Goodness of fit tests with interval censored data

JIAN-JIAN REN

Tulane University and University of Central Florida

ABSTRACT. Cramér–von Mises type goodness of fit tests for interval censored data case 2 are proposed based on a resampling method called the leveraged bootstrap, and their asymptotic consistency is shown. The proposed tests are computationally efficient, and in fact can be applied to other types of censored data, including right censored data, doubly censored data and (mixture of) case k interval censored data. Some simulation results and an example from AIDS research are presented.

Key words: bootstrap, leveraged bootstrap, parametric family goodness of fit test

1. Introduction

Incomplete data are frequently encountered in medical follow-up studies and in reliability studies. Partially motivated by problems arising from these studies, analysis of right censored data has been one of the focal points of statistics in the past three decades. While there exist some earlier works, viz. Gehan (1965), Turnbull (1974), among others, recently statisticians are paying more and more attention to some more complicated types of incomplete data, such as doubly censored data and interval censored data, as these data occur in important clinical trials. For instance, doubly censored data were encountered in recent studies on primary breast cancer (Peer *et al.*, 1993; Ren & Peer, 2000), and interval censored data were encountered in AIDS research (Kim *et al.*, 1993). For doubly censored data (Turnbull, 1974; Mykland & Ren, 1996), a goodness of fit test has been studied by Ren (1995a), and Bickel & Ren (1996, 2001). This current paper is concerned with the goodness of fit test with interval censored data.

Precisely, the interval censored sample we consider in this paper is given by $\mathbf{O}_i = (Y_i, Z_i, \delta_i)$, $i = 1, \ldots, n$, with

$$\delta_i = \begin{cases} 1, & \text{if } Z_i < X_i \leq Y_i \\ 2, & \text{if } X_i > Y_i \\ 3, & \text{if } X_i \leq Z_i \end{cases} \tag{1}$$

where $X_1, \ldots, X_n$ is an independently and identically distributed (i.i.d.) non-negative random sample from an underlying distribution function (d.f.) F, and (Y_i, Z_i) are i.i.d. and independent from X_i with $P\{Z_i < Y_i\} = 1$. This is the interval censoring case 2 considered by Groeneboom & Wellner (1992), among others.

To deal with more complicated interval censored cases encountered in practice, Wellner (1995) studied interval censored case k model. More generally, Schick & Yu (2000) studied a mixture of case k models, which was further investigated by Wellner & Zhang (2000) in the context of the mean function of a counting process. The method developed in this paper for interval censored case 2 data can easily be applied to the models in Wellner (1995), Schick & Yu (2000), Wellner & Zhang (2000), provided that some information on the convergence rate of the estimator for F is available in their cases.

In practice, researchers often like to make parametric assumptions on the underlying distributions, say, normal distribution, exponential distribution, etc. Motivated by this, this

paper provides a method for model checking when interval censored data (1) are encountered. Specifically, for the simple goodness of fit test on F given in (1), we consider testing:

$$H_0: F = F_0 \quad \text{vs.} \quad H_1: F \neq F_0, \tag{2}$$

where F_0 is a known continuous d.f., while for the parametric family goodness of fit test on F, we consider testing:

$$H_0: F \in \mathscr{F}_0 \equiv \{F_0(\cdot;\theta)|\theta \in \Theta\} \quad \text{vs.} \quad H_1: F \notin \mathscr{F}_0, \tag{3}$$

where $F_0(\cdot;\theta)$ is a specified d.f. with unknown parameter $\theta \in \Theta \subset \mathbb{R}^q$.

1.1. Testing the simple goodness of fit (2)

For testing (2), when there is no censoring, the Cramér–von Mises test statistic is given by

$$T_n = n \int_0^\infty (F_n(x) - F_0(x))^2 \, dF_0(x), \tag{4}$$

where F_n is the empirical d.f. of $X_1, \ldots, X_n$, and it is known that under H_0,

$$T_n \xrightarrow{D} W, \quad \text{as } n \to \infty \tag{5}$$

where W has a d.f. given in Shorack & Wellner (1986, p. 147). When a censored sample is observed, one may want to use

$$\hat{T}_n = n \int_0^\infty (\hat{F}_n(x) - F_0(x))^2 \, dF_0(x) \tag{6}$$

as the test statistic, where $\hat{F}_n$ is an estimator of F using censored data. But this functional "plug-in" method fails for interval censored data (1). To see this, let $\hat{F}_n$ be the non-parametric maximum likelihood estimator (NPMLE) of F for data (1) (the method for computing $\hat{F}_n$ can be found in Groeneboom & Wellner, 1992). It is known that the convergence rate of $\hat{F}_n$ is slower than $\sqrt{n}$ (for more general interval censored cases, such as in Schick & Yu (2000), it is not clear what is the convergence rate of the NPMLE $\hat{F}_n$). In Geskus & Groeneboom (1999), it was pointed out that for a given point t_0, the convergence rate of $\hat{F}_n(t_0)$ is sometimes $n^{1/3}$ (Wellner, 1995; Groeneboom, 1996), but sometimes *perhaps* $(n \log n)^{1/3}$ (Groeneboom & Wellner, 1992), depending on whether the observation time distribution has sufficient mass along the diagonal point (t_0, t_0). This means that for a given NPMLE $\hat{F}_n$ computed from observed interval censored data (1), the convergence rate of $\hat{F}_n(t_1)$ could be different from that of $\hat{F}_n(t_2)$ for two different points t_1 and t_2. Thus, $\hat{T}_n$ is not a suitable test statistic for interval censored data, because it does not stabilize under H_0 as $n \to \infty$. This is confirmed by the simulation results presented in Fig. 1. Let $\exp(\mu)$ denote the exponential d.f. with mean μ. For sample size $n = 200$, 500 and 1000, Fig. 1 presents three Monte Carlo curves of $\hat{T}_n$ under H_0 in (2) with $F_0 = \exp(1)$, where for each n, the Monte Carlo curve is based on 1000 samples generated for X_i and Y_i from exp(1) and exp(3), respectively, with $Z_i = [(2/3)Y_i - 2.5]$. Clearly these curves appear to be diverging as n increases.

One may note that although it is shown in th. 3.2 of Geskus & Groeneboom (1999) that for some smooth functional $K(\cdot)$ of the NPMLE $\hat{F}_n$ with interval censored data case 2, $\sqrt{n}(K(\hat{F}_n) - K(F))$ is asymptotically normal, their theorem cannot be used for functional $\hat{T}_n$ given in (6) to construct a test statistic for testing (2). This is because with $K(\hat{F}_n) = n^{-1}\hat{T}_n \geq 0$ always and $K(F) = 0$ under H_0, $\hat{T}_n$ cannot possibly be asymptotically normal no matter what normalization is used on it. In fact, from above we know that a proper normalization for $\hat{T}_n$ may not generally exist unless more assumptions are made on the interval censoring case 2

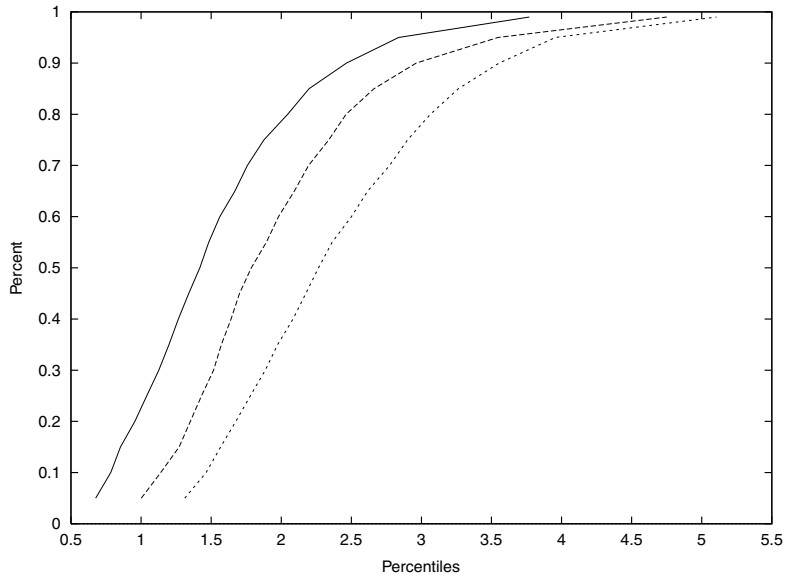

Fig. 1. Monte Carlo curves of $\hat{T}_n$ with interval censored data under H_0: $F = \exp(1)$, where $X \sim \exp(1)$, $Y \sim \exp(3), Z = [(2/3)Y - 2.5]$, $n = 200$ (solid line), $n = 500$ (dashed line), $n = 1000$ (dotted line).

model (1). Thus, the usual methods developed for testing (2) with right censored data (Efron, 1981) or doubly censored data (Ren, 1995a; Bickel & Ren, 2001) do not apply to interval censored data (1). Other resampling methods, such as subsampling (Politis & Romano, 1994; or see a recent book by Politis *et al.*, 1999) and the m out of n bootstrap (Bickel *et al.*, 1997; Bickel & Ren, 2001), do not apply either for the same reason, as all these methods require that the original statistic, suitably normalized (with a known convergence rate), has a limit under the null hypothesis.

In section 3, we propose a method for testing the simple goodness of fit (2) with interval censored data (1) using a new resampling method, called the leveraged bootstrap (Ren, 1995b, 2001), and show that the proposed test is asymptotically consistent with proofs deferred to the appendix. This method is extended to test the parametric family goodness of fit (3) in section 4, while the idea of the leveraged bootstrap (LB) is described in section 2. Some simulation results on the theorems in sections 3 and 4 are presented in section 5, and an example of interval censored data (1) from AIDS research (Kim *et al.*, 1993) is discussed in section 6. Section 7 includes some concluding remarks.

2. Leveraged bootstrap

We begin by recalling the bootstrap principle (Efron, 1979). Suppose the distribution function of statistic $H_n = H_n(X_1, \ldots, X_n)$ is to be approximated, where $X_1, \ldots, X_n$ is an i.i.d. random sample from a distribution function F. Since the empirical distribution function F_n for $X_1, \ldots, X_n$ is asymptotically close (a.s.) to F and since F_n puts equal weight n^{-1} at each observation X_i in the sample, one hopes that if $\hat{X}_{n1}, \ldots, \hat{X}_{nm}$ is a random sample from F_n, where m is a positive integer (see Efron, 1979, for the case $m = n$; and see Bickel *et al.*, 1997, for the case $m = o(n)$), then the distribution of $\hat{H}_m = H_m(\hat{X}_{n1}, \ldots, \hat{X}_{nm})$ is asymptotically close to that of $H_n(X_1, \ldots, X_n)$. However, when we have censored data, such as right censored data, doubly

censored data and interval censored data, the complete sample $X_1, \ldots, X_n$ and F_n are not available.

For a censored sample, the idea of the non-parametric bootstrap (Efron, 1994) is to bootstrap the observed censored data. This resampling method still produces an incomplete sample. Hence, the above statistic H_n formulated for a complete sample is no longer applicable, and the extension of H_n for the particular censoring mechanism under consideration is needed. In some situations, such as the aforementioned goodness of fit test with interval censored data (1), the extension may not be at all obvious.

Extending the idea of Efron's bootstrap principle, we see that one way to look at the problem is that, instead of seeking an extension of H_n for incomplete data, one may try to obtain a complete i.i.d. bootstrap sample, which is asymptotically close to a sample directly drawn from F, so that the statistic H_n for the complete sample may be used for this bootstrap sample. A natural thing to do is to replace the empirical d.f. F_n in the usual bootstrap procedure by an estimator $\hat{F}_n$ of F based on incomplete data. Then, inferences may be done based on the pseudo complete i.i.d. sample $X_{n1}^*, \ldots, X_{nm}^*$ drawn from $\hat{F}_n$. As shown in section 3, one may expect that under some suitable conditions, $H_m(X_{n1}^*, \ldots, X_{nm}^*)$ has approximately the same distribution as $H_n(X_1, \ldots, X_n)$. Next, we describe the leveraged bootstrap (LB), while Fig. 2 shows how the bootstrap, the non-parametric bootstrap and the leveraged bootstrap are related to one another.

Leveraged bootstrap

(LB1) Compute the NPMLE $\hat{F}_n$ using observed incomplete data $\{\mathbf{O}_i, 1 \le i \le n\}$.

(LB2) For an integer m satisfying $m \to \infty$, as $n \to \infty$, obtain an i.i.d. leveraged bootstrap sample $X_{n1}^*, \ldots, X_{nm}^*$, which is drawn from $\hat{F}_n$.

(LB3) For the statistic of interest $H_n(X_1, \ldots, X_n)$ formulated for complete i.i.d. sample, compute $H_m^* = H_m(X_{n1}^*, \ldots, X_{nm}^*)$ and draw inference.

One may note that the above leveraged bootstrap method can be applied to different types of censored data, including interval censored data (1), and is computationally efficient, but the part of "draw inference" in (LB3) often requires some care depending on the situations. One example on this is studied in Ren (2001) in the context of the empirical likelihood inference, while the goodness of fit tests with interval censored data discussed in sections 3 and 4 demonstrate additional examples of the application of the leveraged bootstrap in practice.

3. Simple LB-goodness of fit test

Consider testing the simple goodness of fit (2) with interval censored data (1). From the test statistic T_n in (4) formulated for complete data, the statistic based on the leveraged bootstrap in (LB3) is given by

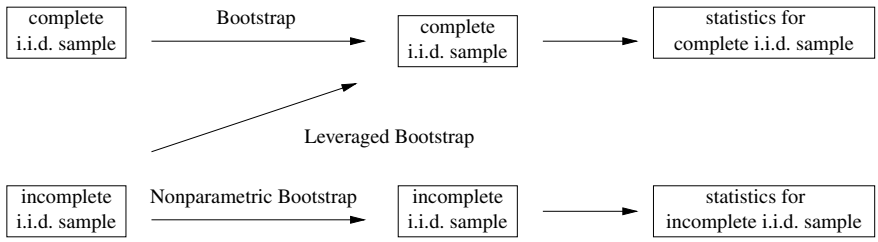

Fig. 2. The relation among bootstrap, non-parametric bootstrap and leveraged bootstrap.

$$T_m^* = m \int_0^\infty (F_{nm}^*(x) - F_0(x))^2 \, dF_0(x), \tag{7}$$

where F_{nm}^* is the empirical d.f. of a leveraged bootstrap sample $X_{n1}^*, \ldots, X_{nm}^*$ from $\hat{F}_n$. We note that the conditional covariance function of $\mathscr{F}_{nm}^* = \sqrt{m}(F_{nm}^* - \hat{F}_n)$ is given by

$$E\{\mathscr{F}_{nm}^*(x)\mathscr{F}_{nm}^*(y)|\mathbf{O}^n\} = \hat{F}_n(x) \wedge \hat{F}_n(y) - \hat{F}_n(x)\hat{F}_n(y), \tag{8}$$

where $\mathbf{O}^n = \{\mathbf{O}_i | 1 \le i \le n\}$. Since $\hat{F}_n$ is a consistent estimator of F (Groeneboom & Wellner, 1992), we know that the limit of (8) is $\Gamma(x,y) = F(x) \wedge F(y) - F(x)F(y)$. Noting that $\Gamma(x,y)$ is the covariance function of $\sqrt{n}(F_n - F)$, with proofs deferred to the appendix the following theorem shows that T_m^* in (7) has the same null limiting distribution as T_n in (4). This is also supported by the simulation results presented in section 5.

Theorem 1

Assume

$$n^\gamma(\hat{F}_n - F) = O_p(1), \quad \textit{for some } \gamma > 0, \tag{AS1}$$

and let $m = o(n^{2\gamma})$ *satisfy* $m \to \infty$, *as* $n \to \infty$. *Then, under* H_0

$$\lim_{n\to\infty} \sup_{0<x<\infty} |P\{T_m^* \le x|\mathbf{O}^n\} - P\{W \le x\}| = 0, \tag{9}$$

in probability, where W *is as in* (5).

One may note that the percentiles of W may be used directly as the critical value when T_m^* in (7) is used as the test statistic for testing (2). Specifically, at $\alpha 100\%$ significance level we

$$\text{Reject } H_0 \quad \text{if} \quad T_m^* \ge C_\alpha, \tag{10}$$

where $P\{W \ge C_\alpha\} = \alpha$. This means that only one leveraged bootstrap sample is used for the decision, thus (10) is called LB1-test.

It is easy to see that if in (LB2) one repeatedly obtains N leveraged bootstrap samples $X_{kn,1}^*, \ldots, X_{kn,m}^*$ from $\hat{F}_n, k = 1, \ldots, N$, and computes T_{km}^* for each of these samples, then frequent occurrence of $T_{km}^* \ge C_\alpha$ should lead to the rejection of H_0. This idea gives another test as follows.

Let

$$\bar{W} = N^{-1} \sum_{k=1}^{N} I\{T_{km}^* \ge C_\alpha\} \quad \text{and} \quad p_n = P_n\{T_{km}^* \ge C_\alpha\}, \tag{11}$$

where "P_n" denotes the conditional probability given $\hat{F}_n$. Then, we know that $N\bar{W}$ has binomial distribution with parameters p_n and N, and is asymptotically normal for large N. If we denote z_α as the $(1-\alpha)100$th percentile of the standard normal distribution $N(0,1)$, and for some ρ such that $0 < \rho < \alpha < 1$ we choose

$$N = \max\left\{1, \frac{p_n(1-p_n)}{[(\alpha - p_n)/(z_{\alpha-\rho} - z_\alpha)]^2}\right\}, \tag{12}$$

then for testing (2), at $\alpha 100\%$ significance level we

$$\text{Reject } H_0 \quad \text{if} \quad \bar{W} \ge \alpha + z_{\alpha-\rho}\sqrt{\frac{\alpha(1-\alpha)}{N}}. \tag{13}$$

This is called LB-goodness of fit test (LB-GOF test), and it is asymptotically consistent under some conditions on m in (7). The following theorem establishes its consistency with proofs

deferred to the appendix, where the derivation of N in (12) is apparent. The choice of m in practice is discussed at the end of this section, and some simulation results are presented in section 5.

Theorem 2
Under the assumptions of theorem 1, LB-GOF test (13) *satisfies:*

(i) under H_0,

$$\lim_{n\to\infty} P\left\{ \bar{W} \geq \alpha + z_{\alpha-\rho}\sqrt{\frac{\alpha(1-\alpha)}{N}} \,\middle|\, H_0 \right\} \leq \alpha; \tag{14}$$

(ii) under fixed alternative H_1: $F = F_1 \neq F_0$,

$$\lim_{n\to\infty} P\left\{ \bar{W} \geq \alpha + z_{\alpha-\rho}\sqrt{\frac{\alpha(1-\alpha)}{N}} \,\middle|\, H_1 \right\} = 1. \tag{15}$$

Remark 1. Theorems 1 and 2 can be applied to different types of censored data, including right censored data, doubly censored data and interval censored data. In fact, under some conditions, assumption (AS1) holds with $\gamma = 1/2$ for right censored data (Gill, 1983) and doubly censored data (Gu & Zhang, 1993), respectively. The situation for interval censored data is more complicated, but we know $\gamma = 1/3$ in (AS1) for interval censoring case 1 (Groeneboom & Wellner, 1992). As pointed out by Geskus & Groeneboom (1999), the convergence rate of $\hat{F}_n$ for interval censoring case 2 and for those cases studied by Wellner (1995), Schick & Yu (2000), Wellner & Zhang (2000) mentioned in section 1 should not be worse than case 1, because one has more information on the location of X_is. Thus, we use $\gamma = 1/3$ for interval censoring case 2 in section 5's simulation studies, which appear to be quite satisfactory.

Choice of m. From the proof of theorem 1, we see that the conditions on m are required to kill the bootstrap bias. In the appendix, this leads to the choice of m under H_0 as:

$$\hat{\boldsymbol{m}} = \min\left\{ \frac{e}{r_n}, \left(\frac{e}{\sqrt{\sigma_{mn}^2 z_{\eta/2}^2 + e r_n n^{2\gamma}} + \sigma_{mn} z_{\eta/2}} \right)^2 n^{2\gamma} \right\}, \tag{16}$$

where for $0 < \alpha, \epsilon, \eta < 1$ and C_α given in (10),

$$e = C_\alpha - C_{\alpha+\epsilon}, \quad r_n = \int_0^\infty (\hat{F}_n - F_0)^2 \, dF_0, \tag{17}$$

and σ_{nm} is the standard deviation (s.d.) of $\int_0^\infty \sqrt{m}(F_{nm}^* - \hat{F}_n) n^\gamma (\hat{F}_n - F_0)\, dF_0$. Hence, in practice one may choose m as:

$$m = \max\{n^\gamma, \hat{\boldsymbol{m}}\}. \tag{18}$$

For $\alpha = 0.05$ and interval censored data (1), based on our simulation studies we recommend the use of $\eta = 0.10, \epsilon = 0.02, \gamma = \frac{1}{3}$ in (16), and estimating σ_{mn} by LB samples (say, 30 samples) with $m = n^\gamma$.

Such a choice of m in (18) is associated with the LB-GOF test (13) as follows. Under H_0, we have $\hat{\boldsymbol{m}} = O_p(n^{2\gamma})$, which is asymptotically larger than n^γ, thus $m = \hat{\boldsymbol{m}}$ in (18). When H_0 is not true, we have $\hat{\boldsymbol{m}} = O_p(1)$, thus $m = n^\gamma$ in (18), which leads to a rejection decision. This selection method of m is used in simulation studies of section 5, and generally performs well.

4. Parametric family LB-goodness of fit test

Consider testing the parametric family goodness of fit (3) with interval censored data (1), and let $\hat{\theta}_n$ be an estimator for θ based on the NPMLE $\hat{F}_n$ (for instance, if θ is the mean of F, $\hat{\theta}_n$ is the mean of $\hat{F}_n$). A natural extension of statistic T_m^* in (7) is given by

$$\tilde{T}_m^* = m \int_0^\infty (F_{nm}^*(x) - F_0(x; \hat{\theta}_n))^2 \, dF_0(x; \hat{\theta}_n), \tag{19}$$

where F_{nm}^* is the empirical d.f. of a leveraged bootstrap sample $X_{n1}^*, \ldots, X_{nm}^*$ drawn from $\hat{F}_n$. With proofs deferred to the appendix, the following theorem shows that theorem 1 holds for this statistic $\tilde{T}_m^*$, while some related simulation results are presented in section 5.

Theorem 3

Let $F_0(x; \theta)$ have a density function $f_0(x; \theta)$, and assume

$$n^\gamma \|F_0(\cdot; \hat{\theta}_n) - F_0(\cdot; \theta)\| = O_p(1) \quad \textit{and} \quad \|f_0(\cdot; \hat{\theta}_n) - f_0(\cdot; \theta)\| = o_p(1). \tag{AS2}$$

Then, under the assumptions of theorem 1, we have that under H_0

$$\lim_{n \to \infty} \sup_{0 < x < \infty} |P\{\tilde{T}_m^* \le x | \mathbf{O}^n\} - P\{W \le x\}| = 0, \tag{20}$$

in probability, where W is as in (5).

Remark 2. In the consistency assumption (AS2), since the estimator $\hat{\theta}_n$ for θ is based on the NPMLE $\hat{F}_n$, we usually could expect the convergence rate of $\hat{\theta}_n$ to be at least the same as $\hat{F}_n$. In fact, as mentioned in section 1, Geskus & Groeneboom (1999) showed that if $\hat{\theta}_n$ is a smooth functional of $\hat{F}_n$, it is asymptotically normal with a better convergence rate than $\hat{F}_n$. In turn, (AS2) holds from the boundedness of f_0 and the uniform continuity of $f_0(\cdot; \theta)$ in θ. On the other hand, one should note that under the assumption of parametric family in (3), we cannot use the usual parametric MLE in the place of $\hat{\theta}_n$ in (AS2) or in (19), because for interval censored data (1), the observed random vector (Y_i, Z_i) does not have any parametric assumption under H_0 in (3). Of course, as the Associate Editor pointed out, a partial likelihood MLE for θ based on $F_0(\cdot; \theta)$ and observed data (Y_i, Z_i, δ_i) could be used in the place of $\hat{\theta}_n$. But the computation of such a partial likelihood-based MLE may be complicated if $F_0(\cdot; \theta)$ has a complicated form.

Based on theorem 3, it is easy to see that to test (3), the LB-GOF test (13) for testing (2) can be extended to the following parametric family LB-goodness of fit test (PF-LB-GOF test): at $\alpha 100\%$ significance level we

$$\text{Reject } H_0 \quad \text{if} \quad \bar{\tilde{W}} \ge \alpha + z_{\alpha - \rho} \sqrt{\frac{\alpha(1-\alpha)}{\tilde{N}}}, \tag{21}$$

where $\bar{\tilde{W}}$ and $\tilde{p}_n$ are obtained by replacing T_m^* with $\tilde{T}_m^*$ in (11), and $\tilde{N}$ is obtained by replacing p_n with $\tilde{p}_n$ in (12). From the proof of theorem 2, it is also easy to see that under the assumptions of theorem 3, theorem 2 holds for the PF-LB-GOF test (21). Thus, (21) is an asymptotically consistent test.

Moreover, from the discussion in section 3, in practice the choice of m for the PF-LB-GOF test (21) may be given by

$$m = \max\{n^\gamma, \hat{\tilde{m}}\}, \tag{22}$$

where $\hat{\tilde{m}}$ is obtained from replacing F_0 by $F_0(\cdot; \hat{\theta}_n)$ in (16)–(17).

5. Simulation studies

This section presents some simulation results on theorems 1–3. In these studies, the choices of m given by (18) and (22) are used for statistics T_m^* and $\tilde{T}_m^*$, respectively, with $\alpha = 0.05$ and $\gamma = 1/3$ for interval censored data (1). Since the function of parameters ϵ in (17) and η in (16) is to reduce the bootstrap bias, shown in (32)–(33) of the appendix, the choice of $\epsilon = 0.02$ and $\eta = 0.10$ seems reasonable enough for this purpose, and our simulation results confirm this. Thus, based on (17) and Shorack & Wellner (1986, p. 147), we have $e = 0.05616$ for $\alpha = 0.05$ and $\epsilon = 0.02$. In (16), the s.d. σ_{nm} or $\tilde{\sigma}_{nm}$ with $F_0(\cdot; \hat{\theta}_n)$ replacing F_0 is estimated based on 30 LB samples with $m = n^\gamma$. Moreover, with an arbitrarily chosen $\rho = \alpha/2$ in (12), the choices of N for LB-GOF test (13) and $\tilde{N}$ for PF-LB-GOF test (21) are computed according to p_n and $\tilde{p}_n$, respectively, which are estimated by 10 000 T_m^*s and $\tilde{T}_m^*$s, respectively.

One may note that the NPMLE $\hat{F}_n$ is not always a proper d.f. for censored data (Mykland & Ren, 1996). In our studies, we always adjust $\hat{F}_n$ to be a proper d.f. by setting $\hat{F}_n = 1$ at the largest observation in the data, and our experiences show that due to this, the simulation results appear to be less biased. This kind of adjustment of the Kaplan–Meier estimator has been adopted by some researchers in literature (Efron, 1967; Miller, 1976).

Simulation on theorems 1 and 3. Table 1 displays the simulation results which compare the percentiles of W in (5) (Shorack & Wellner, 1986, p. 147) with those of T_m^* in (7) for interval censored data case 2. Here, the simulation percentiles of T_m^* are based on 5000 samples, and the absolute value of the differences between the simulation percentiles of T_m^* and the percentiles of W are displayed in the "Errors" column. The same simulation studies are repeated for $\tilde{T}_m^*$ in (19) with interval censored data, and the results are displayed in Table 2.

The results in Tables 1 and 2 show that the leveraged bootstrap performs well, which is consistent with theorems 1 and 3. One may note that $\tilde{T}_m^*$ in (19) for testing the parametric

Table 1. *Percentiles of T_m^* with interval censored data of sample size $n = 200$*

		$X \sim F_0 = \exp(1)$, $Y \sim \exp(3)$, $Z = (2/3)Y - 2.5$		$X \sim F_0 = N(0,1)$, $Y \sim N(1,4)$, $Z = (2/3)Y - 2.5$	
Percentile	W	T_m^*	Errors	T_m^*	Errors
5th	0.03656	0.049818	0.013258	0.049516	0.012956
10th	0.04601	0.060394	0.014384	0.060410	0.014400
15th	0.05426	0.070873	0.016613	0.069615	0.015355
20th	0.06222	0.081346	0.019126	0.079429	0.017209
25th	0.07025	0.090635	0.020385	0.090035	0.019785
30th	0.07860	0.099284	0.020684	0.099979	0.021379
35th	0.08744	0.109160	0.021720	0.109021	0.021581
40th	0.09696	0.119359	0.022399	0.119893	0.022933
45th	0.10736	0.129680	0.022320	0.131418	0.024058
50th	0.11888	0.141144	0.022263	0.144100	0.025220
55th	0.13183	0.155332	0.023502	0.158298	0.026468
60th	0.14663	0.170091	0.023461	0.176190	0.029560
65th	0.16385	0.187482	0.023632	0.197282	0.033432
70th	0.18433	0.209727	0.025397	0.220952	0.036622
75th	0.20939	0.238128	0.028738	0.246602	0.037212
80th	0.24124	0.268220	0.026980	0.276236	0.034996
85th	0.28406	0.307563	0.023503	0.318523	0.034463
90th	0.34730	0.365021	0.017721	0.384534	0.036347
95th	0.46136	0.463432	0.002072	0.491160	0.029800
99th	0.74346	0.716075	0.027385	0.744601	0.001141
		Average of selected $m = 5.1$		Average of selected $m = 5.2$	

Table 2. *Percentiles of $\tilde{T}_m^*$ with interval censored data of sample size $n = 200$*

		$F_0 = \exp(\mu)$, $X \sim \exp(1)$ $Y \sim \exp(3), Z = (2/3)Y - 2.5$		$F_0 = N(\mu, \sigma^2)$, $X \sim N(0,1)$ $Y \sim N(1,4), Z = (2/3)Y - 2.5$	
Percentile	W	$\tilde{T}_m^*$	Errors	$\tilde{T}_m^*$	Errors
5th	0.03656	0.051116	0.014556	0.050676	0.014116
10th	0.04601	0.061572	0.015562	0.061688	0.015678
15th	0.05426	0.071821	0.017561	0.071369	0.017109
20th	0.06222	0.080820	0.018608	0.081972	0.019752
25th	0.07025	0.090404	0.020154	0.093068	0.022818
30th	0.07860	0.099193	0.020593	0.102985	0.024385
35th	0.08744	0.108551	0.021111	0.113126	0.025686
40th	0.09696	0.119536	0.022576	0.124431	0.027471
45th	0.10736	0.130704	0.023344	0.135860	0.028500
50th	0.11888	0.141789	0.022909	0.147220	0.028340
55th	0.13183	0.155115	0.023285	0.161682	0.029852
60th	0.14663	0.171021	0.024391	0.178973	0.032343
65th	0.16385	0.188747	0.024898	0.199824	0.035974
70th	0.18433	0.208789	0.024459	0.220568	0.036238
75th	0.20939	0.232366	0.022976	0.243649	0.034259
80th	0.24124	0.266310	0.025070	0.275699	0.034459
85th	0.28406	0.308877	0.024817	0.320523	0.036463
90th	0.34730	0.373341	0.026041	0.378826	0.031526
95th	0.46136	0.471307	0.009947	0.488538	0.027178
99th	0.74346	0.720468	0.022992	0.719518	0.023942
		Average of selected $m = 5.2$		Average of selected $m = 5.5$	

family GOF (3) and T_m^* in (7) for testing the simple GOF (2) have quite similar performance. Also, as expected, our studies show that the approximation for interval censored data gets better as the sample size n increases. However, we only present the results with $n = 200$ for interval censored data, because it is extremely time-consuming to conduct the simulation study for large samples.

Simulation on theorem 2. The simulation results on theorem 2 for testing the simple GOF (2) are displayed in Figs 3–5, which compare the power curves of the LB-GOF test (13) with those of LB1-test (10). All power curves are the smoothed versions based on 500 simulation runs for sample size $n = 200$, and 300 simulation runs for $n = 500$.

From Figs 3–5, one may notice that in the neighbourhood of H_0, the power of LB-GOF test is generally better, and that as expected, the power increases faster with larger sample size n as F moves away from F_0.

Remark 3. Although not presented, the power curves of PF-LB-GOF test (21) have similar performance to Figs 3–5 for the simple LB-GOF test (13).

6. An example

In De Gruttola & Lagakos (1989), an interval censored data set on

$$X = \text{time of HIV infection} \tag{23}$$

from AIDS research was presented. A brief description of this data set is given below.

Since 1978, 262 people with Type A and B haemophilia have been treated at Hôpital Kremlin Bicêtre and Hôpital Cœur des Yvelines in France. For each individual, the only

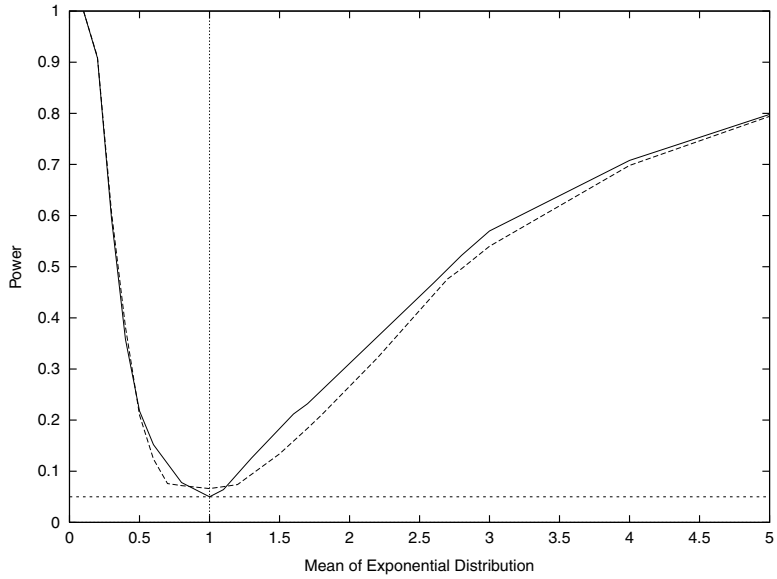

Fig. 3. Power curves of LB-GOF test (solid line) and LB1 test (dashed line) for H_0: $F = F_0 = \exp(1)$, with interval censored data: $X \sim \exp(\mu)$, $Y \sim \exp(3)$, $Z = [(2/3)Y - 2.5]$, $n = 200$; Average N under H_0 is 512.0.

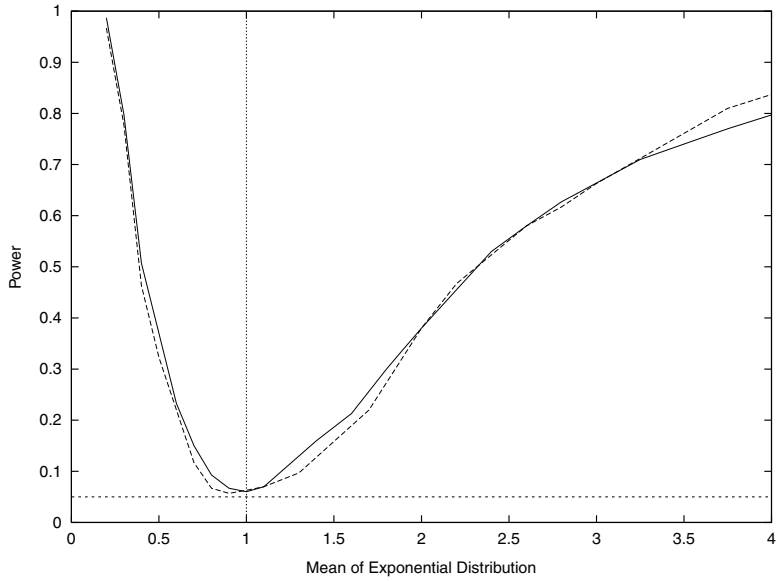

Fig. 4. Power curves of LB-GOF test (solid line) and LB1 test (dashed line) for H_0: $F = F_0 = \exp(1)$, with interval censored data: $X \sim \exp(\mu)$, $Y \sim \exp(3)$, $Z = [(2/3)Y - 2.5]$, $n = 500$; Average N under H_0 is 874.0.

information available on X is $X \in [X_L, X_R]$, while it is assigned $X_L = 1$ if the individual was found to be infected with HIV on his/her first test for infection. Along with the retrospective tests for evidence of HIV infection, observations X_L and X_R were determined by the time at which the blood samples were stored. In this data set, time is measured in 6-months intervals, with $X = 1$ denoting 1 July 1978, and one of the interests of the study is the distribution of X.

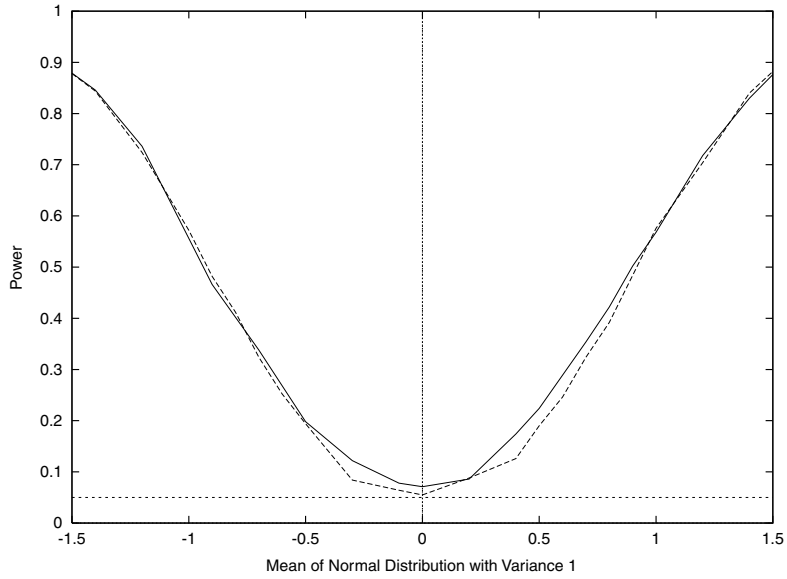

Fig. 5. Power curves of LB-GOF test (solid line) and LB1 test (dashed line) for H_0: $F = F_0 = N(0,1)$, with interval censored data: $X \sim N(\mu, 1)$, $Y \sim N(1,4), Z = [(2/3)Y - 2.5], n = 500$; Average N under H_0 is 505.9.

To demonstrate the application of the parametric family LB-goodness of fit test developed in section 4, we consider the updated version of this data set for 104 individuals in the heavily treated group, i.e. patients who received at least 1000 μg/kg of blood factor for at least one year between 1982 and 1985. The data set is given in Kim *et al.* (1993), and is included in Table 3 for convenience.

Note that the data set in Table 3 always satisfies $X_L < X_R$, and it is associated with interval censored data (1) in the following way:

$$\begin{aligned} 1 < X_L < X_R < \infty \quad &\Leftrightarrow \quad \delta = 1,\ Z = X_L,\ \ Y = X_R \\ 1 < X_L < X_R = \infty \quad &\Leftrightarrow \quad \delta = 2,\ Z = -\infty,\ \ Y = X_L \\ 1 = X_L < X_R < \infty \quad &\Leftrightarrow \quad \delta = 3,\ Z = X_R,\ \ Y = \infty. \end{aligned} \tag{24}$$

Table 3. HIV *observations of* 104 *patients in the heavily treated group*

X_L	X_R	X_L	X_R	X_L	X_R	X_L	X_R	X_L	X_R	X_L	X_R	X_L	X_R	X_L	X_R
15	∞	10	11	14	15	10	11	7	15	12	13	9	11	1	12
15	∞	10	11	14	15	10	12	9	10	12	13	9	12	1	14
17	∞	10	11	14	15	13	14	9	10	12	13	10	12	1	14
17	∞	10	11	14	15	13	14	10	11	12	14	10	12	1	14
1	7	10	15	14	15	16	∞	10	11	13	16	13	15	5	7
1	13	11	13	15	16	16	∞	10	11	14	16	13	15	7	9
1	15	11	13	1	10	17	∞	11	12	14	16	14	15	8	10
1	15	11	13	1	15	1	11	12	13	1	7	16	∞	9	12
3	14	12	14	5	8	1	13	12	13	3	7	1	16	9	12
7	10	13	15	9	13	5	7	12	13	7	9	1	12	9	12
8	15	14	15	9	13	3	15	10	12	15	16	10	12	15	16
11	13	1	7	12	13	5	7	13	15	8	10	13	15	9	12
14	15	10	11	14	15	10	14	15	16	12	13	15	16	13	14

Although the data shown in Table 3 are integers due to the fact that De Gruttola & Lagakos (1989) discretized the time axis into 6-month intervals in their studies, it is obvious that X_L and X_R in (24) are continuous random variables. Moreover, due to the way in which X_L and X_R were determined, we may assume that $[X_L, X_R]$ is independent of X, because the available blood samples were stored purely from haemophilia treatment which had nothing to do with HIV infection. Thus, here we have a data set which is interval censoring case 2 as in (1).

Applying the same procedure for the simulation studies in section 5 to this data set (24), we conduct the PF-LB-GOF tests (21) to test (3) and summarize the results in Table 4.

From Table 4, we conclude that at 5% significance level, there is not sufficient evidence to reject that X in (23) has a normal distribution. Moreover, based on data (24), Fig. 6 compares the curves of the NPMLE $\hat{F}_n$, $N(\hat{\mu}_n, \hat{\sigma}_n^2)$, and $\exp(\hat{\mu}_n)$, where $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are the mean and variance of $\hat{F}_n$, respectively. Evidently, the results of Fig. 6 are consistent with the test results shown in Table 4.

One may note that although Fig. 6 indicates that $\hat{F}_n$ is above $N(\hat{\mu}_n, \hat{\sigma}_n^2)$, our test result does not conclude this. This is because our proposed testing procedure is constructed for the alternative hypothesis H_1: $F \notin \mathscr{F}_0$ in (3), which in the current context means that F is not normal. To test whether F is above $N(\hat{\mu}_n, \hat{\sigma}_n^2)$, a different testing procedure needs to be developed for the corresponding alternative hypothesis, but this is not considered here.

Table 4. *5% parametric family goodness of fit tests with interval censored HIV data*

H_0	$\tilde{\tilde{W}}$	$\alpha + z_{\alpha-\rho}[\alpha(1-\alpha)/\tilde{N}]^{1/2}$	Selected m	Selected $\tilde{N}$
$F = \exp(\mu)$	1.0000	0.477172	6	1
$F = N(\mu, \sigma^2)$	0.0454	0.114399	7	44

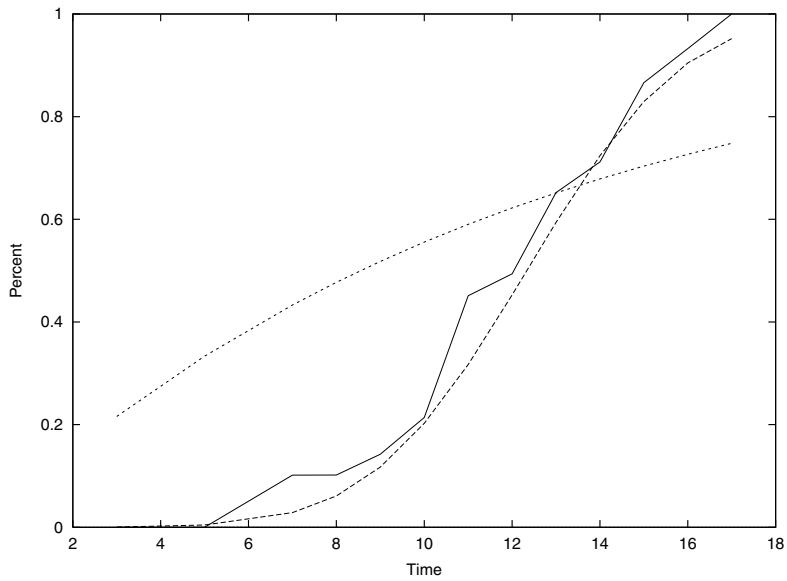

Fig. 6. Distribution curves of $\hat{F}_n$ (solid line), $N(\hat{\mu}_n, \hat{\sigma}_n^2)$ (dashed line) and $\exp(\hat{\mu})$ (dotted line) for interval censored HIV data.

7. Conclusions

Cramér–von Mises type goodness of fit tests for interval censored data are proposed based on a resampling method called the leveraged bootstrap (LB), and their asymptotic consistency is shown mathematically with support from simulation results. The essential difference between the LB tests and the usual testing procedures is that the test statistics of the LB tests are obtained through resampling, in the process of which the leveraged bootstrap transfers censored data through statistic T_n in (4) into some useful information to draw inference. Although the proposed tests can in fact be applied to other types of censored data, including right censored data and doubly censored data, they are mainly meant to fill a blank in the literature for interval censored data. Simulation studies show that the proposed methods are computationally very efficient, because the EM algorithm is used only once in the procedure to compute the NPMLE $\hat{F}_n$. Based on Figs 3–5, one may prefer the LB-GOF test (or PF-LB-GOF test) over the LB1-test in practice. Finally, it should be noted that a better choice of m might be possible in the proposed procedure to improve the power of the tests.

Acknowledgements

This research was partially supported by NSF grants DMS-9510367 and DMS-9626532/DMS-9796229. The author thanks Brad Efron, Peter Bickel and Art Owen for useful discussions while this manuscript was being prepared. Also, the author is grateful for many valuable comments from two referees and the Associate Editor, which helped improve the results of this paper. In particular, one referee's suggestion pushed the author to study the parametric family goodness of fit test and establish the current theorem 3, while many detailed comments from the Associate Editor were particularly helpful for the presentation of this manuscript.

References

Bickel, P. J. & Ren, J. (1996). The m out of n bootstrap and goodness of fit tests with doubly censored data. *Robust statistics, data analysis and computer intensive methods*, Lecture Notes in Statistics **109**, 35–47. Springer-Verlag, New York.

Bickel, P. J. & Ren, J. (2001). The bootstrap in hypothesis testing. *State of the art in statistics and probability theory, festschrift for Willem R. van Zwet*, Lecture Notes in Mathematical Statistics, IMS, **36**, 91–112. Institute of Mathematical statistics, Haywood, CA.

Bickel, P. J., Göetze, F. & van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses and remedies for losses. *Statist. Sinica* **7**, 1–31.

De Gruttola, V. & Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics* **45**, 1–11.

Efron, B. (1967). The two sample problem with censored data. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4**, 831–853.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.

Efron, B. (1981). Censored data and the bootstrap. *J. Amer. Statist. Assoc.* **76**, 312–319.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *J. Amer. Statist. Assoc.* **89**, 463–474.

Gehan, E. A. (1965). A generalized two-sample Wilcoxon test for doubly censored data. *Biometrika* **52**, 650–653.

Geskus, R. & Groeneboom, P. (1999). Asymptotically optimal estimation of smooth functionals for interval censoring, Case 2. *Ann. Statist.* **27**, 627–674.

Gill, R. D. (1983). Large sample behavior of the product-limit estimator on the whole line. *Ann. Statist.* **11**, 49–58.

Groeneboom, P. (1996). Lectures on inverse problems. *Lectures on probability and statistics*. Lecture Notes in Mathematics **1648**, 67–164. Springer, Berlin.

Groeneboom, P. & Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*. Birkhäuser Verlag, Verlag.

Gu, M. G. & Zhang, C. H. (1993). Asymptotic properties of self-consistent estimators based on doubly censored data. *Ann. Statist.* **21**, 611–624.
Kim, M. Y., De Gruttola, V. G. & Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics* **49**, 13–22.
Miller, R. G. (1976). Least squared regression with censored data. *Biometrika* **63**, 449–464.
Mykland, P. A. & Ren, J. (1996). Self-consistent and maximum likelihood estimation for doubly censored data. *Ann. Statist.* **24**, 1740–1764.
Peer, P. G., Van Dijck, J. A., Hendriks, J. H., Holland, R. & Verbeek, A. L. (1993). Age-dependent growth rate of primary breast cancer. *Cancer* **11**, 3547–3551.
Politis, D. N. & Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22**, 2031–2050.
Politis, D. N., Romano, J. P. & Wolf, M. (1999). *Subsampling*. Springer-Verlag, New York.
Ren, J. (1995a). Generalized Cramér–von Mises tests of goodness of fit with doubly censored data. *Ann. Inst. Statist. Math.* **47**, 525–549.
Ren, J. (1995b). Self-consistent estimators, bootstrap and censored data. *IMS Bulle.* **24**, 467.
Ren, J. (2001). Weighted empirical likelihood ratio confidence intervals for the mean with censored data. *Ann. Inst. Statist. Math.* **53**, 498–516.
Ren, J. & Peer, P. G. (2000). A study on effectiveness of screening mammograms. *Internat. J. Epidemiol.* **29**, 803–806.
Schick, A. & Yu, Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scand. J. Statist.* **27**, 45–55.
Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. Wiley, New York.
Shorack, G. R. & Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley, New York.
Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Assoc.* **69**, 169–173.
Wellner, J. A. (1995). Interval censoring Case 2: alternative hypothesis. *Analysis of censored data*, 272–291. Institute of Mathematical Statistics, Hayward, CA.
Wellner, J. A. & Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* **28**, 779–814.

Received October 2000, in final form January 2002

Jian-Jian Ren, Department of Mathematics, University of Central Florida, Orlando, FL 32816-1364, USA.
E-mail: jren@mail.ucf.edu

Appendix

Proof of theorem 1. Let U be the uniform d.f. on $(0,1)$ and let $U_1,\ldots,U_m$ be a random sample from U, then $X^*_{ni}=\hat{F}_n^{-1}(U_i)=\inf\{x;\hat{F}_n(x)\geq U_i\}, 1\leq i\leq m$, is a random sample from $\hat{F}_n$. Denote F_m as the empirical d.f. of $U_1,\ldots,U_m$, and denote

$$\Psi_m(F^*_{nm})=T^*_m=m\int_0^\infty (F^*_{nm}-F_0)^2\,dF_0, \tag{25}$$

where F^*_{nm} is the same as in (7). From Serfling's lemma (1980, p. 3), we have

$$\begin{aligned}F^*_{nm}(x)&=m^{-1}\sum_{i=1}^m I\{X^*_{ni}\leq x\}=m^{-1}\sum_{i=1}^m I\{\hat{F}_n^{-1}(U_i)\leq x\}\\&=m^{-1}\sum_{i=1}^m I\{U_i\leq \hat{F}_n(x)\}=F_m(\hat{F}_n(x)).\end{aligned} \tag{26}$$

Note that assumptions (AS1) and $m=o(n^{2\gamma})$ imply that under H_0,

$$\begin{aligned}\sqrt{m}[F_m(\hat{F}_n(\cdot))-F_0]&=\sqrt{m}[F_m(\hat{F}_n(\cdot))-\hat{F}_n]+\sqrt{m}(\hat{F}_n-F_0)\\&=\sqrt{m}[F_m(\hat{F}_n(\cdot))-\hat{F}_n]+o_p(1),\end{aligned} \tag{27}$$

where $o_p(1)$ uniformly converges to 0 in probability, and note that the limiting process of $\sqrt{m}(F_m - U)$ on $[0,1]$ is the Brownian bridge B with a covariance function given by $\Lambda(s,t) = s \wedge t - st$, where $s, t \in [0,1]$. Thus,

$$\Psi_m(F^*_{nm}) = \Psi_m(F_m \circ \hat{F}_n) = m \int_0^\infty [F_m(\hat{F}_n(\cdot)) - \hat{F}_n]^2 \, dF_0 + o_p(1).$$

Hence, from (AS1), the continuity of F_0, and Shorack & Wellner (1986, pp. 145, 147), we have that under H_0, as $n \to \infty$,

$$P\{\Psi_m(F^*_{nm}) \le x | \mathbf{O}^n\} \to P\left\{ \int_0^\infty (B \circ F_0)^2 \, dF_0 \le x \right\} = P\left\{ \int_0^1 B^2 \, dU \le x \right\} = P\{W \le x\},$$

in probability. Therefore, (9) follows from the continuity of the d.f. of W.

Proof of theorem 2. From theorem 1, we know that under H_0, $p_n \xrightarrow{P} \alpha$, as $n \to \infty$, and it is easy to see that (AS1) implies that under H_1, $T^*_m \xrightarrow{P} \infty$ and $p_n \xrightarrow{P} 1$, as $n \to \infty$. Thus, from the choice of N given in (12), we know that as $n \to \infty$,

$$N \xrightarrow{P} \infty \text{ under } H_0 \quad \text{and} \quad N \xrightarrow{P} 1 \text{ under } H_1. \tag{28}$$

(i) Under H_0, we know that (12) and (28) give

$$\begin{aligned}
&\lim_{n\to\infty} P\left\{ \bar{W} \ge \alpha + z_{\alpha-\rho}\sqrt{\frac{\alpha(1-\alpha)}{N}} \,\middle|\, H_0 \right\} \\
&\quad = \lim_{n\to\infty} P\left\{ \frac{\sqrt{N}(\bar{W} - p_n)}{\sqrt{p_n(1-p_n)}} \ge \frac{\sqrt{N}(\alpha - p_n)}{\sqrt{p_n(1-p_n)}} + z_{\alpha-\rho}\sqrt{\frac{\alpha(1-\alpha)}{p_n(1-p_n)}} \,\middle|\, H_0 \right\} \\
&\quad \le \lim_{n\to\infty} P\left\{ \frac{\sqrt{N}(\bar{W} - p_n)}{\sqrt{p_n(1-p_n)}} \ge -(z_{\alpha-\rho} - z_\alpha) + z_{\alpha-\rho}\sqrt{\frac{\alpha(1-\alpha)}{p_n(1-p_n)}} \,\middle|\, H_0 \right\} = P\{Z \ge z_\alpha\} = \alpha;
\end{aligned}$$

(ii) Under H_1, we know that (12) and (28) give

$$\lim_{n\to\infty} P\left\{ \bar{W} \ge \alpha + z_{\alpha-\rho}\sqrt{\frac{\alpha(1-\alpha)}{N}} \,\middle|\, H_1 \right\} = \lim_{n\to\infty} P\{T^*_m \ge \alpha + z_{\alpha-\rho}\sqrt{\alpha(1-\alpha)} | H_1\} = 1.$$

Derivation of (16). If we denote

$$R^*_m = 2m \int_0^\infty (F^*_{nm} - \hat{F}_n)(\hat{F}_n - F_0) \, dF_0 + mr_n, \tag{29}$$

then (7) is expressed by

$$T^*_m = S^*_m + R^*_m, \quad \text{for} \quad S^*_m = m \int_0^\infty (F^*_{nm} - \hat{F}_n)^2 \, dF_0. \tag{30}$$

From the proof of theorem 1 above, we know that under H_0, as $n \to \infty$

$$S^*_m \overset{D}{\approx} W \quad \text{and} \quad R^*_m \overset{D}{\approx} N(mr_n, 4mn^{-2\gamma}\sigma^2_{mn}). \tag{31}$$

Thus the choice of m should make R^*_m negligible. Note that if $|R^*_m| \le e$, we have

$$\begin{aligned}
|P_n\{T^*_m \le x\} - P_n\{S^*_m \le x\}| &\le \max\{P_n\{x < S^*_m \le x + e\}, P_n\{x - e < S^*_m \le x\}\} \\
&\overset{H_0}{\approx} \max\{P\{x < W \le x + e\}, P\{x - e < W \le x\}\} = G_0(x + e) - G_0(x) \le \epsilon,
\end{aligned} \tag{32}$$

where $x \le C_\alpha$ and G_0 denotes the d.f. of W. Since R^*_m is a random variable for a given $\hat{F}_n$, (32) implies that we may choose m such that

$$P_n\{|R_m^*| \le e\} \ge (1-\eta), \quad \text{where } 0 < \eta < 1 \tag{33}$$

for some small η. Thus, (16) can be easily derived from (29), (33), $m \le e/r_n$ and

$$P_n\{|R_m^*| \le e\} \approx P_n\left\{\left|\frac{2\sqrt{m}}{n^{\gamma}}\sigma_{nm}Z + mr_n\right| \le e\right\} \ge P_n\left\{\left|\frac{2\sqrt{m}}{n^{\gamma}}\sigma_{nm}Z\right| \le (e - mr_n)\right\},$$

where Z stands for a standard normal random variable.

Proof of theorem 3. Since assumptions (AS1), (AS2) and $m = 0(n^{2\gamma})$ imply that under H_0,

$$\begin{aligned}\sqrt{m}(F_m \circ \hat{F}_n - F_0(\cdot;\hat{\theta}_n)) &= \sqrt{m}(F_m \circ \hat{F}_n - \hat{F}_n) + \sqrt{m}(\hat{F}_n - F_0(\cdot;\hat{\theta}_n)) \\ &= \sqrt{m}(F_m \circ \hat{F}_n - \hat{F}_n) + \sqrt{m}(\hat{F}_n - F_0(\cdot;\theta)) + \sqrt{m}(F_0(\cdot;\theta) - F_0(\cdot;\hat{\theta}_n)) \\ &= \sqrt{m}(F_m \circ \hat{F}_n - \hat{F}_n) + o_p(1), \quad \text{as } n \to \infty,\end{aligned}$$

thus for $y = F_0^{-1}(t;\theta)$, the proof follows from (AS2), the proof of theorem 1 and

$$\begin{aligned}\tilde{T}_m^* &= m\int_0^{\infty} (F_m \circ \hat{F}_n - \hat{F}_n)^2 \, dF_0(\cdot;\hat{\theta}_n) + o_p(1) \\ &= m\int_0^{\infty} (F_m \circ \hat{F}_n - \hat{F}_n)^2 \, dF_0(\cdot;\theta) \\ &\quad + m\int_0^1 [F_m \circ \hat{F}_n(y) - \hat{F}_n(y)]^2 [f_0(y;\hat{\theta}_n) - f_0(y;\theta)]\, dt + o_p(1) \\ &= m\int_0^{\infty} (F_m \circ \hat{F}_n - \hat{F}_n)^2 \, dF_0(\cdot;\theta) + o_p(1).\end{aligned}$$