Ann. Inst. Statist. Math.
Vol. 53, No. 2, 262–276 (2001)

KULLBACK-LEIBLER INFORMATION CONSISTENT ESTIMATION FOR CENSORED DATA

AKIO SUZUKAWA* , HIDEYUKI IMAI AND YOSHIHARU SATO**

*Division of Systems and Information Engineering, Hokkaido University,
Kita 13, Nishi 8, Kitaku, Sapporo 060-8628, Japan*

(Received March 11, 1998; revised June 29, 1999)

Abstract. This paper is intended as an investigation of parametric estimation for the randomly right censored data. In parametric estimation, the Kullback-Leibler information is used as a measure of the divergence of a true distribution generating a data relative to a distribution in an assumed parametric model $\mathcal{M}$. When the data is uncensored, maximum likelihood estimator (MLE) is a consistent estimator of minimizing the Kullback-Leibler information, even if the assumed model $\mathcal{M}$ does not contain the true distribution. We call this property minimum Kullback-Leibler information consistency (MKLI-consistency). However, the MLE obtained by maximizing the likelihood function based on the censored data is not MKLI-consistent. As an alternative to the MLE, Oakes (1986, *Biometrics*, **42**, 177–182) proposed an estimator termed approximate maximum likelihood estimator (AMLE) due to its computational advantage and potential for robustness. We show MKLI-consistency and asymptotic normality of the AMLE under the misspecification of the parametric model. In a simulation study, we investigate mean square errors of these two estimators and an estimator which is obtained by treating a jackknife corrected Kaplan-Meier integral as the log-likelihood. On the basis of the simulation results and the asymptotic results, we discuss comparison among these estimators. We also derive information criteria for the MLE and the AMLE under censorship, and which can be used not only for selecting models but also for selecting estimation procedures.

Key words and phrases: Approximate likelihood, information criterion, Kaplan-Meier estimator, maximum likelihood estimation.

1. Introduction

Suppose that $X_1, \dots, X_n$ are i.i.d. random variables from an unknown distribution $F_0(x)$ with probability density $f_0(x)$. Parametric inference is done within an assumed parametric family of densities $\mathcal{M} = \{f(x;\theta); \theta \in \Theta\}$, which may or may not contain the true density f_0. If $\mathcal{M}$ contains f_0, there exists $\theta_0 \in \Theta$ such that $f_0(x) = f(x;\theta_0)$, and θ_0 is called the true parameter value. In this case, our aim is to estimate θ_0 based on the model $\mathcal{M}$. On the other hand, if f_0 is not contained in $\mathcal{M}$, what should be estimated within the model $\mathcal{M}$? A simple answer is that we should try to know the

*Now at Department of Animal Production and Agricultural Economics, Obihiro University, Inada-cho, Obihiro 080-8555, Japan. Supported partially by Grant-in-Aid Scientific Research, No. 10780145, Ministry of Education, Science, Sports and Culture, Japan.

**Supported partially by Grant-in-Aid Scientific Research, No. 11308010, Ministry of Education, Science, Sports and Culture, Japan.

nearest $f(x;\theta)$ to the true density $f_0(x)$. Cramér (1946), Wald (1949) and Akaike (1973) pointed out that the maximum likelihood (ML) principle aims to know a value of θ maximizing $\int f_0(x)\log f(x;\theta)dx$. This means that a purpose of the ML principle is to find a parameter θ which minimizes the Kullback-Leibler information

$$\mathrm{KL}(f_0(\cdot), f(\cdot;\theta)) = \int f_0(x)\log\frac{f_0(x)}{f(x;\theta)}dx, \tag{1.1}$$

which is a measure of the divergence of $f_0(x)$ relative to $f(x;\theta)$. Under suitable regularity conditions, the maximum likelihood estimator (MLE), which is defined as a value of $\theta(\in\Theta)$ maximizing the likelihood function $\prod_{i=1}^n f(X_i;\theta)$, is a consistent estimator of θ_0^* which minimizes (1.1) (Cramér (1946), Wald (1949), Takeuchi (1976)). We call this property *minimum Kullback-Leibler information consistency* (MKLI-consistency). We note that the MKLI-consistency implies the consistency in usual sense when $\mathcal{M}$ includes f_0.

In the analysis of lifetime data, an important problem is censorship of observations. For $i = 1,\dots,n$, let X_i and Y_i be random variables which represent a lifetime and a censoring time of the i-th individual, respectively. In lifetime data analysis, X_i and Y_i are not directly observed, and we can observe

$$(Z_i,\delta_i) = (\min(X_i,Y_i), I(X_i\le Y_i)),$$

where $I(A)$ denotes the indicator function of the set A. We agree that δ_i indicates whether X_i has been censored or not. The set of observations (Z_i,δ_i), $1\le i\le n$ is called randomly right censored data in survival analysis and reliability theory.

Let $G(y)$ be an unknown distribution of the censoring time with density $g(y)$. Suppose that $Y_1,\dots,Y_n$ are i.i.d. from $G(y)$ and X_i's are independent of Y_i's. Our main goal is to draw some inference on the true distribution of X_i, i.e. F_0, while G is a nuisance parameter. The nonparametric maximum likelihood estimator of F_0 is given by the Kaplan-Meier estimator (Kaplan and Meier (1958))

$$\hat{F}_n(x) = 1-\prod_{i=1}^n\left[1-\frac{\delta_{[i]}}{n-i+1}\right]^{I(Z_{(i)}\le x)}, \tag{1.2}$$

where $Z_{(1)}\le\cdots\le Z_{(n)}$ are the ordered values of Z_i, and $\delta_{[i]}$ denotes the concomitant associated with $Z_{(i)}$. In the uncensored case, i.e. all δ_i's equal one, the Kaplan-Meier estimator $\hat{F}_n(x)$ coincides with the empirical distribution function $F_n(x)=\sum_{i=1}^n I(X_i\le x)/n$.

When the parametric model $\mathcal{M}$ is assumed for the distribution of X_i, the log-likelihood function is given by

$$l_n(\theta) = \sum_{i=1}^n\left\{\delta_i\log f(Z_i;\theta) + (1-\delta_i)\log\bar{F}(Z_i;\theta)\right\}, \tag{1.3}$$

where $\bar{F}(z;\theta)=\int I(u>z)f(u;\theta)du$ (see Kalbfleisch and Prentice (1980), Section 3.2). The maximum likelihood estimator is an element $\hat{\theta}_n\in\Theta$ which attains the maximum value of $l_n(\theta)$ in Θ. As mentioned above, when all X_i's are observable, the MLE is MKLI-consistent. However, under random censorship, $\hat{\theta}_n$ is not MKLI-consistent when $\mathcal{M}$ does not contain f_0.

For example, suppose that the true distributions of X_i and Y_i are the Weibull distributions :

$$f_0(x) = \lambda\beta(\lambda x)^{\beta-1}\exp\{-(\lambda x)^\beta\}, \quad g(y) = \xi\beta(\xi y)^{\beta-1}\exp\{-(\xi y)^\beta\}, \quad (x \geq 0, y \geq 0),$$

where $\beta > 0$, $\lambda > 0$, $\xi > 0$, and an assumed model is the exponential distribution model:

$$\mathcal{M} = \{f(x;\theta) = \theta\exp(-\theta x); \theta > 0\}.$$

In this case, the MLE is given by $\hat{\theta}_n = (\sum_{i=1}^n \delta_i)/(\sum_{i=1}^n Z_i)$, and it converges to $\lambda^\beta(\lambda^\beta + \xi^\beta)^{\beta^{-1}-1}/\Gamma(1+\beta^{-1})$ in probability as $n \to \infty$, where $\Gamma(\cdot)$ is the Gamma function. On the other hand, θ_0^* which is a parameter value minimizing (1.1) is $\lambda/\Gamma(1+\beta^{-1})$. Therefore, if $\beta \neq 1$, $\hat{\theta}_n$ is not consistent to θ_0^*. We see that $\beta = 1$ implies that the assumed model contains the true distribution.

In this paper, we consider another estimator $\hat{\theta}_n^*$, which is defined as an element in Θ which maximizes

$$l_n^*(\theta) = n\int \log f(x;\theta)d\hat{F}_n(x).$$

When all X_i's are observable, the log-likelihood function can be expressed as

$$\sum_{i=1}^n \log f(X_i;\theta) = n\int \log f(x;\theta)dF_n(x).$$

Thus $l_n^*(\theta)$ is a natural extension to the censored data in the sense that the empirical distribution F_n is replaced by the Kaplan-Meier estimator $\hat{F}_n$. It is noted that, when all δ_i's equal one (uncensored case), $l_n^*(\theta) = l_n(\theta)$ and therefore $\hat{\theta}_n^* = \hat{\theta}_n$ holds.

This idea of parametric estimation based on censored data was first proposed by Oakes (1986), and which is referred to as approximate maximum likelihood procedure. Oakes (1986) uses Efron's version of the Kaplan-Meier estimator which sets $\hat{F}_n(x) = 1$ after the largest observation. Although in this sense the estimator $\hat{\theta}_n^*$ is slightly different from Oakes', we call $\hat{\theta}_n^*$ approximated maximum likelihood estimator (AMLE). It is also a special case of M-estimators discussed by Wang (1995), in which strong consistency is studied.

Although a considerable number of studies have been made on parametric estimation for censored data, little attention has been given to the misspecification of the parametric model. Our main concern are to consider the parametric estimation under the misspecification and to discuss comparison of the MLE and the AMLE from this point of view.

In Section 2, we give some results on asymptotic properties of the AMLE, which include a result concerning the MKLI-consistency. In Section 3, we report on a simulation study to investigate mean squared errors of the estimators, and discuss comparison of them. In Section 4, we derive information criteria corresponding to the MLE and the AMLE, and which are extensions of Takeuchi (1976)'s TIC.

2. Asymptotic properties

In this section, we discuss the MKLI-consistency and the asymptotic normality of the AMLE $\hat{\theta}_n^*$. We begin with the following assumptions:

(A1) The parameter space Θ is an open interval in $\boldsymbol{R}$.

(A2) $f(x;\theta)$ is continuous for almost every x.

(A3) All probability density functions in the model $\mathcal{M}$ have the same support.

(A4) $\eta^*(\theta) \equiv \int f_0(x)\log f(x;\theta)dx$ has a maximum at θ_0^*, and for any $\theta \neq \theta_0^*$, $\eta^*(\theta_0^*) > \eta^*(\theta)$.

(A5) For any $\theta \neq \theta_0^*$, there exist $d(\theta) > 0$ and a function $h_\theta(X)$ with $\int f_0(x)|h_\theta(x)|dx < \infty$ such that

$$\sup_{\theta':|\theta'-\theta_0^*|<d(\theta)} \log\frac{f(X;\theta')}{f(X;\theta_0^*)} < h_\theta(X).$$

(A6) For a sufficiently large $K > 0$, there exists a function $h_0(X)$ with $\int f_0(x)h_0(x)dx < 0$ such that

$$\sup_{\theta':|\theta'-\theta_0^*|>K} \log\frac{f(X;\theta')}{f(X;\theta_0^*)} < h_0(X).$$

All of the above assumptions are independent of G which is the distribution of the censoring variable Y_i. In the case that all X_i's are observable, the MLE converges to θ_0^* in probability under these assumptions, where θ_0^* is defined by the assumption (A4) and it gives the nearest density in $\mathcal{M}$ to the true density f_0. In this case, the MLE is MKLI-consistent. Considering the MKLI-consistency of the AMLE $\hat{\theta}_n^*$, we also assume

(A7) $\tau_{F_0} \leq \tau_G$ for $\tau_{F_0} = \inf\{x : F_0(x) = 1\}$, $\tau_G = \inf\{y : G(y) = 1\}$.

If $\tau_{F_0} > \tau_G$ holds, $\Pr\{\delta_i = 0 \mid X_i \geq \tau_G\} = 1$, i.e. X_i in $[\tau_G, \infty)$ is certainly censored. The assumption (A7) guarantees observability of X_i on the whole of the support of $f_0(x)$. In a large number of practical situations, $\tau_{F_0} = \tau_G = \infty$, hence the assumption (A7) is satisfied.

The following theorem states the MKLI-consistency of $\hat{\theta}_n^*$, and it can be proved using the law of large numbers of the Kaplan-Meier integral by Stute and Wang (1993).

THEOREM 1. *Under the conditions* (A1)–(A7), *the AMLE* $\hat{\theta}_n^*$ *converges to* θ_0^* *in probability as* $n \to \infty$.

The conditions (A1)–(A6) are well-known conditions under which the MLE is MKLI-consistent for uncensored data (Takeuchi (1974)). In censored case, the assumption (A7) is essential for MKLI-consistency of the AMLE $\hat{\theta}_n^*$.

We next consider the asymptotic distribution of $\hat{\theta}_n^*$. Let $\bar{F}_0(x) = 1 - F_0(x)$, $\bar{G}(y) = 1 - G(y)$ and $\bar{H}(z) = \bar{F}_0(z)\bar{G}(z)$, and we assume the following conditions:

(B1) For every x, the partial derivative of $f(x;\theta)$ of the third order with respect to θ exists.

(B2) For any $\theta' \in \Theta$, there exist a positive number $c > 0$ and a function $M(x)$ such that

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x;\theta)\right| \leq M(x) \qquad \text{for all} \quad \theta \in (\theta' - c, \theta' + c)$$

and $\int M(x)f_0(x)dx < \infty$

(B3) $\int f_0(x)\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)dx < \infty$

(B4) $\int f_0(x)\left\{\frac{\partial}{\partial\theta}\log f(x;\theta)\right\}^2\{\bar{G}(x)\}^{-1}dx < \infty$

(B5) $\int f_0(x) \left|\frac{\partial}{\partial\theta} \log f(x;\theta)\right| \{\bar{H}(x)\}^{-1/2} dx < \infty$

The conditions (B1)and (B2) are the well-known regularity conditions for the asymptotic normality of the MLE in the case that all X_i's are observable. The condition (B3) corresponds to the condition of existence of the Fisher information in the case that $\mathcal{M}$ includes f_0. It is also noted that these three conditions are independent of G. The conditions (B4) and (B5) are essential for the asymptotic normality of $\hat{\theta}_n^*$, and these are needed for the asymptotic normality of the score statistic $\sqrt{n}\partial l_n^*(\theta)/\partial\theta$. In the example of the previous section (f_0 and g are the Weibull distributions and $\mathcal{M}$ is the exponential model), a necessary and sufficient condition for (B4) and (B5) is $\lambda > \xi$, which means $\Pr\{\delta_i = 0\} < 1/2$.

Under the above conditions, we can prove the following theorem by using the results of Stute and Wang (1993) and Stute (1995).

THEOREM 2. *Under the conditions* (A1)–(A7), (B1)–(B5), $\sqrt{n}(\hat{\theta}_n^* - \theta_0^*) \to^{\mathrm{d}} N(0, \sigma^{*2}(\theta_0^*))$, *where* $\to^{\mathrm{d}}$ *denotes convergence in distribution*, $\sigma^{*2}(\theta) = I^*(\theta)/\{J^*(\theta)\}^2$,

$$J^*(\theta) = -\int f_0(x)\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)dx \qquad \textit{and}$$

$$I^*(\theta) = \int \frac{f_0(x)}{\bar{G}(x)}\left\{\frac{\partial \log f(x;\theta)}{\partial\theta}\right\}^2 dx - \int \frac{\bar{F}_0(x)}{\{\bar{H}(x)\}^2}\left\{\int_x^\infty \frac{\partial \log f(u;\theta)}{\partial\theta} dF_0(u)\right\}^2 dG(x).$$

Note that in $I^*(\theta)$, by putting $G \equiv 0$ (no censoring) formally, the first term is equal to

$$\int f_0(x)\left\{\frac{\partial}{\partial\theta}\log f(x;\theta)\right\}^2 dx$$

and the second term vanishes. Thus, in this case $\sigma^{*2}(\theta)$ reduces to the well-known variance formula (Takeuchi (1976)).

When $\mathcal{M}$ includes f_0, the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n^* - \theta_0)$ is normal with mean zero and variance

$$\frac{\int \frac{f(x;\theta_0)}{\bar{G}(x)}\left\{\frac{\partial \log f(x;\theta_0)}{\partial\theta}\right\}^2 dx - \int \frac{\bar{F}(x;\theta_0)}{\{\bar{G}(x)\}^2}\left\{\frac{\partial \log \bar{F}(x;\theta_0)}{\partial\theta}\right\}^2 dG(x)}{\left[\int f(x;\theta_0)\left\{\frac{\partial}{\partial\theta}\log f(x;\theta_0)\right\}^2 dx\right]^2}. \tag{2.1}$$

If there is no censorship ($G \equiv 0$), the asymptotic variance (2.1) reduces to

$$\left[\int f(x;\theta_0)\left\{\frac{\partial}{\partial\theta}\log f(x;\theta_0)\right\}^2 dx\right]^{-1}.$$

We next consider the asymptotic properties of the MLE $\hat{\theta}_n$, which attains the maximum of $l_n(\theta)$ defined by (1.3). Under the suitable regularity conditions (Andersen *et al.* (1993)), the MLE $\hat{\theta}_n$ converges to $\tilde{\theta}_0$ in probability as $n \to \infty$, where $\tilde{\theta}_0$ is defined by

$$\int f_0(x)\bar{G}(x)\log f(x;\tilde{\theta}_0)dx + \int \bar{F}_0(x)\log \bar{F}(x;\tilde{\theta}_0)G(dx)$$

$$= \max_{\theta\in\Theta}\left\{\int f_0(x)\bar{G}(x)\log f(x;\theta)dx + \int \bar{F}_0(x)\log \bar{F}(x;\theta)dG(x)\right\}.$$

In general, $\tilde{\theta}_0$ is not equal to θ_0^*. If there is no censorship ($G \equiv 0$), $\tilde{\theta}_0$ is equal to θ_0^*. And when $\mathcal{M}$ contains f_0, it holds that $\tilde{\theta}_0 = \theta_0^* = \theta_0$ (the true parameter). In this case, the MLE $\hat{\theta}_n$ is a consistent estimator of θ_0, and under suitable conditions, $n^{1/2}(\hat{\theta}_n - \theta_0)$ converges to $N(0, \sigma^2(\theta_0))$ in distribution, where $\sigma^2(\theta) = 1/I(\theta)$ and

$$(2.2)\quad I(\theta) = \int f(x;\theta)\bar{G}(x)\left\{\frac{\partial}{\partial\theta}\log f(x;\theta)\right\}^2 dx + \int \bar{F}(x;\theta)\left\{\frac{\partial}{\partial\theta}\log \bar{F}(x;\theta)\right\}^2 dG(x).$$

When the model $\mathcal{M}$ does not always contain f_0, the AMLE $\hat{\theta}_n^*$ is MKLI-consistent and the MLE $\hat{\theta}_n$ is not. Hence, in this case, the AMLE is better than the MLE. On the other hand, when $\mathcal{M}$ contains f_0, both are consistent estimator of the true θ_0.

THEOREM 3. *When the model $\mathcal{M}$ contains the true density f_0, the asymptotic relative efficiency of the AMLE $\hat{\theta}_n^*$ with respect to the MLE $\hat{\theta}_n$ is not greater than one.*

This theorem relates to loss of information of AMLE pointed out by Oakes (1986) p.182. Although the definition of the AMLE is slightly different from Oakes' as we have mentioned before, the loss of information of $\hat{\theta}_n^*$ is similar to Oakes' AMLE under correct specification of the parametric model.

3. Simulation results

In the previous sections, we consider the estimator $\hat{\theta}_n^*$ by regarding

$$\frac{1}{n}l_n^*(\theta) = \int \log f(x;\theta)d\hat{F}_n(x)$$

as an estimator of $\int f_0(x)\log f(x;\theta)dx$. However, Mauro (1985) and Stute (1994) pointed out that $\int \varphi(x)d\hat{F}_n(x)$ has a nonnegligible bias as an estimator of $\int \varphi(x)f_0(x)dx$, for every integrable φ. Stute and Wang (1994) suggested a jackknife corrected Kaplan-Meier integral

$$\int \varphi(x)d\hat{F}_n(x) + A_n\varphi(Z_{(n)})$$

as an estimator of $\int \varphi(x)f_0(x)dx$, where

$$(3.1)\qquad A_n = \frac{n-1}{n}\delta_{[n]}(1-\delta_{[n-1]})\prod_{j=1}^{n-2}\left(\frac{n-1-j}{n-j}\right)^{\delta_{[j]}}$$

and $Z_{(i)}$ and $\delta_{[i]}$ are defined in (1.2). They reported that this estimator has smaller bias than $\int \varphi(x)d\hat{F}_n(x)$ for $\varphi(x) = x$. However, they also reported that the jackknifing has led to an increase in variance. We consider an estimator $\hat{\theta}_n^{*JK}$ which attains the maximum of $l_n^{*JK}(\theta)$ in Θ, where

$$l_n^{*JK}(\theta) = l_n^*(\theta) + nA_n\log f(Z_{(n)};\theta).$$

Table 1. Simulation results for MSE and SB of $\hat{\theta}_n$, $\hat{\theta}_n^*$ and $\hat{\theta}_n^{*JK}$; $\lambda = 2.0$

β		1/3	5/6	1	5/4	2
$\Pr\{\delta = 0\}$		0.442	0.360	0.333	0.296	0.200
			$n = 20$			
$\hat{\theta}_n$	MSE	7.345	**0.629**	**0.371**	**0.251**	0.157
	SB	2.969	0.105	0.011	0.006	0.044
$\hat{\theta}_n^*$	MSE	4.894	0.736	0.461	0.273	**0.090**
	SB	1.613	0.156	0.071	0.026	0.003
$\hat{\theta}_n^{*JK}$	MSE	**3.884**	0.678	0.465	0.306	0.112
	SB	1.114	0.039	0.009	0.001	0.000
			$n = 50$			
$\hat{\theta}_n$	MSE	1.935	**0.228**	**0.131**	0.104	0.095
	SB	1.221	0.052	0.002	0.014	0.052
$\hat{\theta}_n^*$	MSE	0.733	0.233	0.158	**0.096**	**0.034**
	SB	0.319	0.035	0.015	0.005	0.000
$\hat{\theta}_n^{*JK}$	MSE	**0.576**	0.250	0.187	0.123	0.043
	SB	0.203	0.002	0.000	0.000	0.000
			$n = 100$			
$\hat{\theta}_n$	MSE	1.128	0.123	**0.063**	0.062	0.076
	SB	0.855	0.039	0.000	0.018	0.054
$\hat{\theta}_n^*$	MSE	0.263	**0.110**	0.075	**0.046**	**0.017**
	SB	0.121	0.012	0.005	0.001	0.000
$\hat{\theta}_n^{*JK}$	MSE	**0.206**	0.135	0.100	0.065	0.021
	SB	0.070	0.000	0.000	0.000	0.000

The smallest MSE is written in bold script.

In this section, we discuss comparison of the estimators based on a simulation study. Following the example of the previous sections, we assume that

$$X_i \sim f_0(x) = \lambda\beta(\lambda x)^{\beta-1}\exp\{-(\lambda x)^\beta\}, \qquad Y_i \sim g(y) = \beta y^{\beta-1}\exp\{-y^\beta\}$$

and

$$\mathcal{M} = \{f(x;\theta) = \theta\exp(-\theta x); \theta > 0\}.$$

Thus the parameter value we should estimate is $\theta_0^* = \lambda/\Gamma(1+\beta^{-1})$, and the MLE is given by $\hat{\theta}_n = (\sum_{i=1}^n \delta_i)/(\sum_{i=1}^n Z_i)$. From an expression of the Kaplan-Meier integral by Stute and Wang (1994), the AMLE is given by $\hat{\theta}_n^* = 1/\sum_{i=1}^n W_i Z_{(i)}$, where

$$W_i = \frac{\delta_{[i]}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j]}}, \qquad i = 1, 2, \ldots, n.$$

The estimator $\hat{\theta}_n^{*JK}$ is expressed as

$$\hat{\theta}_n^{*JK} = \frac{1 + A_n}{A_n Z_{(n)} + \sum_{i=1}^n W_i Z_{(i)}}.$$

Table 2. Simulation results for MSE and SB of $\hat{\theta}_n$, $\hat{\theta}_n^*$ and $\hat{\theta}_n^{*JK}$; $\beta = 1.0$

n		20			50			100		
λ		1/2	1	2	1/2	1	2	1/2	1	2
$\Pr\{\delta = 0\}$		2/3	1/2	1/3	2/3	1/2	1/3	2/3	1/2	1/3
$\hat{\theta}_n$	MSE	**0.045**	**0.122**	**0.371**	**0.016**	**0.043**	**0.131**	**0.008**	**0.021**	**0.063**
	SB	0.001	0.003	0.011	0.000	0.000	0.002	0.000	0.000	0.000
$\hat{\theta}_n^*$	MSE	0.125	0.199	0.461	0.048	0.070	0.158	0.025	0.034	0.075
	SB	0.066	0.061	0.071	0.025	0.018	0.015	0.013	0.007	0.005
$\hat{\theta}_n^{*JK}$	MSE	0.074	0.173	0.465	0.033	0.077	0.187	0.022	0.047	0.100
	SB	0.007	0.005	0.009	0.000	0.000	0.000	0.000	0.000	0.000

The smallest MSE is written in bold script.

Sample size are $n = 20$, 50, and 100. For a fixed $\lambda = 2.0$, varying β's and each n, 100000 samples were drawn and the mean squared error (MSE) and the squared bias (SB) of each estimator were computed; results are shown in Table 1. The probability that X_i is censored is $\Pr(\delta_i = 0) = 1/(1 + \lambda^{\beta})$, and $\lambda = 2.0$ was chosen so that the probability is 1/3 when $\beta = 1.0$ (this means $f_0 \in \mathcal{M}$). After exploring various β, five values of β (1/3, 5/6, 1, 5/4, 2) were chosen. When $\beta = 5/6$ or 5/4, the assumed model $\mathcal{M}$ can be considered to be near to the true distribution. When $\beta = 1/3$ or 2, it is far from the true.

Similarly, a simulation for a fixed $\beta = 1.0$ and varying λ's was carried out. Results for $\lambda = 1/2$, 1 and 2 are shown in Table 2. The values of $\lambda = 1/2$, 1 and 2 represent 2/3, 1/2 and 1/3 censoring probability, respectively.

From Table 1, it is seen that an estimator which has the smallest MSE for all β does not exist. When β is near to one, the MLE $\hat{\theta}_n$ is best among three estimators. However it has larger bias than $\hat{\theta}_n^{*JK}$. When β is far from one, either $\hat{\theta}_n^*$ or $\hat{\theta}_n^{*JK}$ is best, and superiority of them to $\hat{\theta}_n$ is remarkable for large sample size. For large β, $\hat{\theta}_n^*$ is best, and reversely $\hat{\theta}_n^{*JK}$ is best for small β. It occurs because the censoring probability $\Pr\{\delta = 0\}$ decreases with β. If the probability is small, both $\hat{\theta}_n^*$ and $\hat{\theta}_n^{*JK}$ have small biases, and their MSEs mainly depend on the variances. This is the reason $\hat{\theta}_n^*$ is best for large β. On the other hand, if the censoring probability is large, the biases are serious and hence $\hat{\theta}_n^{*JK}$ is best. On the whole, $\hat{\theta}_n^{*JK}$ has small bias and large variance.

From Table 2, we can see that the MLE $\hat{\theta}_n$ is best among three estimators when the assumed model contains the true density. This superiority is remarkable in case of heavy censorship. This implies that in such case much information is lost by AMLE relative to MLE. In comparison between $\hat{\theta}_n^*$ and $\hat{\theta}_n^{*JK}$, $\hat{\theta}_n^{*JK}$ is better than $\hat{\theta}_n^*$ in case of heavy censorship.

4. Information criteria under censorship

In this section, we derive information criteria under right censorship. To do this, we consider the multiparameter case. Let $f(x; \boldsymbol{\theta})$ be an assumed parametric density with p-dimensional vector of unknown parameters. The information criteria such as AIC (Akaike (1973)), TIC (Takeuchi (1976)) and GIC (Konishi and Kitagawa (1996)) are derived as asymptotically unbiased estimators of

$$\int f_0(x) \log f(x; \boldsymbol{\eta}) dx,$$

where $\boldsymbol{\eta}$ is some estimator of $\boldsymbol{\theta}$. In AIC and TIC, the MLE is used as $\boldsymbol{\eta}$, and more general estimator (functional estimator) is used in GIC. In the censored case, we consider the AMLE $\hat{\boldsymbol{\theta}}_n^*$ and the MLE $\hat{\boldsymbol{\theta}}_n$ as $\boldsymbol{\eta}$.

We first consider estimation of

$$\int f_0(x)\log f(x;\hat{\boldsymbol{\theta}}_n^*)dx. \tag{4.1}$$

Expanding $\log f(x;\hat{\boldsymbol{\theta}}_n^*)$ as a Taylor series about $\boldsymbol{\theta}_0^*$, we have an approximation

$$\int f_0(x)\log f(x;\hat{\boldsymbol{\theta}}_n^*)dx \sim \int f_0(x)\log f(x;\boldsymbol{\theta}_0^*)dx - \frac{1}{2}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*)'\boldsymbol{J}^*(\boldsymbol{\theta}_0^*)(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*), \tag{4.2}$$

where

$$\boldsymbol{J}^*(\boldsymbol{\theta}_0^*) = -\int \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(x;\boldsymbol{\theta}_0^*)dF_0(x). \tag{4.3}$$

The first term on the right-hand side of (4.2) can be similarly approximated as

$$\begin{aligned}
&\int f_0(x)\log f(x;\boldsymbol{\theta}_0^*)dx \\
&= \int \log f(x;\hat{\boldsymbol{\theta}}_n^*)d\hat{F}_n(x) + \int\{\log f(x;\boldsymbol{\theta}_0^*) - \log f(x;\hat{\boldsymbol{\theta}}_n^*)\}d\hat{F}_n(x) - M \\
&\sim \int \log f(x;\hat{\boldsymbol{\theta}}_n^*)d\hat{F}_n(x) - \left\{\int \frac{\partial}{\partial\boldsymbol{\theta}'}\log f(x;\boldsymbol{\theta}_0^*)d\hat{F}_n(x)\right\}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*) \\
&\quad -\frac{1}{2}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*)'\left\{\int \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(x;\boldsymbol{\theta}_0^*)d\hat{F}_n(x)\right\}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*) - M \\
&\sim \int \log f(x;\hat{\boldsymbol{\theta}}_n^*)d\hat{F}_n(x) - \frac{1}{2}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*)'\boldsymbol{J}^*(\boldsymbol{\theta}_0^*)(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*) - M,
\end{aligned} \tag{4.4}$$

where

$$M = \int \log f(x;\boldsymbol{\theta}_0^*)d\hat{F}_n(x) - \int f_0(x)\log f(x;\boldsymbol{\theta}_0^*)dx.$$

From (4.2) and (4.4), we have

$$\int f_0(x)\log f(x;\hat{\boldsymbol{\theta}}_n^*)dx \sim \int \log f(x;\hat{\boldsymbol{\theta}}_n^*)d\hat{F}_n(x) - (\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*)'\boldsymbol{J}^*(\boldsymbol{\theta}_0^*)(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*) - M.$$

Put

$$\begin{aligned}
\boldsymbol{I}^*(\boldsymbol{\theta}_0^*) &= \int \left\{\frac{\partial \log f(x;\boldsymbol{\theta}_0^*)}{\partial\boldsymbol{\theta}}\right\}\left\{\frac{\partial \log f(x;\boldsymbol{\theta}_0^*)}{\partial\boldsymbol{\theta}'}\right\}\{\bar{G}(x)\}^{-1}dF_0(x) \\
&\quad - \int \frac{\bar{F}_0(x)}{\{\bar{H}(x)\}^2}\left\{\int_x^\infty \frac{\partial \log f(u;\boldsymbol{\theta}_0^*)}{\partial\boldsymbol{\theta}}dF_0(u)\right\}\left\{\int_x^\infty \frac{\partial \log f(u;\boldsymbol{\theta}_0^*)}{\partial\boldsymbol{\theta}'}dF_0(u)\right\}dG(x),
\end{aligned}$$

then the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*)$ is given by $\boldsymbol{J}^*(\boldsymbol{\theta}_0^*)^{-1}\boldsymbol{I}^*(\boldsymbol{\theta}_0^*)\boldsymbol{J}^*(\boldsymbol{\theta}_0^*)^{-1}$. Hence we have an approximation

$$E[(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*)'\boldsymbol{J}^*(\boldsymbol{\theta}_0^*)(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*)] \sim \frac{1}{n}\mathrm{trace}\left\{\boldsymbol{J}^*(\boldsymbol{\theta}_0^*)^{-1}\boldsymbol{I}^*(\boldsymbol{\theta}_0^*)\right\}.$$

If there is no censorship ($G \equiv 0$) and the assumed parametric model contains the true density, $E[(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*)'\boldsymbol{J}^*(\boldsymbol{\theta}_0^*)(\hat{\boldsymbol{\theta}}_n^* - \boldsymbol{\theta}_0^*)] \sim p/n$, since $\boldsymbol{J}^*(\boldsymbol{\theta}_0^*) = \boldsymbol{I}^*(\boldsymbol{\theta}_0^*)$. Thus in this case we can estimate (4.1) by

$$\frac{1}{n}\left\{\sum_{i=1}^{n}\log f(X_i;\hat{\boldsymbol{\theta}}_n^*) - p\right\},$$

since $E[M] = 0$. This is Akaike (1973)'s AIC.

In general case, we have to estimate $\boldsymbol{J}^*(\boldsymbol{\theta}_0^*)$, $\boldsymbol{I}^*(\boldsymbol{\theta}_0^*)$ and $E[M]$. By substituting $\hat{\boldsymbol{\theta}}_n^*$, $\hat{F}_n$ and $\hat{G}_n$ into $\boldsymbol{\theta}_0^*$, F_0, G, where $\hat{G}_n$ is the Kaplan-Meier estimator of G, $\boldsymbol{J}^*(\boldsymbol{\theta}_0^*)$ and $\boldsymbol{I}^*(\boldsymbol{\theta}_0^*)$ are estimated by $\hat{\boldsymbol{J}}^*(\hat{\boldsymbol{\theta}}_n^*)$ and $\hat{\boldsymbol{I}}^*(\hat{\boldsymbol{\theta}}_n^*)$, respectively. On the other hand, $E[M]$ is a bias of a Kaplan-Meier integral $\int \log f(x;\boldsymbol{\theta}_0^*)d\hat{F}_n(x)$. Hence we can estimate $E[M]$ by $-A_n \log f(Z_{(n)};\hat{\boldsymbol{\theta}}_n^*)$ using Stute and Wang (1994)'s jackknife bias correction, where A_n is defined in (3.1). Thus we can estimate (4.1) by

$$\int \log f(x;\hat{\boldsymbol{\theta}}_n^*)d\hat{F}_n(x) - \frac{1}{n}\mathrm{trace}\{\hat{\boldsymbol{J}}^*(\hat{\boldsymbol{\theta}}_n^*)^{-1}\hat{\boldsymbol{I}}^*(\hat{\boldsymbol{\theta}}_n^*)\} + A_n \log f(Z_{(n)};\hat{\boldsymbol{\theta}}_n^*), \tag{4.5}$$

and which can be regarded as an information criterion for the AMLE under censorship.

If there is no censorship, the criterion (4.5) coincides with Takeuchi (1976)'s TIC, since the third term is zero, and $\hat{\boldsymbol{J}}^*(\hat{\boldsymbol{\theta}}_n^*)$ and $\hat{\boldsymbol{I}}^*(\hat{\boldsymbol{\theta}}_n^*)$ reduce to

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(X_i;\hat{\boldsymbol{\theta}}_n^*) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X_i;\hat{\boldsymbol{\theta}}_n^*)\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}'}\log f(X_i;\hat{\boldsymbol{\theta}}_n^*)\right\},$$

respectively.

We next derive an information criterion corresponding to the MLE $\hat{\boldsymbol{\theta}}_n$, and which is an asymptotically unbiased estimator of

$$\int f_0(x)\log f(x;\hat{\boldsymbol{\theta}}_n)dx,$$

where the MLE $\hat{\boldsymbol{\theta}}_n$ is a solution of the likelihood equation

$$\sum_{i=1}^{n}\left\{\delta_i\frac{\partial}{\partial\boldsymbol{\theta}}\log f(Z_i;\hat{\boldsymbol{\theta}}_n) + (1-\delta_i)\frac{\partial}{\partial\boldsymbol{\theta}}\log \bar{F}(Z_i;\hat{\boldsymbol{\theta}}_n)\right\} = \boldsymbol{o}.$$

Under some regularity conditions, the MLE $\hat{\boldsymbol{\theta}}_n$ converges in probability to $\tilde{\boldsymbol{\theta}}_0$ as $n \to \infty$, where $\tilde{\boldsymbol{\theta}}_0$ is a solution of an equation

$$\int \bar{G}(x)\frac{\partial}{\partial\boldsymbol{\theta}}\log f(x;\tilde{\boldsymbol{\theta}}_0)dF_0(x) + \int \bar{F}_0(x)\frac{\partial}{\partial\boldsymbol{\theta}}\log \bar{F}(x;\tilde{\boldsymbol{\theta}}_0)dG(x) = \boldsymbol{o}.$$

Expanding $\log f(x;\hat{\boldsymbol{\theta}}_n)$ as a Taylor series about $\tilde{\boldsymbol{\theta}}_0$, we obtain an approximation

$$\begin{aligned}\int f_0(x)\log f(x;\hat{\boldsymbol{\theta}}_n)dx \sim &\int f_0(x)\log f(x;\tilde{\boldsymbol{\theta}}_0)dx \\ &+ \left\{\int f_0(x)\frac{\partial}{\partial\boldsymbol{\theta}'}\log f(x;\tilde{\boldsymbol{\theta}}_0)dx\right\}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_0)\end{aligned}$$

$$-\frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_0)'\boldsymbol{J}^*(\tilde{\boldsymbol{\theta}}_0)(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_0).$$

The first term on the right-hand side can be approximated as

$$\begin{aligned}
&\int f_0(x)\log f(x;\tilde{\boldsymbol{\theta}}_0)dx \\
&\quad = \int \log f(x;\hat{\boldsymbol{\theta}}_n)d\hat{F}_n(x) + \int\{\log f(x;\tilde{\boldsymbol{\theta}}_0) - \log f(x;\hat{\boldsymbol{\theta}}_n)\}d\hat{F}_n(x) - \tilde{M} \\
&\quad \sim \int \log f(x;\hat{\boldsymbol{\theta}}_n)d\hat{F}_n(x) - \left\{\int \frac{\partial}{\partial\boldsymbol{\theta}'}\log f(x;\tilde{\boldsymbol{\theta}}_0)d\hat{F}_n(x)\right\}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_0) \\
&\qquad +\frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_0)'\boldsymbol{J}^*(\tilde{\boldsymbol{\theta}}_0)(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_0) - \tilde{M},
\end{aligned}$$

where $\boldsymbol{J}^*(\boldsymbol{\theta})$ is defined in (4.3) and

$$\tilde{M} = \int \log f(x;\tilde{\boldsymbol{\theta}}_0)d\hat{F}_n(x) - \int f_0(x)\log f(x;\tilde{\boldsymbol{\theta}}_0)dx.$$

Thus we have

$$\begin{aligned}
(4.6)\qquad \int f_0(x)\log f(x;\hat{\boldsymbol{\theta}}_n)dx \sim {} & \int \log f(x;\hat{\boldsymbol{\theta}}_n)d\hat{F}_n(x) - \tilde{M} \\
& +\left\{\int f_0(x)\frac{\partial}{\partial\boldsymbol{\theta}'}\log f(x;\tilde{\boldsymbol{\theta}}_0)dx \right.\\
& \left. -\int \frac{\partial}{\partial\boldsymbol{\theta}'}\log f(x;\tilde{\boldsymbol{\theta}}_0)d\hat{F}_n(x)\right\}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_0).
\end{aligned}$$

Let $\boldsymbol{v}_i(\boldsymbol{\theta})$ be a p-dimensional random vector defined by

$$\boldsymbol{v}_i(\boldsymbol{\theta}) = \delta_i\frac{\partial}{\partial\boldsymbol{\theta}}\log f(Z_i;\boldsymbol{\theta}) + (1-\delta_i)\frac{\partial}{\partial\boldsymbol{\theta}}\log \bar{F}(Z_i;\boldsymbol{\theta}).$$

Then it holds that

$$(4.7)\qquad \hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_0 = \frac{1}{n}\boldsymbol{J}(\tilde{\boldsymbol{\theta}}_0)^{-1}\sum_{i=1}^{n}\boldsymbol{v}_i(\tilde{\boldsymbol{\theta}}_0) + o_p(n^{-1/2}),$$

where

$$\boldsymbol{J}(\boldsymbol{\theta}) = \int \bar{G}(x)\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(x;\boldsymbol{\theta})dF_0(x) + \int \bar{F}_0(x)\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log \bar{F}(x;\boldsymbol{\theta})dG(x).$$

From Theorem 1.1 of Stute (1995), we have an expression

$$(4.8)\quad \int \frac{\partial}{\partial\boldsymbol{\theta}}\log f(x;\tilde{\boldsymbol{\theta}}_0)d\hat{F}_n(x) - \int f_0(x)\frac{\partial}{\partial\boldsymbol{\theta}}\log f(x;\tilde{\boldsymbol{\theta}}_0)dx = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}_i(\tilde{\boldsymbol{\theta}}_0) + o_p(n^{-1/2}),$$

where

$$\boldsymbol{u}_i(\boldsymbol{\theta}) = \delta_i\{\bar{G}(Z_i)\}^{-1}\frac{\partial\log f(Z_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} + (1-\delta_i)\gamma_1(Z_i;\boldsymbol{\theta}) - \gamma_2(Z_i;\boldsymbol{\theta})$$

$$-\int f_0(x)\frac{\partial \log f(x;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}dx,$$
$$\gamma_1(x;\boldsymbol{\theta}) = \{\bar{H}(x)\}^{-1}\int_x^{\infty}\frac{\partial \log f(u;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}dF_0(u) \quad \text{and}$$
$$\gamma_2(x;\boldsymbol{\theta}) = \int_{-\infty}^{x}\frac{\gamma_1(u;\boldsymbol{\theta})}{\bar{G}(u)}dG(u).$$

Noting $E\{\boldsymbol{v}_i(\tilde{\boldsymbol{\theta}}_0)\} = E\{\boldsymbol{u}_i(\tilde{\boldsymbol{\theta}}_0)\} = \boldsymbol{o}$, from (4.7) and (4.8), we can approximate the expectation of the third term on the right-hand side of (4.6) by

$$(4.9) \qquad -\frac{1}{n}E\{\boldsymbol{u}_i'(\tilde{\boldsymbol{\theta}}_0)\boldsymbol{J}(\tilde{\boldsymbol{\theta}}_0)^{-1}\boldsymbol{v}_i(\tilde{\boldsymbol{\theta}}_0)\} = -\frac{1}{n}\text{trace}[\boldsymbol{J}(\tilde{\boldsymbol{\theta}}_0)^{-1}E\{\boldsymbol{v}_i(\tilde{\boldsymbol{\theta}}_0)\boldsymbol{u}_i'(\tilde{\boldsymbol{\theta}}_0)\}].$$

Put

$$\boldsymbol{K}(\boldsymbol{\theta}) = \int\left\{\frac{\partial}{\partial \boldsymbol{\theta}}\log f(x;\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial \boldsymbol{\theta}'}\log f(x;\boldsymbol{\theta})\right\}dF_0(x)$$

and

$$\boldsymbol{\xi}(x;\boldsymbol{\theta}) = \int_x^{\infty}\bar{G}(u)\frac{\partial}{\partial \boldsymbol{\theta}}\log f(u;\boldsymbol{\theta})dF_0(u) + \int_x^{\infty}\bar{F}_0(u)\frac{\partial}{\partial \boldsymbol{\theta}}\log \bar{F}(u;\boldsymbol{\theta})dG(u),$$

then we have

$$(4.10) \qquad E\{\boldsymbol{v}_i(\tilde{\boldsymbol{\theta}}_0)\boldsymbol{u}_i'(\tilde{\boldsymbol{\theta}}_0)\} = \boldsymbol{K}(\tilde{\boldsymbol{\theta}}_0) + \boldsymbol{D}(\tilde{\boldsymbol{\theta}}_0),$$

where

$$\boldsymbol{D}(\boldsymbol{\theta}) = \int \bar{F}_0(x)\left[\frac{\partial}{\partial \boldsymbol{\theta}}\log \bar{F}(x;\boldsymbol{\theta}) - \{\bar{H}(x)\}^{-1}\boldsymbol{\xi}(x;\boldsymbol{\theta})\right]\gamma_1'(x;\boldsymbol{\theta})dG(x).$$

From (4.6), (4.9) and (4.10), we obtain an approximation

$$E\left\{\int f_0(x)\log f(x;\hat{\boldsymbol{\theta}}_n)dx - \int \log f(x;\hat{\boldsymbol{\theta}}_n)d\hat{F}_n(x)\right\}$$
$$\sim -\frac{1}{n}\text{trace}[\boldsymbol{J}(\tilde{\boldsymbol{\theta}}_0)^{-1}\{\boldsymbol{K}(\tilde{\boldsymbol{\theta}}_0) + \boldsymbol{D}(\tilde{\boldsymbol{\theta}}_0)\}] - E(\tilde{M}).$$

If $f_0 \in \mathcal{M}$, then $f_0(x) = f(x;\tilde{\boldsymbol{\theta}}_0)$ and hence

$$\{\bar{H}(x)\}^{-1}\boldsymbol{\xi}(x;\tilde{\boldsymbol{\theta}}_0) = \{\bar{F}(x;\tilde{\boldsymbol{\theta}}_0)\}^{-1}\int_x^{\infty}f(u;\tilde{\boldsymbol{\theta}}_0)\frac{\partial}{\partial \boldsymbol{\theta}}\log f(u;\tilde{\boldsymbol{\theta}}_0)du = \frac{\partial}{\partial \boldsymbol{\theta}}\log \bar{F}(x;\tilde{\boldsymbol{\theta}}_0).$$

Thus in this case $\boldsymbol{D}(\tilde{\boldsymbol{\theta}}_0) = \boldsymbol{O}$. It is also obvious that $\boldsymbol{D}(\tilde{\boldsymbol{\theta}}_0) = \boldsymbol{O}$ if there is no censorship ($G \equiv 0$). In these special cases, we do not have to estimate $\boldsymbol{D}(\tilde{\boldsymbol{\theta}}_0)$. In general case, however, we have to estimate not only $\boldsymbol{J}(\tilde{\boldsymbol{\theta}}_0)$, $\boldsymbol{K}(\tilde{\boldsymbol{\theta}}_0)$ and $E(\tilde{M})$ but also $\boldsymbol{D}(\tilde{\boldsymbol{\theta}}_0)$. Since $E(\tilde{M})$ is a bias of a Kaplan-Meier integral, we can estimate $E(\tilde{M})$ by $-A_n \log f(Z_{(n)};\hat{\boldsymbol{\theta}}_n)$. The matrices $\boldsymbol{J}(\tilde{\boldsymbol{\theta}}_0)$, $\boldsymbol{K}(\tilde{\boldsymbol{\theta}}_0)$ and $\boldsymbol{D}(\tilde{\boldsymbol{\theta}}_0)$ are estimated by substituting $\hat{\boldsymbol{\theta}}_n$, $\hat{F}_n$ and $\hat{G}_n$ into $\tilde{\boldsymbol{\theta}}_0$, F_0 and G, respectively. Denote by $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_n)$, $\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}}_n)$ and $\hat{\boldsymbol{D}}(\hat{\boldsymbol{\theta}}_n)$ the estimators of $\boldsymbol{J}(\tilde{\boldsymbol{\theta}}_0)$, $\boldsymbol{K}(\tilde{\boldsymbol{\theta}}_0)$ and $\boldsymbol{D}(\tilde{\boldsymbol{\theta}}_0)$, respectively. We obtain an information criterion corresponding to the MLE as

$$(4.11) \quad \int \log f(x;\hat{\boldsymbol{\theta}}_n)d\hat{F}_n(x) - \frac{1}{n}\text{trace}[\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}_n)^{-1}\{\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}}_n) + \hat{\boldsymbol{D}}(\hat{\boldsymbol{\theta}}_n)\}] + A_n \log f(Z_{(n)};\hat{\boldsymbol{\theta}}_n).$$

If there is no censorship, the criterion (4.11) coincides with TIC.

The two information criteria (4.5) and (4.11) can be used to select estimation procedures. If (4.5) is greater than (4.11), then the AMLE is better than the MLE, otherwise the MLE better than the AMLE.

5. Concluding remarks

In many practical situations, it seems that the assumed model $\mathcal{M}$ does not contain the true density f_0. In the parametric estimation for the censored data, it is very important that whether $\mathcal{M}$ contains f_0 or not. When $\mathcal{M}$ does not contain f_0, the maximum likelihood estimation does not provide the nearest density to f_0 even if the sample size n is sufficiently large. The parametric estimation for the censored data has not been discussed from this point of view. In this paper, we showed that the AMLE is MKLI-consistent but the MLE is not. The AMLE is worse than the MLE when $\mathcal{M}$ contains f_0, and it is better when not so.

As pointed out by Miller (1983), the Kaplan-Meier estimator can be inefficient compared to parametric survival estimators. If we can firmly believe that the assumed model is correct, we should analyze data by using the parametric model and MLE. If the parametric model is roughly correct, it is worth considering the parametric model and AMLE. In this sense, AMLE is an intermediate between the nonparametric approach (Kaplan-Meier estimator) and the parametric MLE approach.

On these grounds we have come to the conclusion that the assumed model must be checked carefully in analysis of censored data. In particular, if the values of MLE and AMLE are significantly different for large n, the possibility of misspecification is strong. In this paper we derived information criteria for MLE and AMLE, and which are extensions of Takeuchi (1976)'s TIC. The information criteria can be used not only for selecting parametric models but also for selecting estimation procedures between MLE and AMLE.

6. Proofs

Proof of Theorem 1. Appling the law of large numbers of Stute and Wang (1993), it can be shown

$$\lim_{n\to\infty} \Pr\left\{ \sup_{|\theta-\theta_0^*|>\epsilon} \int \log \frac{f(x;\theta)}{f(x;\theta_0^*)} d\hat{F}_n(x) < 0 \right\} = 1,$$

in a similar way to the proof of Lemma in Takeuchi (1974), p. 156. From this, we have

$$\Pr\{|\hat{\theta}_n^* - \theta_0^*| > \epsilon\} \le \Pr\left\{ \sup_{|\theta-\theta_0^*|>\epsilon} \int \log \frac{f(x;\theta)}{f(x;\theta_0^*)} d\hat{F}_n(x) \ge 0 \right\} \to 0$$

for any $\epsilon > 0$. □

Proof of Theorem 2. Under the assumptions, it holds that $l_n^{*'}(\hat{\theta}_n^*) = 0$, and expanding $l_n^{*'}(\hat{\theta}_n^*)$ as a Taylor series about θ_0^*, we have

$$0 = l_n^{*'}(\hat{\theta}_n^*) = l_n^{*'}(\theta_0^*) + l_n^{*''}(\theta_0^*)(\hat{\theta}_n^* - \theta_0^*) + \frac{1}{2} l_n^{*'''}(\tilde{\theta})(\hat{\theta}_n^* - \theta_0^*)^2,$$

where $\tilde{\theta}$ lies between $\hat{\theta}_n^*$ and θ_0^*, so that

$$\sqrt{n}(\hat{\theta}_n^* - \theta_0^*) = \frac{-\dfrac{1}{\sqrt{n}} l_n^{*'}(\theta_0^*)}{\dfrac{1}{n} l_n^{*''}(\theta_0^*) + \dfrac{1}{2n} l_n^{*'''}(\tilde{\theta})(\hat{\theta}_n^* - \theta_0^*)}. \tag{6.1}$$

From the law of large numbers of Stute and Wang (1993), $l_n^{*''}(\theta_0^*)/n \to -J^*(\theta_0^*)$ a.s.. Under (B4) and (B5), which correspond to the condition (1.9) and (1.10) of Stute (1995), respectively, the numerator of (6.1)

$$-\frac{1}{\sqrt{n}} l_n^{*'}(\theta_0^*) = -\sqrt{n} \int \frac{\partial}{\partial \theta} \log f(x; \theta_0^*) d\hat{F}_n(x)$$

converges in distribution to $N(0, I^*(\theta_0^*))$ by the central limit theorem of Stute (1995). From the condition (B2), it follows that $l_n^{*'''}(\tilde{\theta})/n$ is $O_p(1)$. Hence the quantity $l_n^{*'''}(\tilde{\theta})(\hat{\theta}_n^* - \theta_0^*)/(2n)$ converges to 0 in probability. □

Proof of Theorem 3. From Theorem 2.1 of Lehmann (1983), the asymptotic relative efficiency of $\hat{\theta}_n^*$ with respect to $\hat{\theta}_n$ is given by $\sigma^2(\theta_0)/\sigma_0^{*2}(\theta_0)$, where $\sigma^2(\theta_0)$ and $\sigma_0^{*2}(\theta_0)$ are defined in (2.2) and (2.1), respectively. Put

$$J_0^*(\theta_0) = \int f(x;\theta_0) \left\{ \frac{\partial}{\partial \theta} \log f(x;\theta_0) \right\}^2 dx \quad \text{and}$$
$$I_0^*(\theta_0) = \int \frac{f(x;\theta_0)}{\bar{G}(x)} \left\{ \frac{\partial \log f(x;\theta_0)}{\partial \theta} \right\}^2 dx - \int \frac{\bar{F}(x;\theta_0)}{\{\bar{G}(x)\}^2} \left\{ \frac{\partial \log \bar{F}(x;\theta_0)}{\partial \theta} \right\}^2 dG(x),$$

then $\sigma_0^{*2}(\theta_0) = I_0^*(\theta_0)/\{J_0^*(\theta_0)\}^2$ and $I_0^*(\theta_0)$ can be expressed as

$$I_0^*(\theta_0) = E[\{(\bar{G}(Z_i))^{-1} L_i\}^2],$$

where

$$L_i = \delta_i \frac{\partial}{\partial \theta} \log f(Z_i;\theta_0) + (1-\delta_i) \frac{\partial}{\partial \theta} \log \bar{F}(Z_i;\theta_0).$$

On the other hand, $I(\theta_0)$ defined by (2.4) can be written as $I(\theta_0) = E[L_i^2]$. Hence we obtain

$$\frac{\sigma^2(\theta_0)}{\sigma_0^{*2}(\theta_0)} = \frac{\{J_0^*(\theta_0)\}^2}{E[L_i^2] E[\{(\bar{G}(Z_i))^{-1} L_i\}^2]}.$$

Noting

$$\begin{aligned} E[(\bar{G}(Z_i))^{-1} L_i^2] &= J_0^*(\theta_0) + E\left[(1-\delta_i)(\bar{G}(Z_i))^{-1} \left\{ \frac{\partial}{\partial \theta} \log \bar{F}(Z_i;\theta_0) \right\}^2 \right] \\ &\geq J_0^*(\theta_0) > 0, \end{aligned}$$

we have

$$\frac{\sigma^2(\theta_0)}{\sigma_0^{*2}(\theta_0)} \leq \frac{\{E[(\bar{G}(Z_i))^{-1} L_i^2]\}^2}{E[L_i^2] E[\{(\bar{G}(Z_i))^{-1} L_i\}^2]} \leq 1.$$

The last inequality holds from Cauchy-Schwarz inequality. □

Acknowledgements

Thanks are due to referees who made a number of helpful comments and bought to our attentions several references.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory* (eds. B. N. Petrov, and F. Csaki), 267–281, Akademiai Kiado, Budapest. (Reproduced (1992) *Breakthroughs in Statistics* (eds. S. Kotz and N. L. Johnson), **1**, 610–624, Springer, New York.)

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer, New York.

Cramèr, H. (1946). *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N. J.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*, Wiley, New York.

Kaplan, E. L. and Meier, P. (1958). Non-parametric estimation from incomplete observations, *J. Amer. Statist. Assoc.*, **53**, 457–481.

Konishi, S. and Kitagawa, G. (1996). Generalized information criteria in model selection, *Biometrika*, **83**, 875–890.

Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley, New York.

Mauro, D. (1985). A combinatoric approach to the Kaplan-Meier estimation, *Ann. Statist.*, **13**, 142–149.

Miller, R. G., Jr. (1983). "What Price Kaplan-Meier?", *Biometrics*, **39**, 1077–1081.

Oakes, D. (1986). An approximate likelihood procedure for censored data, *Biometrics*, **42**, 177–182.

Stute, W. (1994). The bias of Kaplan-Meier integrals, *Scand. J. Statist.*, **21**, 475–484.

Stute, W. (1995). The central limit theorem under random censorship, *Ann. Statist.*, **23**, 422–439.

Stute, W. and Wang, J. L. (1993). The strong law under random censorship, *Ann. Statist.*, **21**, 1591–1607.

Stute, W. and Wang, J. L. (1994). The jackknife estimate of a Kaplan-Meier integral, *Biometrika*, **81**, 602–606.

Takeuchi, K. (1974). *Toukeiteki Suitei no Zenkin Riron*, Kyoiku-Shuppan, Japan (in Japanese).

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models, *Mathematical Sciences*, **153**, 12–18 (in Japanese).

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.*, **20**, 595–601.

Wang, J. L. (1995). M-estimators for censored data: strong consistency, *Scand. J. Statist.*, **22**, 197–205.