows that the EM algorithm will converge. We now move on to specific examples.

We first consider the case of the Gamma distribution. In this case the log-likelihood can be written as

$$l(\alpha, \beta) = an_1 \ln b - n_1 \ln (\Gamma(a)) + (a - 1) \sum_{i=1}^{n_1} \ln (t_i) - b \sum_{i=1}^{n_1} t_i$$

$$+ \sum_{i=n_1+1}^{n_1+n_2} \ln [F (r_i | a, b) - F (l_i | a, b)] \tag{14}$$

where and $F(t | a, b)$ is the CDF of a Gamma (a, b) distribution. We now need the conditional expectations for the incomplete data in order to use the EM algorithm. The two necessary expectations are

$$E[T | L < T < R] = \frac{\int_L^R t \frac{b^a}{\Gamma(a)} t^{a-1} e^{-bt} dt}{F_1(R | a, b) - F_1(L | a, b)} \tag{15}$$

$$E[\ln T | L < T < R] = \frac{\int_L^R \ln t \frac{b^a}{\Gamma(a)} t^{a-1} e^{-bt} dt}{F_2(R | a, b) - F_2(L | a, b)} \tag{16}$$

The above equation does not have a closed-form solution and so we solve numerically to obtain the solution. This can be used in the E-Step in the EM algorithm, and then the pseudo log-likelihood will be

$$l^*(\theta) = an \ln b - n \ln (\Gamma(a)) + (a - 1) \left[\sum_{i=1}^{n_1} \ln (t_i) + \sum_{i=n_1+1}^{n_1+n_2} \ln (t_i^*) \right]$$

$$-b \left[\sum_{i=1}^{n_1} t_i + \sum_{i=n_1+1}^{n_1+n_2} t_i^* \right] \tag{17}$$

where the t_i^*'s are found using Equations (15) and (16).

Thus, the EM Algorithm can be set up as follows. Choose $(a, b)_{(0)}$ to be the MLE of the uncensored data. Update the estimates with the following steps:

- Step 1: Suppose that $(a, b)_{(j)}$ is the jth estimate.
- Step 2: Compute T_i^* using Equation (15) and 16 with $(a, b) = (a, b)_{(j)}$.
- Step 3: Solve Equation (17) for its maximum and set $(a, b)_{(j+1)}$ as that maximum.
- Step 4: Repeat until convergence criteria are met.

Since there is no explicit form for the MLEs of a Gamma distribution, the maximum must either be solved iteratively or with a built-in numerical solver.

The same procedure works for the case of the Weibull distribution. We re-label b^a as b while carrying out the simulations. In this case the log likelihood for the Weibull lifetimes is given by

$$l(a, b) = n_1 \ln a + n_1 \ln b + (a - 1) \sum_{i=1}^{n_1} \ln (t_i) - b \sum_{i=1}^{n_1} t_i^a$$

$$+ \sum_{i=n_1+1}^{n_1+n_2} \ln \left(\exp \left[-bl_i^a \right] - \exp \left[-br_i^a \right] \right) \tag{18}$$

The desired conditional expectations are given by

$$E[T^a | L < T < R] = \frac{\int_L^R t^a \, ab \, t^{a-1} \exp\left[-b \, t^a\right] dt}{\exp\left[-bl_i^a\right] - \exp\left[-br_i^a\right]} \tag{19}$$

$$E[\ln T | L < T < R] = \frac{\int_L^R \ln t \, ab \, t^{a-1} \exp\left[-b \, t^a\right] dt}{\exp\left[-bl_i^a\right] - \exp\left[-br_i^a\right]} \tag{20}$$

This will lead to the following pseudo log-likelihood

$$l^*(a, \, b) \propto n \ln a + n \ln b + (a - 1) \left[\sum_{i=1}^{n_1} \ln (t_i) + \sum_{i=n_1+1}^{n_1+n_2} \ln \left(t_i^*\right) \right]$$

$$-b \left[\sum_{i=1}^{n_1} t_i^a + \sum_{i=n_1+1}^{n_1+n_2} \left(t_i^*\right)^a \right] \tag{21}$$

The rest of the procedure is identical to the previous case. ☐

3. Accelerated failure time models

In this section, we will consider the problem of MLE of the parameters of a $p-$parameter AFT model where the baseline distribution is exponential, gamma or Weibull distributed. The AFT models are known to be more robust to the estimation of covariate effects (e.g. see Keiding and Andersen, 1997), and are more easy to interpret than hazard rates. For instance in a clinical trial where mortality is the endpoint, one could translate the result as a certain percentage increase in future life expectancy on the new treatment compared to the baseline. As before suppose the middle-censored data are in the form

$$t_1, \ldots, t_{n_1}, \left(l_{n_1+1}, \, r_{n_1+1}\right), \ldots, \left(l_{n_1+n_2}, \, r_{n_1+n_2}\right)$$

Associated with each observation is an observed vector Z_i representing the covariates. The MLE is done as earlier using the EM algorithm, which requires the conditional expectation of the unobserved data given that it falls in a particular interval. As before, the convergence of the EM algorithm is a consequence of the continuity of the log-likelihood function. The respective probability density function of the observations for the three models, namely the exponential, Gamma and Weibull are given below:

$$f(t|Z, a) = \exp\left[\theta^T \mathbf{Z_i}\right] \exp\left\{-a \exp\left[\theta^T \mathbf{Z_i}\right] t\right\}, \qquad t > 0, \tag{22}$$

$$f(t|Z, a, b) = \frac{1}{\Gamma(a) \left(b \exp\left[-\theta^T \mathbf{Z}\right]\right)^a} t^{a-1} \exp\left[\frac{-t}{b \exp\left[-\theta^T \mathbf{Z}\right]}\right], \qquad t > 0, \tag{23}$$

$$f(t|Z, a, b) = a \left(b \exp\left[a \, \theta^T \mathbf{Z}\right]\right) t^{a-1} \exp\left[-\left(b \exp\left[a \, \theta^T \mathbf{Z}\right]\right) t^a\right], \qquad t > 0. \tag{24}$$

Table 1. Numerical results for Gamma ($a = 2, b = 1$) lifetimes.

	(α, β)	(1, 1)	(0.5, 0.5)	(0.5, 1)	(1.25, 0.75)
n					
50	*a* est	2.1150	2.0914	2.1245	2.1062
	b est	0.9980	0.9982	0.9757	0.9880
	MSE *a*	0.2084	0.1805	0.2064	0.1868
	MSE *b*	0.0453	0.0479	0.0479	0.0438
	Censored	(0.08, 0.44)	(0.04, 0.50)	(0.04, 0.44)	(0.14, 0.50)
100	*a* est	2.0676	2.0665	2.0534	2.0562
	b est	0.9882	0.9814	0.9935	0.9959
	MSE *a*	0.0962	0.0904	0.0856	0.0944
	MSE *b*	0.0239	0.0237	0.0232	0.0255
	Censored	(0.13, 0.40)	(0.13, 0.43)	(0.08, 0.31)	(0.20, 0.47)
500	*a* est	2.0144	2.0135	2.0013	2.010
	b est	0.9960	0.9963	0.9968	0.9989
	MSE *a*	0.0153	0.0159	0.0147	0.0168
	MSE *b*	0.0046	0.0049	0.0045	0.0050
	Censored	(0.19, 0.33)	(0.23, 0.36)	(0.14, 0.25)	(0.25, 0.39)

4. Simulation study

To illustrate and validate the procedure, we simulate data under the assumption that the left end point and the length of the censoring intervals are independent and exponentially distributed with parameters α and β, respectively. For each sample size *n*, $N = 1000$ samples were simulated. Each sample was then censored, and the EM algorithm was applied to the censored data. The *a* and *b* estimates reported are the average value of the $N = 1000$ estimates obtained.

See Table 1 for the results of these simulations when the lifetimes are from a Gamma distribution. The row *Censored* in the table provides the smallest proportion of censoring and largest proportion of censoring in the $N = 1000$ simulated samples. The simulation results for the Weibull case are summarized in Table 2. The estimates for the Weibull model also appear to converge very well. The procedure performs reasonably well even with a large proportion of censored observations.

Table 2. Numerical results for Weibull ($a = 2, b = 1$) lifetimes.

|	(α, β)	(1, 1)	(0.5, 0.5)	(0.5, 1)	(1.25, 0.75)
n					
50	*a* est	1.9273	1.9635	1.9659	1.8799
	b est	1.1211	1.1504	1.1000	1.1569
	MSE *a*	0.0952	0.0905	0.0667	0.1172
	MSE *b*	0.0716	0.0730	0.0500	0.1038
	Censored	(0.12, 0.54)	(0.10, 0.44)	(0.06, 0.40)	(0.16, 0.60)
100	*a* est	2.0275	2.0305	2.0417	1.9930
	b est	1.1015	1.1386	1.0859	1.1193
	MSE *a*	0.0107	0.0098	0.0102	0.0160
	MSE *b*	0.0391	0.0416	0.0224	0.0508
	Censored	(0.19, 0.47)	(0.14, 0.39)	(0.10, 0.33)	(0.22, 0.56)
500	*a* est	2.0008	2.0024	2.0042	2.0002
	b est	1.0414	1.1032	1.0468	1.0639
	MSE *a*	0.00016	0.00048	0.00084	0.00004
	MSE *b*	0.01076	0.02344	0.01040	0.02004
	Censored	(0.26, 0.39)	(0.19, 0.32)	(0.15, 0.26)	(0.34, 0.47)

Plot of ECDF vs. Fitted CDF
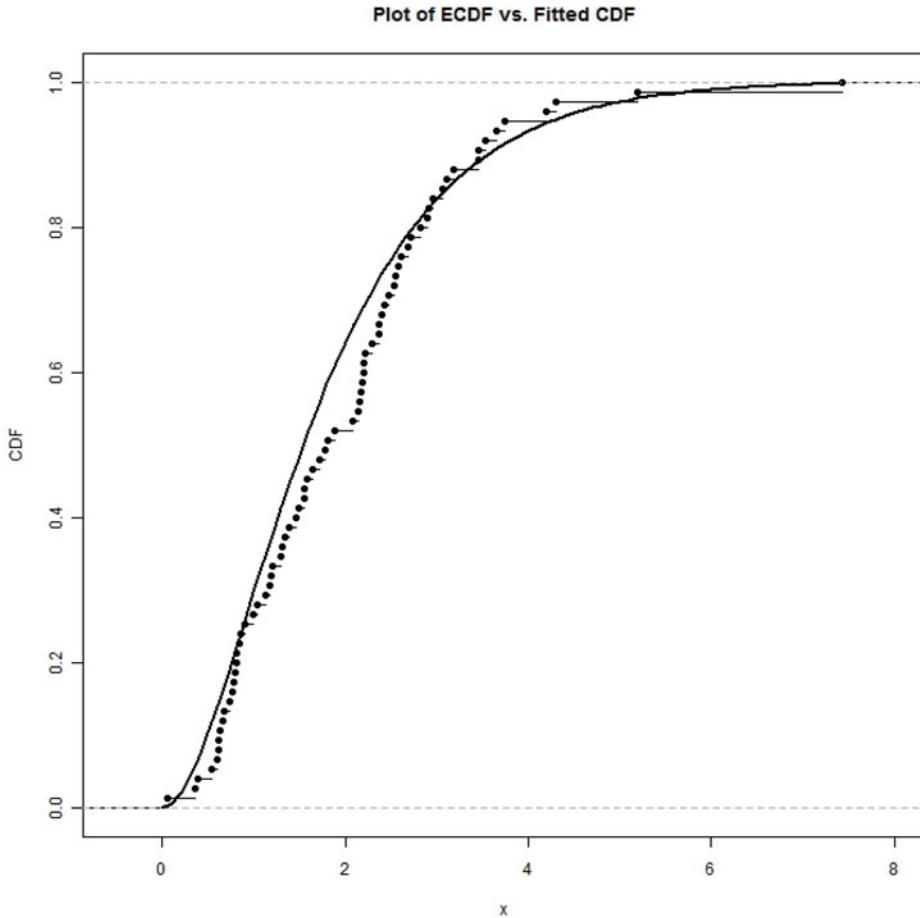
Figure 1. Goodness of fit using the Gamma distribution.

Also examined was the goodness of fit of the estimated model. To study this, a sample of size $n = 100$ was created from a Gamma distribution. Using the aforementioned process, these data were middle-censored, resulting in 25% of the data being censored. The MLEs were calculated using the proposed EM algorithm. The empirical CDF of the uncensored data and fitted CDF are given in Figure 1. The two curves appear to be very similar. Furthermore, a Kolmogorov Smirnov test was performed using the fitted Gamma distribution and uncensored data which yielded a p-value of 0.433 indicating no lack of fit.

A simulation study was performed to illustrate the usefulness of the approach outlined above for the AFT models. Simulations were carried out in R using $N = 1000$ replications with a common sample size of $n = 100$. The censoring mechanism is the same as used previously. Specifically, the left endpoint of the censored interval is Exponentially distributed with mean 1; the length of the censored interval is also Exponentially distributed with mean 1. Three covariate values were used. The first two covariates, Z_1 & Z_2, were generated from a Binomial distribution with one trial and probability of success equal to 0.5. The third covariate, Z_3, was

Table 3. Numerical results for exponential AFT model.

	a	θ_1	θ_2	θ_3
True value	1	1	1	1
MLE	1.0334	1.0003	0.9979	0.9961
MSE	0.0367	0.0465	0.0464	0.0115
True value	1	1	0	0
MLE	1.0318	1.0139	0.0018	0.0078
MSE	0.0395	0.0478	0.0452	0.0123
True value	1	0	0	1
MLE	1.0445	−0.0065	0.0035	0.9985
MSE	0.0422	0.0468	0.0453	0.0110

Table 4. Numerical results for Gamma AFT model.

	a	b	θ_1	θ_2	θ_3
True value	2	1	1	1	1
MLE	2.2761	0.8961	1.0005	1.0021	1.0012
MSE	0.1927	0.0415	0.0238	0.0227	0.0057
True value	2	1	1	0	0
MLE	2.3025	0.8842	1.0102	−0.0092	0.0009
MSE	0.2146	0.0421	0.0225	0.0236	0.0058
True value	2	1	0	0	1
MLE	2.2383	0.9166	0.0022	0.0115	0.9986
MSE	0.1641	0.0377	0.0230	0.0208	0.0060

generated from a Standard Normal distribution. Similar to Pan (1999), three cases for the true covariate effects were considered. They are $\theta = (1,\ 1,\ 1)$, $\theta = (1,\ 0,\ 0)$, and $\theta = (0,\ 0,\ 1)$. These three cases were chosen since they represent the case where all covariates have an equal effect, where only one Bernoulli covariate has an effect, and where only the Normally distributed covariate had an effect. In the exponential case, between 7% and 36% of the observations were censored, between 9% and 42% in the case of the Gamma distribution and between 8% and 40% in the Weibull case. Tables 3–5 report the results from these simulations. The MLEs of all the parameters in all cases are very close to the actual value, and the mean-squared errors are also small.

Finally we evaluate the theoretical convergence to normality of the estimators via simulations. For this purpose, we simulate $N = 100$ samples of size $n = 100$ each from the Gamma AFT and Weibull AFT models described in the previous paragraph. We compute the MLEs for the samples in the usual fashion. The Q-Q plots of the estimated values for the shape and

Table 5. Numerical results for Weibull AFT model.

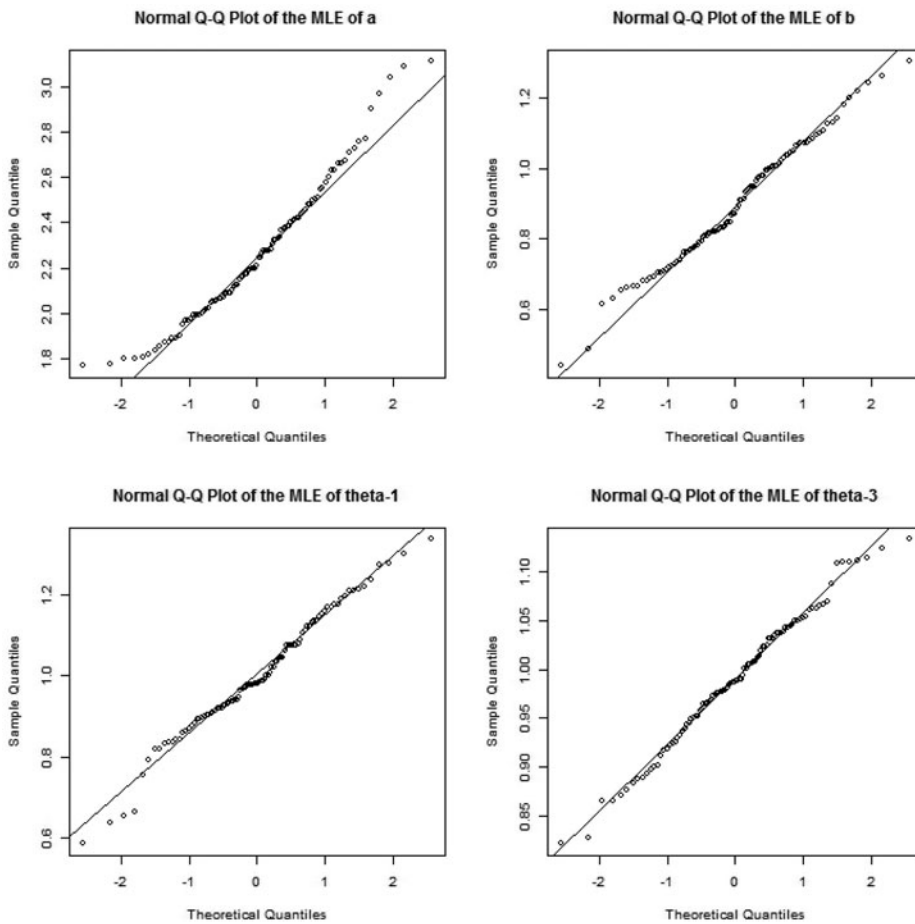|	a	b	θ_1	θ_2	θ_3
True value	2	1	1	1	1
MLE	1.8582	1.0556	1.0452	1.0579	1.0502
MSE	0.0303	0.0411	0.0190	0.0218	0.0081
True value	2	1	1	0	0
MLE	1.9251	1.0796	1.1554	−0.0094	0.0040
MSE	0.0197	0.0470	0.0431	0.0245	0.0066
True value	2	1	0	0	1
MLE	1.8714	0.9847	−0.0136	0.0008	1.1471
MSE	0.0277	0.0457	0.0319	0.0319	0.0275

Figure 2. $Q - Q$ plot of MLEs from Gamma AFT model.

scale parameters, and first and third elements of θ are displayed in Figures 2 and 3. These plots seem to indicate a fairly good fit to normality.

5. Data analysis

To highlight the usefulness of the methods developed in the previous sections, we will now consider a dataset from Environmental Economics. The data studied is from a Contingent Valuation study conducted by Cecilia Hakansson from Sweden and Katja Parkkila from Finland in 2004. People in Finland were asked how much they were willing-to-pay (WTP) to increase the salmon stock in a particular river basin. Participants were allowed to either give an exact amount that they were WTP or provide an interval which contained their WTP if they preferred.

A total of 205 Finnish subjects provided data for their WTP and income. Of the 205 responders, 57 gave intervals, and thus 27.8% of the data are middle censored. We fit a Weibull AFT model to these data, using Equation (24) with 1 covariate. The fitted values for this are as follows: $\hat{a} = 1.4407$, $\hat{b} = 0.0149$, $\hat{\theta} = -0.1610$. To transform the parameter value onto the
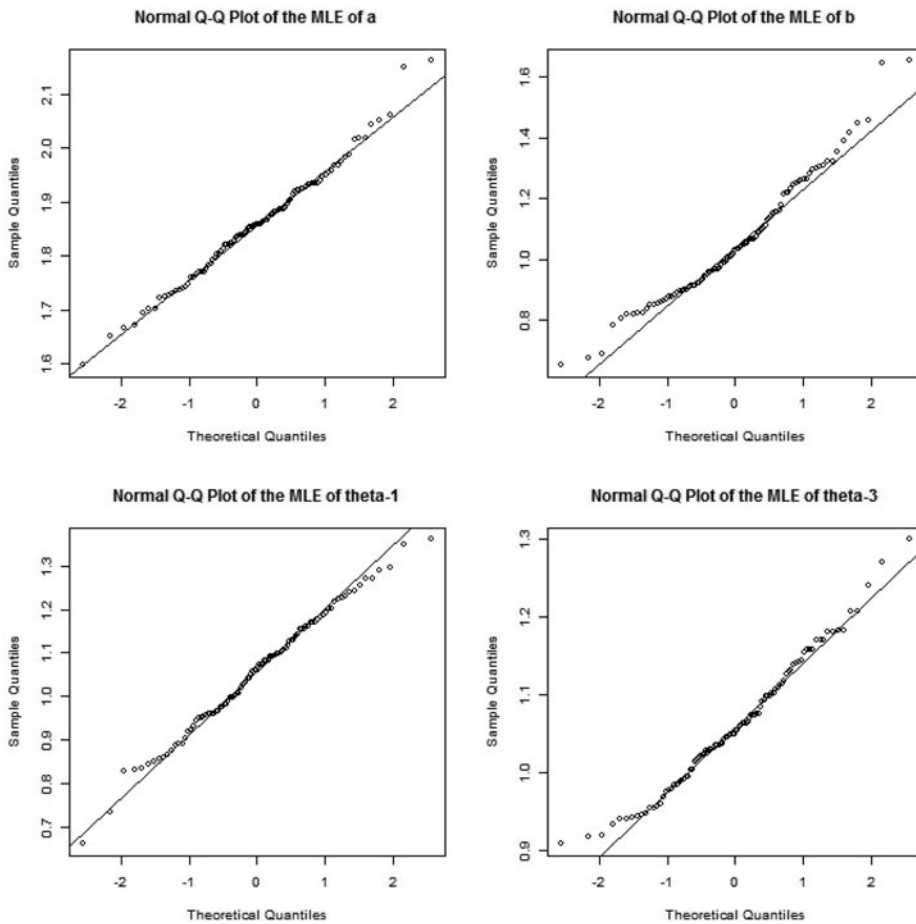
Figure 3. $Q - Q$ plot of MLEs from Weibull AFT model.

WTP scale, we must look at $\exp[\hat{\theta}] = 0.8513$. In this dataset, this means that people with higher incomes have a lower WTP.

6. Conclusion

In conclusion, we prove that the MLEs from a large family of distributions will converge in the case of middle censoring, and we give their large-sample properties. Additionally, we also consider the case of parametric models with the presence of covariates and again provide the large-sample properties of these estimators. In both cases, simulation studies are presented illustrating the usefulness and accuracy of these methods.

The MLEs of the regression coefficients are very close to the true value in all cases, but the MLEs of the parameters from the Gamma distribution are slightly off. Again, in all cases, the mean-squared errors are small for all parameters except for $\hat{a}_{MLE}$. Table 5 reports the results from these simulations. The MLEs of the regression coefficients are very close to the true value in the case of equal effects of all covariates, but the estimates are slightly when only one covariate has an effect. The MLEs of the parameters from the Weibull distribution are fairly good, but $\hat{a}_{MLE}$ was consistently underestimated. Again, in all cases, the mean-squared errors

are quite small for all the parameters, demonstrating that the estimation procedures work well. Finally, this methodology is very flexible and applicable to many different areas of research, as demonstrated by the Contingent Valuation example.

References

Dempster, A.P., Laird, N.M., Rubin, D.B. (1997). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B.* 39:1–22.

Gentleman, R., Geyer, C.J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika* 81:618–623.

Iyer, S.K., Jammalamadaka, S. Rao, Kundu, D. (2008). Analysis of middle-censored data with exponential lifetime distributions. *J. Stat. Plann. Inference* 138(11):3550–3560.

Jammalamadaka, S. Rao, Mangalam, V. (2003). Nonparametric estimation for middle censored data. *J. Nonparamet. Stat.* 15(2):253–265.

Jammalamadaka, S. Rao, Mangalam, V. (2009). A general censoring scheme for circular data. *Stat. Methodol.* 15(3):280–289.

Keiding, N., Andersen, P.K., Klein, J.P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Stat. Med.* 16(1–3):215–224.

Kundu, D., Pradhan, B. (2014). Analysis of interval-censored data with Weibull lifetime distribution. *Sankhya B* 76(1):120–139.

Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd Ed. New York: John Wiley & Sons.

Tsai, W.Y., Crowley, J. (1985). A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency. *Ann. Stat.* 13(4):1317–1334.

Wei, P. (1999). Extending the iterative convex minorant algorithm to the cox model for interval-censored data. *J. Comput Graph. Stat.* 8(1):109–120.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Ann. Stat.* 11:95–103.