Electronic Journal of Statistics
Vol. 7 (2013) 2550–2576
ISSN: 1935-7524
DOI: 10.1214/13-EJS853

A goodness of fit test for the survival function under random right censoring

Dimitrios Bagkavos

Accenture, 17 Rostoviou Str, 11526, Athens, Greece
e-mail: dimitrios.bagkavos@accenture.com

Dimitrios Ioannides

University of Macedonia, 156 Egnatia Str, 54006, Thessaloniki, Greece

and

Aglaia Kalamatianou

Panteion University, 136 Syggrou Ave, 176 75, Athens, Greece

Abstract: The present article contributes a goodness of fit test for the survival function under random right censoring. The test is based on a central limit theorem for the Integrated Square Error of an already existing in the literature kernel survival function estimate. Establishment of its asymptotic distribution yields the proposed test statistic for drawing decision on the null hypothesis of correctness of the assumed survival function. Numerical simulations quantify the empirical nominal level and power of the suggested test for various sample sizes and amounts of censoring and facilitate comparison with the power of the data driven Neyman goodness of fit test for censored samples.

AMS 2000 subject classifications: Primary 62G10; secondary 62N03.
Keywords and phrases: Survival function, goodness of fit, censoring, kernel.

Received May 2012.

1. Introduction

This article considers the goodness of fit testing of the survival function by a fully specified estimate under the random right censored data setting. The intention is to provide a statistical test applicable to a wide variety of situations for assessing the validity of a parametric model. This is achieved by employing as the test basis, a kernel survival function estimate, which by definition does not depend on distributional assumptions on the underlying data set.

Typically, formulation of a hypothesis test involves establishing the asymptotic distribution of a measure of accuracy of the smooth estimate, under the null hypothesis that the true survival function is fully specified by a given parametric model. The preferred measure here is the Integrated Square Error (ISE) because it quantifies the performance of the estimate for the available data set

at hand. The smooth estimate is taken to be the estimate of [15], given in Section 2 below. The absence in the literature of a goodness of fit test based on the estimate employed here, together with the advantages this estimate offers, discussed in detail in [2], necessitate the development of such a test and extend its scope of application.

Based on the ISE, development of the test statistic is done in lines parallel to those of [9] for the density setting which was also used by [12] for the fixed design nonparametric regression setting. The result reveals that the form of the obtained statistic is determined by the amount of smoothing applied to the data. Furthermore, the empirical power of the proposed goodness of fit test is investigated via an extensive simulation study and is compared under the same settings with the power of Neyman's test.

The methodological contributions of the present research include a central limit theorem for the ISE of the smooth estimate of [15] as well as numerical evidence on the null performance and power of the proposed test for various sample sizes and amounts of censoring.

The rest of the paper is organized as follows. The framework of study together with definition of the smooth estimate is given in Section 2. Section 3 is devoted to the central limit theorem for the ISE of the estimate and development of the suggested test. The simulation studies on the nominal level and the power of both the proposed and the Neyman test are given in Section 4. Section 5 demonstrates how to utilize the test in practice with real data. Section 6 contains a discussion of the present work with emphasis on when and how the test should be used. Technical proofs are deferred for the last Section.

2. Notation and preliminaries

Let $T_1, T_2, \ldots, T_n$ be a sample of n i.i.d. survival times censored at the right by n i.i.d. random variables $U_1, U_2, \ldots, U_n$, independent of the T_i's. Let f and F be the density and distribution function of the T_i's and H the distribution function of the U_i's. The observed data are then the pairs (X_i, Δ_i), $i = 1, 2, \ldots, n$ with $X_i = \min\{T_i,\ U_i\}$ and $\Delta_i = 1_{\{T_i \leq U_i\}}$ where $1_{\{\cdot\}}$ is the indicator random variable of the event $\{\cdot\}$. The observed data form an i.i.d. sample with probability density g and distribution function G which satisfies $1-G = (1-F)(1-H)$. An estimate of the unknown survival function $S(x) = P(T > x)$ for a continuous duration T is $\hat{S}(x) = 1 - \hat{F}(x)$ where

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i}{1 - H(X_i)} W\left(\frac{x - X_i}{h}\right),$$
$$W(x) = \int_{-\infty}^{x} K(u)\, du.$$

The real valued function K is called kernel and integrates to 1, while h is called bandwidth and controls the amount of smoothing applied to the estimate. Estimator $\hat{S}(x)$ cannot be used directly in practice as it involves the unknown

censoring distribution $H(x)$. One solution is to reverse the intuitive role played by T_i and U_i and estimate $1-H(x)$ by the (sightly modified) Kaplan–Meier, [16], estimator. The result is

$$1-\hat{H}(x)=\begin{cases}1, & 0\leq x\leq Z_1\\ \prod_{i=1}^{k-1}\left(\frac{n-i+1}{n-i+2}\right)^{1-\Lambda_i}, & Z_{k-1}<x\leq Z_k, k=2,\ldots,n\\ \prod_{i=1}^{n}\left(\frac{n-i+1}{n-i+2}\right)^{1-\Lambda_i}, & Z_n<x,\end{cases}$$

where (Z_i,Λ_i) are the ordered X_i's, along with their censoring indicators Δ_i, $i=1,\ldots,n$. This gives rise to the practically useful estimator

$$\hat{S}_n(x)=1-\frac{1}{n}\sum_{i=1}^{n}\frac{\Delta_i}{1-\hat{H}(X_i)}W\left(\frac{x-X_i}{h}\right), \tag{1}$$

which is used next as the basis of the proposed goodness of fit test.

3. Assumptions and main results

The main result of this research is presented in this section in the form of Theorem 1 below which is then applied to drawing inference about the true survival curve. Prior to that, the necessary notation and assumptions are given.

First, denote with $\mu_i(K)$ the ith moment, $i=0,1,2$ of the function K and with $R(K)$ the integral of the real function K^2 over its domain. The following conditions are assumed throughout

1. $S(x)$ is twice differentiable and $S''(x)$ is bounded and uniformly continuous.
2. For $l=0,1,2$, the lth derivative of K, $K^{(l)}$, is bounded and absolutely integrable with finite second moments.
3. $R(K)<+\infty$ and $\mu_0(K)=1,\mu_1(K)=0,\mu_2(K)<+\infty$, i.e. the kernel K is of order 2.
4. There exists small enough h such that $W((y-x)h^{-1})/(1-G(y))$ is uniformly bounded for $|y-x|>M$, for any $M>0$.

A consequence of condition 2 is that $W(x)$ is bounded, while condition 3 and particularly $\mu_0(K)=1$ implies

$$\lim_{x\to-\infty}W(x)=0,\ \text{and}\ \lim_{x\to+\infty}W(x)=1.$$

Conditions 1–3 are satisfied by virtually all kernels in use in practice, see for example [19]. Condition 4 essentially means that there should be enough censored data at the right end of the estimation region for the asymptotics to apply. It has to be noted that it is automatically satisfied when the kernel has bounded support.

A widely used measure of closeness of two curves is the ISE. In the case of estimator $\hat{S}(x)$ it is defined as

$$I_n \equiv ISE(\hat{S}, S) = \int (\hat{S}(x) - S(x))^2 \, dx = \int (\hat{F}(x) - F(x))^2 \, dx.$$

Set $k = \mu_2(K)/2$,

$$B(u) = \int W(t)W(u+t) \, dt$$

and

$$d(n) = \begin{cases} nh^{-3/2} & \text{if } nh^3 \to 0 \\ n^{-1/2}h^{-3} & \text{if } nh^3 \to +\infty \\ n^{3/2} & \text{if } nh^3 \to \lambda \neq 0. \end{cases}$$

Also let Z denote an asymptotic standard normal random variable. Then, the limiting distribution of I_n is given in the next theorem

Theorem 1. *Assume conditions 1–4. Also assume that $h \to 0$ and $nh \to +\infty$ as $n \to +\infty$. Then*

$$d(n)\,(I_n - \mathbb{E}I_n) \to \begin{cases} 2^{1/2}\sigma_1 Z & \text{if} \quad nh^3 \to 0 \\ 2k\sigma_4 Z & \text{if} \quad nh^3 \to +\infty \\ (2^{1/2}\lambda^{1/2}\sigma_1 + 2\lambda^{1/3}k\sigma_4)Z & \text{if} \quad nh^3 \to \lambda \end{cases}$$

with $0 < \lambda < +\infty$ and

$$\sigma_1^2 = R\left(f(z)(1-H(z))^{-1}\right) R(B(u))$$
$$\sigma_4^2 = \left\{ \int F''(x)^2 \frac{f(x)}{1-H(x)} \, dx \right\} R(B(u)) - \int F''(x)^2 F(x) \, dx.$$

Remark 1. It is evident from Theorem 1 that the limiting distribution of the centered and scaled I_n depends on the amount of smoothing applied to the data.

If $nh^3 \to 0$, then the data are undersmoothed and the stochastic part of the deviation dominates the systematic part. In this situation the stochastic behavior of the centered and scaled ISE is determined by the term

$$\int (\hat{F}(x) - \mathbb{E}\hat{F}(x))^2 \, dx.$$

In the case of oversmoothing ($nh^3 \to +\infty$), the stochastic part of $d(n)\,(I_n - \mathbb{E}I_n)$ is asymptotically negligible compared to the systematic part. Then, the limiting distribution of the centered and scaled ISE is determined by the term

$$\int (\hat{F}(x) - \mathbb{E}\hat{F}(x))(\mathbb{E}\hat{F}(x) - F(x)) \, dx.$$

For optimal smoothing, i.e. $nh^3 \to \lambda$ neither term dominates as asymptotically they both have the same magnitude. In this case the limiting distribution centered and scaled ISE is the sum of the distributions of the two terms above.

Theorem 1 is applied next in creating a goodness of fit test. The test is given in the form of the simple null hypothesis $H_0 : S(x) = S_0(x)$ against the alternative $H_1 : S(x) \neq S_0(x)$. $S_0(x)$ denotes the assumed true parametric model and is completely known. By Corollary 1 in [2], Theorem 1 and under H_0, the proposed test statistic, T_n, is obtained by replacing $\hat{S}(x)$ by $\hat{S}_n(x)$ and $S(x)$ by $S_0(x)$ in the expression of I_n. By lemma 3, the scaled and centered T_n, say $T_{n,*}$, can be written as

$$T_{n,*} = d(n)\left(\int (\hat{S}_n(x) - S_0(x))^2\, dx - c(n)\right)$$

where

$$\begin{aligned} c(n) &= \int (\mathbb{E}\hat{S}_n(x) - S_0(x))^2\, dx + \sigma_2^2 \qquad (2)\\ \sigma_2^2 &= n^{-1}\Big\{hR(W)\int \frac{f(x)}{1-H(x)}\, dx \\ &\quad - h^2\left(\iint K(t)K(t+v)\, dt\, dv\right)\int F(x)F(x+hv)\, dx\Big\}. \end{aligned}$$

Thus, in testing H_0 against H_1 with significance level a we have

$$T_{n,*}/\sqrt{\mathbb{V}\mathrm{ar}\{T_{n,*}\}} \to N(0,1)$$

where

$$\mathbb{V}\mathrm{ar}\{T_{n,*}\} = \begin{cases} (2^{1/2}\sigma_1)^2 & \text{if} \quad nh^3 \to 0 \\ (2k\sigma_4)^2 & \text{if} \quad nh^3 \to +\infty \\ (2^{1/2}\lambda^{1/2}\sigma_1 + 2\lambda^{1/3}k\sigma_4)^2 & \text{if} \quad nh^3 \to \lambda. \end{cases} \qquad (3)$$

Consequently, the test suggests rejection of H_0 when $T_{n,*}(\mathbb{V}\mathrm{ar}\{T_{n,*}\})^{-1/2} > z_a$ where z_a is the standard normal quantile at level a. Implementation of the test is discussed in the next section.

4. Simulations

The primary objective of this section is the study of the power function of the proposed test for small samples. Efficient implementation of the test in practice is discussed first, together with the test's operational characteristics. The power function is then simulated for various sample sizes, amounts of censoring and lifetime distributions in order to provide numerical evidence for its small sample practical performance.

In order to obtain a computationally convenient formula for realizing the test, the sample density is employed as a weight function for the integrands in $T_{n,*}$'s definition. This has the additional benefit of trimming out low density areas. Repeated use of the fact that weighting the integrand of a functional

by the population's density leads to its expected value, which in turn can be reasonably estimated by its sample mean, yields the sample version of $T_{n,*}$,

$$\tilde{T}_{n,*} = d(n)\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{S}_n(X_i) - S_0(X_i)\right)^2 - \frac{h^2\mu_2^2(K)}{2n}\sum_{i=1}^{n}\hat{S}_n''(X_i)^2 - \hat{\sigma}_2^2\right\} \quad (4)$$

where the second term on the RHS of (4) results from applying the bias expression of $\hat{S}_n(x)$ (Theorem 1, [2]) on (2) and

$$\hat{\sigma}_2^2 = n^{-2}\Big\{hR(W)\sum_{i=1}^{n}\frac{f_n(X_i)}{1-\hat{H}(X_i)} - h^2R(K)\sum_{i=1}^{n}(1-\hat{S}_n(X_i))^2\,dx\Big\}. \quad (5)$$

The quantities involved in (4) and (5) are discussed next. First, the Epanechnikov kernel

$$K(x) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1-\frac{x^2}{5}\right) & \text{for } |x| \le \sqrt{5} \\ 0 & \text{otherwise} \end{cases}$$

is used throughout this section. Its integrated version, $W(x)$, is given by

$$W(x) = \int_{-\infty}^{x} K(u)\,du = \begin{cases} 0 & \text{for } x < -\sqrt{5} \\ -\frac{1}{100}\sqrt{5}x^3 + \frac{1}{2} + \frac{3}{20}\sqrt{5}x & \text{for } -\sqrt{5} \le x \le \sqrt{5} \\ 1 & \text{for } x > \sqrt{5}. \end{cases}$$

By direct calculation, the constants in (4), (5) and elsewhere throughout this section are given by

$$\mu_2^2(K) = 1, R(W) = \frac{26}{35}\sqrt{5}, \text{ and } R(K) = \frac{3}{25}\sqrt{5}. \quad (6)$$

Now, $\hat{S}_n(x)$, defined in (1), uses the MSE optimal (local) bandwidth of [15], (3.3.1). Note here that this is also the bandwidth h used in (4) and (5). As this expression involves unknown quantities, following [15], a practical version is

$$h = \left\{\frac{2\tilde{f}(x)\int xK(x)W(x)\,dx}{n(1-\hat{H}(x))(\tilde{f}'(x))^2\mu_2^2(K)}\right\}^{\frac{1}{3}}. \quad (7)$$

In (7), $\tilde{f}(x)$ and $\tilde{f}'(x)$ are estimates of the unknown $f(x)$ and $f'(x)$ respectively based on an exponential reference distribution with its mean estimated by maximum likelihood. That is,

$$\tilde{f}(x) = e^{-x\hat{\theta}^{-1}}, \quad \tilde{f}'(x) = -\hat{\theta}^{-2}e^{-x\hat{\theta}^{-1}}, \quad \hat{\theta} = \frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n}\Delta_i}. \quad (8)$$

Further, by direct calculation based on the Epanechnikov kernel, (7) is implemented with

$$\int_{-\sqrt{5}}^{\sqrt{5}} xK(x)W(x)\,dx = \frac{9}{70}\sqrt{5}.$$

Next, $f_n(x)$ is the estimate of the underlying density which is discussed extensively in [21]. It is given by

$$f_n(x) = \sum_{i=1}^{n} \frac{\Delta_i}{n(1-\hat{H}(X_i))h_0} K\left(\frac{x-X_i}{h_0}\right) \tag{9}$$

and is implemented with MSE optimal (local) bandwidth (see also [21], page 1525)

$$h_0 = \left\{ \frac{R(K)\tilde{f}(x)}{n(1-\hat{H}(x))\mu_2^2(K)\{\tilde{f}''(x)\}^2} \right\}^{\frac{1}{5}},$$

where $\tilde{f}''(x)$ denotes an estimate of the second derivative of the underlying density. Following the rationale that was used for obtaining (8), $\tilde{f}''(x) = \hat{\theta}^{-3}e^{-x\hat{\theta}^{-1}}$. Furthermore, in implementing $f_n(x)$, the reflection method of [13] has been employed to reduce boundary bias. This means that in practice, $K\left((x-X_i)h_0^{-1}\right)$ in (9) is replaced by

$$K\left((x-X_i)h_0^{-1}\right) + K\left((-x-X_i)h_0^{-1}\right).$$

Estimator $\hat{S}_n''(x)$ in (4) corresponds to an estimate of the negative of $f'(x)$. Based on [13], page 143, (8.3), it is straightforward to derive

$$f_n'(x) = -\frac{1}{nh_d^2\mu_2(K)} \sum_{i=1}^{n} h_d^{-1}(x-X_i)K\left((x-X_i)h_d^{-1}\right) \frac{\Delta_i}{1-\hat{H}(X_i)} \tag{10}$$

as the second derivative of $\hat{S}_n(x)$, adjusted for boundary bias. In (10), $h_d = (4/(5n))^{1/7}\,\tilde{\sigma}$ where $\tilde{\sigma}$ is the standard deviation of the sample. That is, h_d is the univariate version of the normal scale bandwidth selector of [6], page 815, (3.2). It is important to note that a separate simulation on the performance of $f_n'(x)$ (not reported here) indicated that the use of reflection further improves its performance at the boundary.

Attention is now shifted to the implementation of the denominator of the test statistic, i.e. $\mathbb{V}\mathrm{ar}\{T_{n,*}\}^{-1/2} \equiv 1/\sigma(\tilde{T}_{n,*})$. By the same weighting argument as in (4) and (5), and assuming the optimal smoothing version of (3) due to h and h_0, we have

$$\sigma(\tilde{T}_{n,*}) = \hat{\sigma}_1\sqrt{2\hat{\lambda}} + 2\hat{\lambda}^{1/3}k\hat{\sigma}_4 \tag{11}$$

where, $\hat{\lambda} = nh^3$ and

$$\hat{\sigma}_1^2 = \frac{s}{n}\sum_{i=1}^{n}\left(\frac{f_n(X_i)}{1-\hat{H}(X_i)}\right)^2$$

$$\hat{\sigma}_4^2 = \frac{s}{n}\sum_{i=1}^{n}\left(\hat{S}_n''(X_i)^2\frac{f_n(X_i)}{1-\hat{H}(X_i)} - \hat{S}_n''(X_i)^2(1-\hat{S}_n(X_i))\right) \tag{12}$$

$$s = R(B(u)) = \frac{136277}{29400}\sqrt{5}. \tag{13}$$

Note that (12) is realized by using the negative of (10) in place of $\hat{S}_n''(X_i)$ and (13) is obtained by direct calculation specifically for the Epanechnikov distribution function. Combining (4) and (11), the test statistic used in practice is $\tilde{T}_{n,*}/\sigma(\tilde{T}_{n,*})$.

Performance of the test statistic under the null and alternative hypotheses is discussed next. Three families of common lifetime distributions are assumed for this purpose: the exponential with rate λ ($\exp(\lambda)$) and the Weibull and Gamma distributions with shape parameter κ and scale parameter λ ($\mathrm{W}(\kappa,\lambda)$ and $\mathrm{G}(\kappa,\lambda)$ respectively). The null hypothesis is formulated as $H_0 : S(x) = S_0(x)$ where $S_0(x) = 1 - e^{-x}$. This corresponds to the survival function of any of the null distributions $\exp(1)$, $\mathrm{W}(1,1)$ or $\mathrm{G}(1,1)$. In order to estimate the power of the test, the probability of rejecting the null hypothesis given that the alternative is true is approximated at each one of 12 specific alternatives. Each alternative survival function, say $S_1(x)$, is seen as a member of a sequence which all belong to the same family of distributions with $S_0(x)$ and their parameter(s) selected so that the Kullback–Liebler divergence between $S_0(x)$ and $S_1(x)$ becomes increasingly deviant. Thus for each distribution, sample size and amount of censoring, the conditional probability of rejecting $H_0 : S(x) = S_0(x)$ given that $H_1 : S(x) = S_1(x)$ is true is approximated from $m = 10,000$ independent random samples by

$$P(\tilde{T}_{n,*}/\sigma(\tilde{T}_{n,*}) > \text{cut-off}) = \frac{\#\{\tilde{T}_{n,*}/\sigma(\tilde{T}_{n,*}) > \text{cut-off}\}}{m}, \tag{14}$$

where, for each specific alternative, $S_0(x)$ in the definition of $\tilde{T}_{n,*}$ in (4) is replaced by $S_1(x)$. Assuming significance level $a = 5\%$ (probability of type I error), the cut-off points in (14) are estimated by the 95% quantile of the numerical distribution of $\tilde{T}_{n,*}/\sigma(\tilde{T}_{n,*})$. The distribution of the statistic is approximated by generating 100,000 values for each different sample size, assuming no censoring and under H_0. Definition 7 of [11] is then used for obtaining the cut–off point. It has to be noted that the cut-off points vary by sample size but not by censoring level so as to get an indication on how censoring affects power.

The results of the simulation are presented in Fig. 1 which contains the power of the test for each combination of distribution, sample size and censoring level. The simulation results indicate that, as expected, power increases as divergence from H_0 increases. In parallel, power increases as sample size increases. On average, the test is consistent with the nominal level under no censoring. However, censoring has a drastic effect on the test's nominal level and power. For as low as 10% censoring, the nominal level of the test is (on average) doubled. As expected, the more the censoring increases, the more the type I error and power increases. We conjugate though that by using cut–off points based on the test statistic's distribution under censoring – calculated in an obvious manner – to get nominal level and power figures closer to the uncensored ones. Censoring aside, the rejection percentages of Fig. 1 suggest that the further the divergence from the null, the more sensitive the test becomes in detecting the inappropriate fit resulting from H_1. Furthermore it has to be noted that even though many

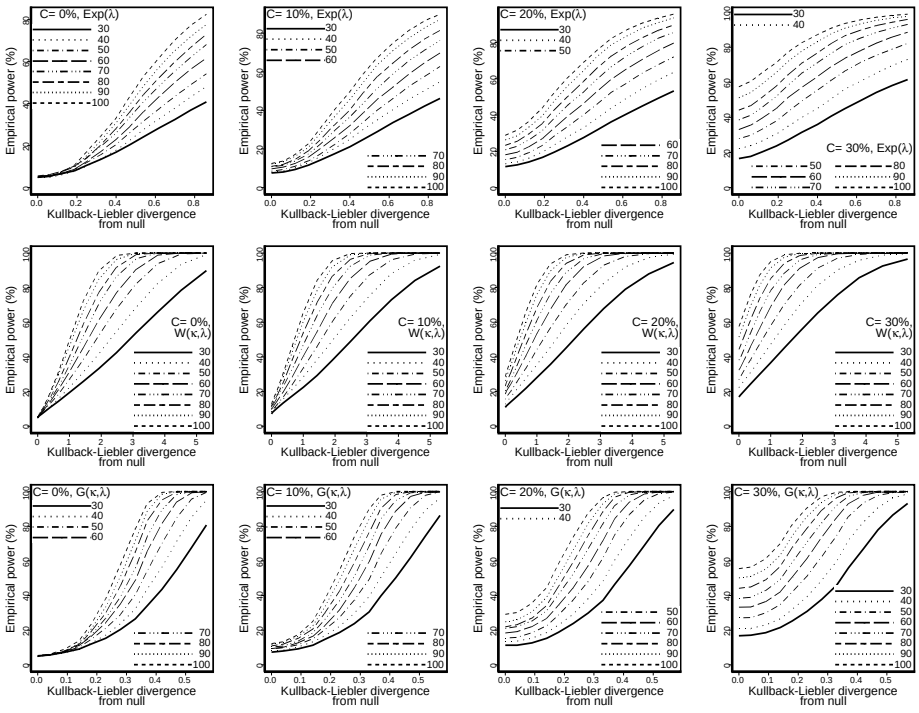

FIG 1. *Rejection percentages, based on 10,000 replications, of the suggested goodness–of–fit test when testing* $H_0 : S(x) = 1 - e^{-x}$ *versus*

- ***Top row:*** $H_1 : S(x) = 1 - \exp\{-\lambda x\}$ *(Exponential, denoted by: Exp(*λ*)) for increasing values of* $\lambda (= 1, 1.1, 1.2, \ldots, 2.2)$,
- ***Middle row:*** $H_1 : S(x) = 1 - \exp\{-\left(x\lambda^{-1}\right)^{\kappa}\}$ *(Weibull, denoted by: W(*κ*,* λ*)) for increasing values of* $\kappa = \lambda = 1, 1.1, 1.2, \ldots, 2.2$,
- ***Bottom row:*** $H_1 : S(x) = 1 - \frac{\gamma(\kappa, x/\lambda)}{\Gamma(\kappa)}$, *(the Gamma distribution, denoted by: G(*κ*,* λ*), where* γ *is the lower incomplete gamma function and* Γ *is the gamma function), for increasing values of* $\kappa = \lambda = 1, 1.05, 1.1, 1.15, \ldots, 1.6$.

Censoring (denoted by C) is either 0% (no censoring), 10%, 20% or 30% and the sample sizes considered are $n = 30, 40, \ldots, 100$ *at significance level* $a = 5\%$.

data driven estimates are used in (4) and (5), this does not prevent the test from possessing reasonable power. Theoretical explanation for this, at least asymptotically, is provided by first noticing that using $\hat{H}(x), f_n(x)$ and $f_n'(x)$ instead of $H(x), f(x)$ and $f'(x)$ respectively in (4) and (5) results in an error of order $o_p(n^{-1/2}), o(n^{-2/5})$ and $o(n^{-2/7})$ in each of these estimations. In (4), the substitution of $f'(x)$ by $f_n'(x)$ on the second term of the RHS (which involves the derivative estimate) will incur an additional term of order $n^{-1.62}$ (assuming optimal bandwidth of order $n^{-1/3}$) which converges to zero very quickly. Similarly in (5), the substitution of $H(x)$ and $f(x)$ by $\hat{H}(x)$ and $f_n(x)$ respectively will

result in an additional bias term of rate $n^{-2.7}$ which again, is negligible. Therefore, in practice it is expected that the test statistic, at least asymptotically, will perform quite reasonably.

In addition, for comparison purposes, the same examples under exactly the same settings as in Fig. 1, have been replicated using the Neyman data driven goodness of fit test for completely specified distributions. The test is implemented in the `R` package `surv2sample`, [20]. The package is not currently available from `CRAN`, however can be incorporated into existing MS Windows `R` installations via the `Rtools` package.

The comparison between the two methods indicates that the test suggested in the present research is more sensitive to subtle differences between the null and alternative distributions. However, as divergence from the null increases, the simulation results indicate that Neyman's test appears to be more powerful, albeit the rejection percentages of the present test are reasonable too. The source `R` code for obtaining the numbers used in Fig. 1–2 together with the code used for obtaining the cut-off points, as well as full instructions and comments, is available from the Journal's website.

5. Real data analysis

As real data analysis, the air conditioning unit failure data set of [22] is presented here to exhibit the practical usefulness of the test when employed to validate a parametric model. The data set consists of the time intervals, in hours, between successive failures of the air conditioning system of each member of a fleet of 13 Boeing 720 jet airplanes. The aim of [22] was to find a characterization of the distribution of the failure times at fleet level for maintenance purposes. After pooling the data (yielding thus a total of 213 observations), [22] investigated fitting an exponential survival function with the parameter of the distribution estimated by the mean of the failure intervals. The appropriateness of the fit was assesed by testing the null hypothesis

$$H_0 : S(x) = e^{-\frac{t}{93.14}} \tag{15}$$

which although not rejected by the Kolmogorov–Smirnov test, was found to be unreasonable for practical use as it suggests that failures decline with time; this is counter intuitive in view of wear out and ageing effects of air conditioning units. It further implies that all failure times follow the same distribution irrespective of which plane they come from, something questionable too as different planes might be exposed to different conditions which affect failure occurrences. Prompted by these issues, [14] developed models individually for each plane and used those as a basis for suggestions at the fleet level. Moreover, prompted by [7], another suggestion of [14] was the use of mixture distributions as an approximation of aggregated individual survival functions for the pooled data. This route was further followed by [1], [18] and [23], who all assessed the resulting fits graphically. As noted by [14] though, use of mixture distributions can be uncertain as adding or deleting one plane from the sample would change the

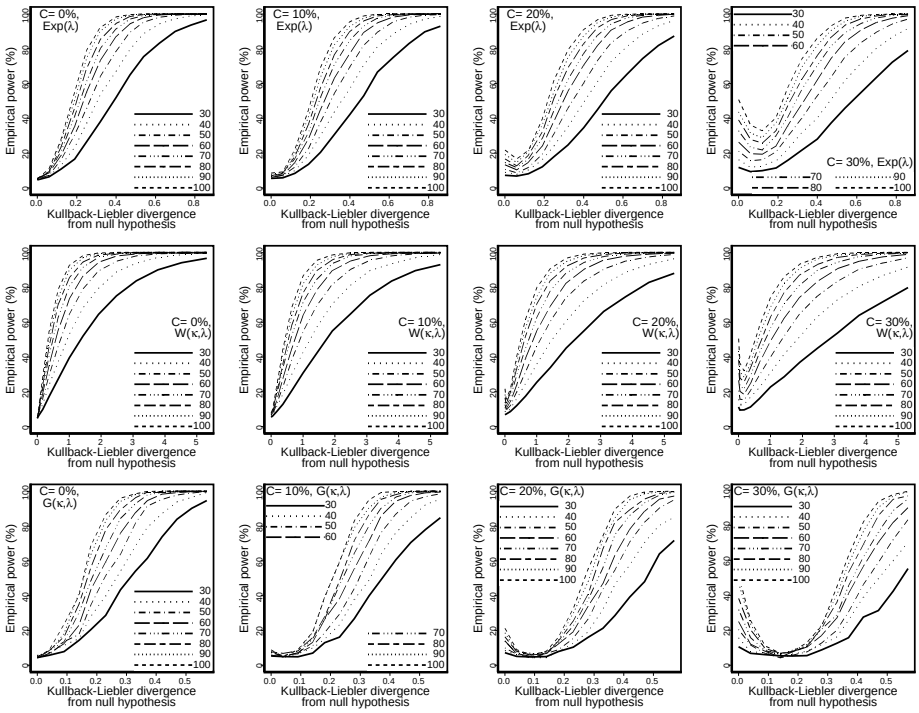

FIG 2. *Rejection percentages, based on 10,000 replications, of the Neyman smooth test for censored data when testing* $H_0 : S(x) = 1 - e^{-x}$ *versus*

- ***Top row:*** $H_1 : S(x) = 1 - \exp\{-\lambda x\}$ *(Exponential, denoted by: Exp(*λ*)) for increasing values of* $\lambda (= 1, 1.1, 1.2, \ldots, 2.2)$,
- ***Middle row:*** $H_1 : S(x) = 1 - \exp\{-\left(x\lambda^{-1}\right)^{\kappa}\}$ *(Weibull, denoted by: W(*κ*,* λ*)) for increasing values of* $\kappa = \lambda = 1, 1.1, 1.2, \ldots, 2.2$,
- ***Bottom row:*** $H_1 : S(x) = 1 - \frac{\gamma(\kappa, x/\lambda)}{\Gamma(\kappa)}$, *(the Gamma distribution, denoted by: G(*κ*,* λ*), where* γ *is the lower incomplete gamma function and* Γ *is the gamma function), for increasing values of* $\kappa = \lambda = 1, 1.05, 1.1, 1.15, \ldots, 1.6$.

Censoring (denoted by C) is either 0% (no censoring), 10%, 20% or 30% and the sample sizes considered are $n = 30, 40, \ldots, 100$ *at significance level* $a = 5\%$.

mixture distribution. However, while the last examples correspond to distributions with decreasing failure rate, which despite any ageing or wear out effects corroborates with the data pattern, increasing failure rate distributions for specific planes on the dataset have been investigated too. As an example, [17] and [8] tested the null hypothesis of exponentiality versus increasing failure rates based on the gamma distribution. In both cases the null is rejected. Yet another approach was based on the fatigue life distribution (also known as the Birnbaum–Saunders distribution, (BS)), which is used extensively in reliability applications to model failure times. Specifically, [5] tested (via the likelihood

ratio statistic) the null hypothesis that the data are coming from the BS distribution against the alternative that the underlying distribution is a β–BS. The null was rejected under any usual significance level ($p < 0.01$).

The fact that all aforementioned approaches are parametric in nature together with the fact that $\hat{S}_n(x)$ is geared towards revealing the pattern of the underlying survival makes the suggested goodness-of-fit test an impartial asset for testing the validity of a given parametric model. For example, using $\tilde{T}_{n,*}/\sigma(\tilde{T}_{n,*})$ to test (15) leads to a p-value of 0.1178 (the statistic's value is equal to 1.1857) which leads to the outcome that the null model may not be the best option to adopt. On the other hand, testing

$$H_0 : S_{\mathrm{AL}}(x) = (1 - 0.4276)e^{-0.00801x}(1 - 0.4276e^{-0.00801x})^{-1},$$

suggested by [1], leads to a p-value of 0.03247 ($\tilde{T}_{n,*}/\sigma(\tilde{T}_{n,*}) = 1.8456$) which gives strong evidence in favor of the model.

6. Discussion

This article has proposed a kernel based goodness–of–fit test, useful for assessing how well a parametric survival estimate matches the true curve under the random right censoring data setting.

The test is based on the asymptotic distribution of the discrepancy between the estimated and actual survival functions. Thus, it summarizes the divergence between observed values and the values expected under the model in question and as a result it is of a very general scope. However, specific applications include: a) comparison of the distribution of a data set to a normal distribution or in general to a fully specified distribution/survival function, b) testing the normality assumption in analysis of variance, c) testing the normality of residuals in regression and d) whether outcome frequencies follow a specified distribution.

Additionally, by straightforward adjustments, it can serve as a goodness–of–fit test for Cox proportional hazards models when they are expressed in terms of the survival function. Further, it can be potentially used as an alternative approach to resampling techniques as well as to graphical goodness–of–fit tests when an objective view of the discrepancy between the estimated and the actual model is needed.

7. Proof of Theorem 1

The present Section starts with an outline description of the proof and continues with its full mathematical details in Subsections 7.1 and 7.2.

The proof begins by decomposing $\hat{F}(x)$'s ISE expression into four terms: (a) a symmetric U-statistic, (b) its diagonal, (c) the integrated square bias of $\hat{F}(x)$ and (d) the integrated product of the stochastic and systematic parts of the deviation of $\hat{F}(x)$ from $F(x)$. These terms are studied in Lemmas 1–4 and the resulting expressions are substituted back so as to obtain the central limit theorem for I_n.

In detail, Lemma 1 shows that (a) is an asymptotically zero mean normal random variable. This is established by expressing the symmetric U-statistic as a martingale and then applying a suitable central limit theorem to quantify the distribution of the statistic. Validation of the theorem's applicability is done through two conditional Lindeberg and variance conditions given in Corollary 3.1 of [10]. Proceeding with the term in (b), Lemma 2 uses direct calculation to show that the expression there equals σ_2^2 plus a negligible (under optimal bandwidth) term. Lemma 3 shows that the expected ISE value equals the term in (c) plus the result of Lemma 2. Thus, the result of Lemma 3 is used to shape the LHS of Theorem 1 as well as to cancel out the σ_2^2 term resulting from Lemma 2. Finally, in Lemma 4 the term in (d) is written as a sum of independent and identically distributed random variables and so, by a central limit theorem, verification of the theorem's conditions and straightforward calculations, is proved that the term is asymptotically normally distributed.

Putting back the results of all four lemmas in the decomposed I_n expression and noting that for optimal bandwidth the asymptotic terms are negligible concludes the proof.

7.1. ISE decomposition

The proof of Theorem 1 starts with the ISE decomposition. We have

$$\begin{aligned} I_n &= \int (\hat{F}(x) - \mathbb{E}\hat{F}(x) + \mathbb{E}\hat{F}(x) - F(x))^2 \, dx \\ &= \int (\hat{F}(x) - \mathbb{E}\hat{F}(x))^2 \, dx + \int (\mathbb{E}\hat{F}(x) - F(x))^2 \, dx \\ &\quad + 2 \int (\hat{F}(x) - \mathbb{E}\hat{F}(x))(\mathbb{E}\hat{F}(x) - F(x)) \, dx. \end{aligned}$$

Let $Y_i = (X_i, \Delta_i), i = 1, \ldots, n$ denote the observed sample. Then,

$$\int (\hat{F}(x) - \mathbb{E}\hat{F}(x))^2 \, dx = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} H_n(Y_i, Y_j)$$

where

$$\begin{aligned} H_n(Y_i, Y_j) &= \int R_i(x) R_j(x) \, dx \\ R_i(x) &= \frac{\Delta_i}{1 - H(X_i)} W\left(\frac{x - X_i}{h}\right) - \mathbb{E}\left\{\frac{\Delta_i}{1 - H(X_i)} W\left(\frac{x - X_i}{h}\right)\right\}. \end{aligned}$$

Note also that

$$n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} H_n(Y_i, Y_j) = n^{-2} \sum_{i=1}^{n} \sum_{i \neq j}^{n} H_n(Y_i, Y_j) + n^{-2} \sum_{i=1}^{n} H_n(Y_i, Y_i).$$

Then the ISE becomes

$$\begin{aligned} I_n &= n^{-2}\sum_{i=1}^{n}\sum_{i\neq j}^{n} H_n(Y_i,Y_j) \\ &\quad + n^{-2}\sum_{i=1}^{n} H_n(Y_i,Y_i) + \int (\mathbb{E}\hat{F}(x) - F(x))^2\,dx \\ &\quad + 2\int (\hat{F}(x) - \mathbb{E}\hat{F}(x))(\mathbb{E}\hat{F}(x) - F(x))\,dx \\ &\equiv \hat{I}_1 + \hat{I}_2 + \hat{I}_3 + 2\hat{I}_4. \end{aligned} \tag{16}$$

Apply Lemmas 1, 2, 3 and 4 to (16) to get

$$I_n - \mathbb{E}I_n = 2^{1/2}n^{-1}h^{3/2}\sigma_1 Z + \sigma_2^2 - \sigma_2^2 + n^{-1/2}h^3 k\sigma_4 Z + O_p(n^{-3/2}h)$$

and note that for optimal bandwidth $h \sim n^{-1/3}$, $O_p(n^{-3/2}h) = O_p(n^{-7/6})$ which is asymptotically negligible. This completes the proof.

7.2. Auxiliary Lemmas

Lemma 1. *As $n \to +\infty$, we have $\hat{I}_1 \sim N(0, 2n^{-2}h^3\sigma_1^2)$.*

Proof. The proof is analogous to the proof of Theorem 1, [9]. Note that,

$$H_n(Y_i,Y_j) = \int R_i(x)R_j(x)\,dx = \int R_j(x)R_i(x)\,dx = H_n(Y_j,Y_i)$$

and so $\hat{I}_1$ is a symmetric U-statistic. Therefore

$$\hat{I}_1 = n^{-2}\sum_{i=1}^{n}\sum_{i\neq j} H_n(Y_i,Y_j) = 2n^{-2}\sum_{i=2}^{n}\sum_{j=1}^{i-1} H_n(Y_i,Y_j) = 2n^{-2}\sum_{i=2}^{n} T_i$$

where

$$T_i = \sum_{j=1}^{i-1} H_n(Y_i,Y_j).$$

Corollary 3.1 of [10] will be used to prove the asymptotic distribution of $\hat{I}_1$. Note that it is readily established that

$$\sum_{i=2}^{k} T_i, k = 2,\ldots,n$$

is a zero mean, square integrable martingale with differences T_i. Moreover, in the present setting the conditional Lindeberg condition of corollary 3.1, [10], is equivalent to its unconditional version ((3.6) in [10], page 53). This means that the proof will be completed by showing

(a) For all $\varepsilon > 0$, as $n \to +\infty$

$$s_n^{-2} \sum_{i=1}^{n} \mathbb{E}\left(T_i^2 I(|T_i| > \varepsilon s_n)\right) \xrightarrow{p} 0$$

where $s_n^2 = \mathbb{E}(\hat{I}_1^2)$ and

(b)

$$s_n^{-2} V_n^2 = \sum_{i=1}^{n} \mathbb{E}\left(T_i^2 | Y_1, Y_2, \ldots, Y_{i-1}\right) \xrightarrow{p} 1.$$

First, note that

$$\begin{aligned}
\mathbb{E}T_i^2 &= \mathbb{E}\left(\sum_{j=1}^{i-1} H_n(Y_i, Y_j)\right)^2 \\
&= \mathbb{E}\left(\sum_{j=1}^{i-1} H_n(Y_i, Y_j)\right)\left(\sum_{k=1}^{i-1} H_n(Y_i, Y_k)\right) \\
&= \sum_{j=1}^{i-1}\sum_{k=1}^{i-1} \mathbb{E}H_n(Y_i, Y_j) H_n(Y_i, Y_k) \\
&= \sum_{j=1}^{i-1} \mathbb{E}H_n^2(Y_i, Y_j) + 2\sum\sum_{j<k} \mathbb{E}H_n(Y_i, Y_j) H_n(Y_i, Y_k) \\
&= \sum_{j=1}^{i-1} \mathbb{E}H_n^2(Y_i, Y_j) \text{ by (48) in Lemma 6} \\
&= (i-1)\mathbb{E}H_n^2(Y_i, Y_j) \text{ for any } i, j
\end{aligned}$$

since the $Y_i, i = 1, \ldots, n$ are i.i.d. Hence

$$s_n^2 = \sum_{i=2}^{n} \mathbb{E}T_i^2 = \frac{1}{2} n(n-1)\mathbb{E}H_n^2(Y_i, Y_j). \tag{17}$$

It is immediately then seen that (a) and (b) imply $\hat{I}_1 / s_n \sim N(0,1)$ which in turn verifies the statement of the lemma. Starting with the proof of (a), first note that by Chebyshev's inequality

$$s_n^{-2} \sum_{i=1}^{n} \mathbb{E}\left(T_i^2 I(|T_i| > \varepsilon s_n)\right) \leq \varepsilon^{-2} s_n^{-4} \sum_{i=1}^{n} \mathbb{E}\left(|T_i|^4\right). \tag{18}$$

Now,

$$\sum_{i=1}^{n} \mathbb{E}\left(|T_i|^4\right) = \sum_{i=1}^{n} \mathbb{E}\left|\sum_{j=1}^{i-1} H_n(Y_i, Y_j)\right|^4. \tag{19}$$

Expanding the summand of the RHS of (19) and using (49) and (50) of Lemma 6 yields

$$\left|\sum_{j=1}^{i-1} H_n(Y_i,Y_j)\right|^4 = \sum_{j=1}^{i-1}\left|H_n^4(Y_i,Y_j)\right| + 3\sum_{1\le j,k\le i-1;j\neq k}\sum\sum\sum\left|H_n^2(Y_i,Y_j)H_n^2(Y_i,Y_k)\right|. \tag{20}$$

Combining (19) and (20) we get

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\left(|T_i|^4\right) &= \sum_{i=1}^{n}\mathbb{E}\sum_{j=1}^{i-1}\left|H_n^4(Y_i,Y_j)\right| \\ &\quad + 3\sum_{i=1}^{n}\mathbb{E}\sum_{1\le j,k\le i-1;j\neq k}\sum\sum\left|H_n^2(Y_i,Y_j)H_n^2(Y_i,Y_k)\right| \\ &= \sum_{i=1}^{n}(i-1)\mathbb{E}\left|H_n^4(Y_i,Y_j)\right| \\ &\quad + 3\sum_{i=1}^{n}(i-1)(i-2)\mathbb{E}\left|H_n^2(Y_i,Y_j)H_n^2(Y_i,Y_k)\right| \quad j,k \text{ fixed} \\ &\le n^2C\mathbb{E}\left(H_n^4(Y_i,Y_j)\right) + n^3C\mathbb{E}\left(H_n^2(Y_i,Y_j)H_n^2(Y_i,Y_k)\right)\end{aligned}$$

where C is a positive generic constant. Now, by (36) in Lemma 5 we have that for optimally chosen bandwidth ($h \sim n^{-1/3}$), $n^3\mathbb{E}\left(H_n^2(Y_i,Y_j)H_n^2(Y_i,Y_k)\right) = O(n)$ which is smaller than $n^3C\mathbb{E}\left(H_n^4(Y_i,Y_k)\right) = O(n^{4/3})$. Thus we conclude

$$\sum_{i=1}^{n}\mathbb{E}\left(|T_i|^4\right) \le n^3C\mathbb{E}\left(H_n^4(Y_i,Y_j)\right). \tag{21}$$

Using (17), (21) and (37) in (18) proves (a). We now turn our attention to prove (b). First,

$$\begin{aligned}V_n^2 &= \sum_{i=2}^{n}\left(\mathbb{E}\left\{\sum_{j=1}^{i-1}H_n(Y_i,Y_j)\right\}^2 |Y_1,\ldots,Y_{i-1}\right) \\ &= \sum_{i=2}^{n}\mathbb{E}\left(\sum_{j=1}^{i-1}H_n(Y_i,Y_j)^2\Big|Y_1,\ldots,Y_{i-1}\right) \\ &\quad + 2\sum_{i=2}^{n}\mathbb{E}\left(\sum_{j=1}^{i-1}\sum_{l=j+1}^{i-1}H_n(Y_i,Y_j)H_n(Y_i,Y_l)\Big|Y_1,\ldots,Y_{i-1}\right).\end{aligned}$$

It is easily seen that (b) will be proved if we show that

$$s_n^{-4}\mathbb{E}(V_n^2 - s_n^2)^2 \to 0. \tag{22}$$

Set

$$v_i = \mathbb{E}\left(\sum_{j=1}^{i-1} H_n(Y_i, Y_j)^2 \Big| Y_1, \ldots, Y_{i-1}\right).$$

Then

$$\mathbb{E}V_n^4 = 2\sum_{2\le i\le j\le n}\sum \mathbb{E}(v_i v_j) + \sum_{i=2}^{n} \mathbb{E}v_i^2$$

Working in exactly the same way as in as in [9], page 5 we have

$$\begin{aligned}\mathbb{E}V_n^4 &= 2\sum_{i=2}^{n}(i-1)(i-2)(2n-2i+1)\mathbb{E}\left(\mathbb{E}H_n^*(X_i,x)H_n^*(X_j,y)\right)^2\\ &+\sum_{i=2}^{n}(i-1)(2n-2i+1)\mathbb{V}\mathrm{ar}\{(\mathbb{E}H_n^*(X_i,x)H_n^*(X_i,y))\}\\ &+\frac{1}{2}n(n-1)\mathbb{E}\left(\mathbb{E}H_n^*(X_i,x)H_n^*(X_i,y)\right)^2\end{aligned}$$

with

$$\begin{aligned}H_n^*(X_i,x) &= \int A(u,x)A(u,y)\,du,\\ A(u,x) &= \frac{1}{1-H(u)}W\left(\frac{u-x}{h}\right) - \mathbb{E}\frac{1}{1-H(X_i)}W\left(\frac{u-X_i}{h}\right)\end{aligned}$$

Then, for a positive generic constant C

$$\mathbb{E}(V_n^2 - s_n^2)^2 \le C\left\{n^4\mathbb{E}\left(\mathbb{E}H_n^*(X_i,x)H_n^*(X_j,y)\right)^2 + n^3\mathbb{E}H_n^2(X_i,X_j)\right\}. \quad (23)$$

Working as in the proof of (4.6) of [9], we get that

$$\mathbb{E}\left(\mathbb{E}H_n^*(X_i,x)H_n^*(X_j,y)\right)^2 = O(h^7) \quad (24)$$

Now, (23), (24) and (36) prove (22) which in turn proves (b). □

For the remainder of this Section and in order to simplify notation define

$$w_i(x) = \frac{\Delta_i}{1-H(X_i)}W\left(\frac{x-X_i}{h}\right).$$

Lemma 2. *As* $n \to +\infty$

$$\hat{I}_2 = \sigma_2^2 + O_p(n^{-3/2}h).$$

Proof. Write

$$\hat{I}_2 = n^{-2}\sum_{i=1}^{n} L_i \equiv n^{-2}\sum_{i=1}^{n}\int \left(w_i(x) - \mathbb{E}w_i(x)\right)^2 dx.$$

Now,

$$\begin{aligned}\mathbb{E}L_i &= \mathbb{E}\int \left(w_i(x) - \mathbb{E}w_i(x)\right)^2 dx \\ &= \int \mathbb{E}w_i^2(x)\,dx + \int \left\{\mathbb{E}w_i(x)\right\}^2 dx - 2\mathbb{E}\int \left\{w_i(x)\right\}\left\{\mathbb{E}w_i(x)\right\} dx \\ &= \int \mathbb{E}w_i^2(x)\,dx - \int \left\{\mathbb{E}w_i(x)\right\}^2 dx.\end{aligned}$$

By the definition of $w_i(x)$, (40) for $x = y$ and (41),

$$\begin{aligned}\mathbb{E}L_i &= \int \left(h\int W^2(t)\frac{f(x-ht)}{1-H(x-ht)}\,dt\right) dx \\ &\quad - \int \left(h\int K(t)F(x-ht)\,dt\right)^2 dx \\ &= hR(W)\int \frac{f(x)}{1-H(x)}\,dx \\ &\quad - h^2\left(\iint K(t)K(t+v)\,dt\,dv\right)\int F(x)F(x+hv)\,dx + o(h^2) \qquad (25)\end{aligned}$$

after using a Taylor series expansion around x on the term $f(x)(1-H(x))^{-1}$ in the third step above. Also, for a positive generic constant C_1,

$$\begin{aligned}\mathbb{E}L_i^2 &= \iint \mathbb{E}\Big(\left\{w_i(x) - \mathbb{E}w_i(x)\right\}^2\left\{w_i(y) - \mathbb{E}w_i(y)\right\}^2\Big)\,dx\,dy \\ &\le C_1\iint \mathbb{E}w_i^2(x)w_i^2(y)\,dx\,dy + C_1\iint \mathbb{E}w_i(x)w_i^2(y)\,dx\,dy \\ &\qquad + C_1\left(\int \mathbb{E}w_i(x)\right)^2. \qquad (26)\end{aligned}$$

Applying the definition of $w_i(x)$ and (40) for $x = y$ yields

$$\left(\int \mathbb{E}w_i(x)\right)^2 = O(h^2). \qquad (27)$$

Similarly

$$\begin{aligned}\iint \mathbb{E}w_i(x)w_i^2(y)\,dx\,dy &= h^2\left(\iint W(t)W^2(t+v)\,dt\,dv\right) \\ &\quad \times \int \frac{f(x+hv)}{(1-H(x+hv))^2}\,dx + o(h^2) = O(h^2) \qquad (28)\end{aligned}$$

and

$$\begin{aligned}\iint \mathbb{E}w_i^2(x)w_i^2(y)\,dx\,dy &= h^2\left(\iint W^2(t)W^2(t+v)\,dt\,dv\right) \\ &\quad \times \int \frac{f(x+hv)}{(1-H(x+hv))^3}\,dx + o(h^2) = O(h^2). \qquad (29)\end{aligned}$$

Using (27)–(29) back to (26) yields $\mathbb{E}L_i^2 = O(h^2)$ and so

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left\{\hat{I}_2\right\} &= \mathbb{V}\mathrm{ar}\left\{n^{-2}\sum_{i=1}^{n} L_i\right\} \\ &= n^{-4}n\mathbb{V}\mathrm{ar}\left\{L_i\right\}, \text{ since } Y_i \text{ and therefore } L_i \text{ are i.i.d.,} \\ &= n^{-3}\left\{\mathbb{E}L_i^2 - (\mathbb{E}L_i)^2\right\} = n^{-3}\mathbb{E}L_i^2 = n^{-3}O(h^2) = O(n^{-3}h^2). \quad (30)\end{aligned}$$

By (25), setting $\sigma_2^2 = n^{-1}\mathbb{E}L_i$ and using (30) completes the proof. □

Lemma 3.

$$\mathbb{E}\int\left(\hat{F}(x) - F(x)\right)^2 dx = \int\left(\mathbb{E}\hat{F}(x) - F(x)\right)^2 dx + \sigma_2^2 + O_p(n^{-3/2}h)$$

Proof. This is verified by straight calculation

$$\begin{aligned}\mathbb{E}\int\left(\hat{F}(x) - F(x)\right)^2 dx &= \mathbb{E}\int\left(\hat{F}(x) - \mathbb{E}\hat{F}(x) + \mathbb{E}\hat{F}(x) - F(x)\right)^2 dx \\ &= \mathbb{E}\int\left(\hat{F}(x) - \mathbb{E}\hat{F}(x)\right)^2 dx + \mathbb{E}\int\left(\mathbb{E}\hat{F}(x) - F(x)\right)^2 dx \\ &\quad -2\mathbb{E}\int\left(\hat{F}(x) - \mathbb{E}\hat{F}(x)\right)\left(\mathbb{E}\hat{F}(x) - F(x)\right) dx \\ &= \mathbb{E}\int\left(\hat{F}(x) - \mathbb{E}\hat{F}(x)\right)^2 dx + \mathbb{E}\int\left(\mathbb{E}\hat{F}(x) - F(x)\right)^2 dx \\ &\quad -\mu_2(K)h^2\mathbb{E}\int\left(\hat{F}(x) - \mathbb{E}\hat{F}(x)\right)F''(x)\,dx,\end{aligned}$$

by Theorem 1 in [2]. Now, note that

$$\int\left(\hat{F}(x) - \mathbb{E}\hat{F}(x)\right)^2 dx = \hat{I}_2$$

and that

$$\mathbb{E}\int\left(\hat{F}(x) - \mathbb{E}\hat{F}(x)\right)F''(x)\,dx = 0.$$

Then

$$\mathbb{E}\int\left(\hat{F}(x) - F(x)\right)^2 dx = \mathbb{E}\int\left(\mathbb{E}\hat{F}(x) - F(x)\right)^2 dx + \hat{I}_2$$

and so, by applying Lemma 2, the result follows immediately. □

Lemma 4. *Under conditions 1, 2, 3 and under* $H_0 : F(x) = F_0(x)$,

$$\hat{I}_4 \sim N(0, n^{-1}h^6k\sigma_4^2). \quad (31)$$

Proof. Let

$$\hat{I}_4 = \int(\hat{F}(x) - \mathbb{E}\hat{F}(x))(\mathbb{E}\hat{F}(x) - F(x))\,dx = n^{-1}\sum_{i=1}^{n} Z_i$$

with

$$Z_i = \int (w_i(x) - \mathbb{E}w_i(x))\,(\mathbb{E}\hat{F}(x) - F(x))\,dx.$$

Write $Z_i = D_i - \mathbb{E}D_i$ with

$$D_i = \int w_i(x)(\mathbb{E}\hat{F}(x) - F(x))\,dx.$$

Obviously $\mathbb{E}Z_i = 0$. Now, set $D_{i,1} = \mathbb{E}D_i, D_{i,2} = \mathbb{E}D_i^2$ and note that

$$\begin{aligned}\mathbb{E}Z_i^2 &= \mathbb{E}(D_i - \mathbb{E}D_i)^2 = \mathbb{E}(D_i^2 - 2D_i\mathbb{E}D_i + (\mathbb{E}D_i)^2\\ &= \mathbb{E}D_i^2 - 2(\mathbb{E}D_i)^2 + (\mathbb{E}D_i)^2 = \mathbb{E}D_i^2 - (\mathbb{E}D_i)^2 = D_{i,2} - D_{i,1}^2. \quad (32)\end{aligned}$$

From Theorem 1, [2],

$$\mathbb{E}\hat{F}(x) - F(x) = \frac{h^2}{2}F''(x)\mu_2(K) + O(h^4). \quad (33)$$

Then we have

$$\begin{aligned}D_{i,1} &= \int \mathbb{E}\frac{\Delta_i}{1 - H(X_i)}W\left(\frac{x - X_i}{h}\right)(\mathbb{E}\hat{F}(x) - F(x))\,dx\\ &= \int (\mathbb{E}\hat{F}(x) - F(x))\,dx \int \frac{f(y)}{1 - H(y)}W\left(\frac{x - y}{h}\right)dx\\ &= \frac{h^3}{2}\mu_2(K)\int F''(x)\,dx \int K(t)F(x - ht)\,dt + o(h^2) \text{ by (33) and (41)}\\ &= \frac{h^3}{2}\mu_2(K)\int F''(x)F(x)\,dx + o(h^2). \quad (34)\end{aligned}$$

Also,

$$\begin{aligned}D_{i,2} &= \iint \mathbb{E}\frac{\Delta_i}{(1 - H(X_i))^2}W\left(\frac{x - X_i}{h}\right)W\left(\frac{y - X_i}{h}\right)\\ &\qquad \times (\mathbb{E}\hat{F}(x) - F(x))(\mathbb{E}\hat{F}(y) - F(y))\,dx\,dy\\ &= h\iiint W(t)W\left(\frac{y - x + ht}{h}\right)\frac{f(x - ht)}{1 - H(x - ht)}\\ &\qquad \times (\mathbb{E}\hat{F}(x) - F(x))(\mathbb{E}\hat{F}(y) - F(y))\,dt\,dx\,dy \text{ by (40)}.\end{aligned}$$

Using the change of variable $y - x = hu$ yields,

$$\begin{aligned}D_{i,2} = h^2 &\iint (\mathbb{E}\hat{F}(x) - F(x))(\mathbb{E}\hat{F}(x + hu) - F(x + hu))\,dx\,du\\ &\times \int W(t)W\,(u + t)\,\frac{f(x - ht)}{1 - H(x - ht)}\,dt\end{aligned}$$

$$= \frac{h^6}{4}\mu_2^2(K) \iint F''(x)F''(x+hu)\,dx\,du \times \int W(t)W\,(u+t)\,\frac{f(x-ht)}{1-H(x-ht)}\,dt + o(h^6) \text{ by (33)}.$$

Expanding in Taylor series around x gives

$$D_{i,2} = \frac{h^6}{4}\mu_2^2(K)\left\{\int (F''(x))^2\frac{f(x)}{1-H(x)}\,dx\right\} R(B(u)) + o(h^6). \tag{35}$$

By (34) and (35), (32) becomes

$$\mathbb{E}Z_i^2 = \frac{h^6}{4}\mu_2^2(K)\left\{\int (F''(x))^2\frac{f(x)}{1-H(x)}\,dx\right\} R(B(u)) - \frac{h^6}{4}\mu_2^2(K)\int F''(x)^2F(x)\,dx + o(h^2).$$

Working in the same way it is shown that $\mathbb{E}Z_i^4 = O(h^{12})$. By setting

$$s_n^2 = \sum_{i=1}^n \mathbb{E}Z_i^2$$

we get

$$s_n^{-2}\sum_{i=1}^n \mathbb{E}\left\{Z_i^2 I(|Z_i| > \varepsilon s_n)\right\} \le \varepsilon^{-2}s_n^{-4}\sum_{i=1}^n \mathbb{E}Z_i^4 \to 0$$

as $n \to +\infty$. Therefore, $\hat{I}_4$ is normally distributed with zero mean and variance

$$\begin{aligned}\mathbb{V}\text{ar}\left\{\hat{I}_4\right\} &= \mathbb{V}\text{ar}\left\{n^{-1}\sum_{i=1}^n Z_i\right\}\\ &= n^{-2}n\mathbb{V}\text{ar}\left\{Z_i\right\}, \text{ since } Y_i \text{ and therefore } Z_i \text{ are i.i.d.},\\ &= n^{-1}(\mathbb{E}Z_i^2 - (\mathbb{E}Z_i)^2) = n^{-1}\mathbb{E}Z_i^2.\end{aligned}$$

□

Lemma 5. *For i, j fixed*

$$\mathbb{E}H_n^2(Y_i, Y_j) = h^3 R\left(f(x)(1-H(x))^{-1}\right) R(B(u)) + O(h^5) \tag{36}$$

$$\mathbb{E}H_n^4(Y_i, Y_j) = O(h^5). \tag{37}$$

Proof. The proof is based on conditioning on the number of the uncensored observations of the observed sample $Y_i = (X_i, \Delta_i), i = 1, \ldots, n$. If N denotes the number of the uncensored observations then $N \sim$Binomial(n, p) where $p = \int f(x)(1-H(x))\,dx$. For given $N = \nu$, $(X_i : \Delta_i = 1)$ is a set of i.i.d random variables with density $f(x)(1-H(x))/p$ for $\nu = 1, 2, \ldots, n$. Now,

$$\mathbb{E}H_n^2(Y_i, Y_j) = \iint \left\{\mathbb{E}R_i(x)R_i(y)\right\}^2\,dx\,dy \tag{38}$$

For fixed i,

$$\begin{aligned}\mathbb{E}R_i(x)R_i(y) &= \mathbb{E}\left\{w_i(x)-\mathbb{E}w_i(x)\right\}\left\{w_i(y)-\mathbb{E}w_i(y)\right\}\\ &= \mathbb{E}\left\{w_i(x)w_i(y)\right\}-w_i(x)\mathbb{E}w_i(y)-w_i(y)\mathbb{E}w_i(x)+\mathbb{E}\left\{w_i(y)\mathbb{E}w_i(x)\right\}\\ &= \mathbb{E}\left\{w_i(x)w_i(y)\right\}-\mathbb{E}\left\{w_i(y)\mathbb{E}w_i(x)\right\}. \end{aligned} \tag{39}$$

Since i is fixed, $(X_i : \Delta_i = 1)$ is a Bernouli random variable with mean p. Then,

$$\begin{aligned}&\mathbb{E}w_i(x)w_i(y)\\ &= \mathbb{E}\left\{\mathbb{E}\left(\frac{\Delta_i}{1-H(X_i)}W\left(\frac{x-X_i}{h}\right)\frac{\Delta_i}{1-H(X_i)}W\left(\frac{y-X_i}{h}\right)\Big|\Delta_i=1\right)\right\}\\ &= \mathbb{E}\left\{\int\frac{1}{1-H(z)}W\left(\frac{x-z}{h}\right)\frac{1}{1-H(z)}W\left(\frac{y-z}{h}\right)\frac{f(z)(1-H(z))}{p}\,dz\right\}\\ &= \mathbb{E}\left\{\frac{1}{p}\int\frac{f(z)}{1-H(z)}W\left(\frac{x-z}{h}\right)W\left(\frac{y-z}{h}\right)\,dz\right\}\\ &= p\frac{1}{p}\int\frac{f(z)}{1-H(z)}W\left(\frac{x-z}{h}\right)W\left(\frac{y-z}{h}\right)\,dz\\ &= h\int W(t)W\left(\frac{y-x+ht}{h}\right)\frac{f(x-ht)}{1-H(x-ht)}\,dt \end{aligned} \tag{40}$$

where in the last step above the change of variable $x-z=ht$ was used. Similarly,

$$\begin{aligned}\mathbb{E}\left(w_i(x)\right) &= \mathbb{E}\left\{\mathbb{E}\left(\frac{\Delta_i}{1-H(X_i)}W\left(\frac{x-X_i}{h}\right)\Big|\Delta_i=1\right)\right\}\\ &= \mathbb{E}\left\{\int\frac{1}{1-H(z)}W\left(\frac{x-z}{h}\right)\frac{f(z)(1-H(z))}{p}\,dz\right\}\\ &= \mathbb{E}\left\{\frac{1}{p}\int f(z)W\left(\frac{x-z}{h}\right)\,dz\right\} = p\frac{1}{p}\int W\left(\frac{x-z}{h}\right)f(z)\,dz\\ &= h\int W(t)f(x-ht)\,dt = h\int K(t)F(x-ht)\,dt. \end{aligned} \tag{41}$$

By (41),

$$\begin{aligned}\mathbb{E}w_i(x)\mathbb{E}w_i(y) &= \left(h\int K(t)F(x-ht)\,dt\right)\left(h\int K(t)F(y-ht)\,dt\right)\\ &= h^2\left(\int K(t)F(x-ht)\,dt\right)\left(\int K(t)F(y-ht)\,dt\right). \end{aligned} \tag{42}$$

Substitute (40) and (42) back to (39) to get

$$\begin{aligned}\mathbb{E}R_i(x)R_i(y) &= h\int W(t)W\left(\frac{y-x+ht}{h}\right)\frac{f(x-ht)}{1-H(x-ht)}\,dt\\ &\quad - h^2\left(\int K(t)F(x-ht)\,dt\right)\left(\int K(t)F(y-ht)\,dt\right). \end{aligned} \tag{43}$$

Substitute (43) back to (38) to get

$$\mathbb{E}H_n^2(Y_i,Y_j) = \iint \Big\{ h \int W(t) W\left(\frac{y-x+ht}{h}\right) \frac{f(x-ht)}{1-H(x-ht)}\, dt - h^2 \left(\int K(t)F(x-ht)\, dt\right)\left(\int K(t)F(y-ht)\, dt\right)\Big\}^2 dx\, dy$$

and use the change of variable $y - x = uh$ to get

$$\begin{aligned}\mathbb{E}H_n^2(Y_i,Y_j) &= h \iint h^2 \Big\{ \int W(t)W(u+t)\frac{f(x-ht)}{1-H(x-ht)}\, dt \\ &\quad - h\left(\int K(t)F(x-ht)\, dt\right) \\ &\quad \times \left(\int K(t)F(x+uh-ht)\, dt\right)\Big\}^2 dx\, du \\ &= h^3 \iint \Big\{ \int W(t)W(u+t)\frac{f(x-ht)}{1-H(x-ht)}\, dt \\ &\quad - h\left(\int K(t)F(x-ht)\, dt\right) \\ &\quad \times \left(\int K(t)F(x+uh-ht)\, dt\right)\Big\}^2 dx\, du. \end{aligned} \tag{44}$$

Note that the terms

$$h^2 \left(\int K(t)F(x-ht)\, dt \int K(t)F(x+uh-ht)\, dt\right)^2$$

and

$$h\left\{\int W(t)W(u+t)\frac{f(x-ht)}{1-H(x-ht)}\, dt\right\}\left\{\int K(t)F(x-ht)\, dt\right\} \times \left\{\int K(t)F(x+uh-ht)\, dt\right\}$$

are for optimal bandwidth $h \sim n^{-1/3}$, asymptotically negligible compared to

$$\left(\int W(t)W(u+t)\frac{f(x-ht)}{1-H(x-ht)}\, dt\right)^2.$$

With this observation into account, (44) becomes

$$\begin{aligned}\mathbb{E}H_n^2(Y_i,Y_j) &= h^3 \iint \Big\{ \int W(t)W(u+t)\frac{f(x-ht)}{1-H(x-ht)}\, dt\Big\}^2 dx\, du + O(h^5) \\ &= h^3 \left\{\int \left(\frac{f(x)}{1-H(x)}\right)^2 dx\right\}\left\{\int \left(\int W(t)W(u+t)\, dt\right)^2 du\right\} \\ &\quad + O(h^5)\end{aligned}$$

and hence the proof of (36) is completed. In proving (37), first note that

$$\mathbb{E}H_n^4(Y_i, Y_j) = \iiiint (\mathbb{E}R_i(x)R_i(y)R_i(z)R_i(u))^2 \, dx\, dy\, dz\, du.$$

The product of the above summand consists of several terms, each one being of order h^5. As an illustration, the first term is

$$\iiiint \left(\mathbb{E}w_i(x)w_i(y)w_i(z)w_i(u)\right)^2 \, dx\, dy\, dz\, du. \tag{45}$$

Now,

$$\begin{aligned}
&\mathbb{E}w_i(x)w_i(y)w_i(z)w_i(u)\\
&\quad = \int W\left(\frac{x-v}{h}\right) W\left(\frac{y-v}{h}\right) W\left(\frac{z-v}{h}\right) W\left(\frac{u-v}{h}\right) \frac{f(v)}{(1-H(v))^3}\, dv\\
&\quad = h\int W(t) W\left(\frac{y-v}{h}\right) W\left(\frac{z-v}{h}\right) W\left(\frac{u-v}{h}\right) \frac{f(x-th)}{(1-H(x-th))^3}\, dt
\end{aligned} \tag{46}$$

after applying the change of variable $x - v = ht$. In view of (46), (45) becomes

$$\begin{aligned}
h^2 \iiiint \Big\{ \int W(t) W\left(\frac{y-v}{h}\right) W\left(\frac{z-v}{h}\right) W\left(\frac{u-v}{h}\right)\\
\times \frac{f(x-th)}{(1-H(x-th))^3}\, dt\Big\}^2 \, dx\, dy\, dz\, du.
\end{aligned} \tag{47}$$

Set $y - v = wh$, $z - v = sh$ and $u - v = rh$. Then, in view of (47), (45) becomes

$$\begin{aligned}
h^5 \iiiint \Big\{ \int W(t)W(w+t)W(s+t)W(r+t) \frac{f(x-th)}{(1-H(x-th))^3}\, dt\Big\}^2\\
\times\, dx\, ds\, dw\, dr\\
\le h^5 \left\{ \int \left(\frac{f(x)}{(1-H(x))^3}\right)^2 dx\right\}\\
\times \iiiint \Big\{W(t)W(w+t)W(s+t)W(r+t)\Big\}^2 \, dt\, ds\, dw\, dr = O(h^5).
\end{aligned}$$

The rest terms are treated in exactly the same way which concludes in proving (37). □

Lemma 6. *For i, j, k, l, r fixed and all different,*

$$\mathbb{E}(H_n(Y_i, Y_j)H_n(Y_i, Y_k) = 0 \tag{48}$$

$$\mathbb{E}(H_n(Y_i, Y_j)H_n(Y_i, Y_k)H_n(Y_i, Y_l), H_n(Y_i, Y_r)) = 0 \tag{49}$$

$$\mathbb{E}H_n(Y_i, Y_j)H_n^3(Y_i, Y_k) = 0. \tag{50}$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}(H_n(Y_i,Y_j)H_n(Y_i,Y_k))\\
&\quad=\mathbb{E}\left(\int R_i(x)R_j(x)\,dx\right)\left(\int R_i(x)R_k(x)\,dx\right)\\
&\quad=\mathbb{E}\iint R_i(x)R_j(x)\int R_i(y)R_k(y)\,dx\,dy\\
&\quad=\mathbb{E}\iint(w_i(x)-\mathbb{E}w_i(x))(w_j(x)-\mathbb{E}w_j(x))(w_i(y)\\
&\qquad-\mathbb{E}w_i(y))(w_k(y)-\mathbb{E}w_k(y))\,dx\,dy\\
&\mathbb{E}\iint(w_i(x)w_j(x)-w_i(x)\mathbb{E}w_j(x)-w_j(x)\mathbb{E}w_i(x)+\mathbb{E}w_i(x)\mathbb{E}w_j(x))\\
&\qquad\times(w_i(y)w_k(y)-w_i(y)\mathbb{E}w_k(y)-w_k(y)\mathbb{E}w_i(y)+\mathbb{E}w_i(y)\mathbb{E}w_k(y))\,dx\,dy\\
&\quad=\mathbb{E}\iint(w_i(x)w_j(x)w_i(y)w_k(y)-w_i(x)w_j(x)w_i(y)\mathbb{E}w_k(y)\\
&\qquad-w_i(x)w_j(x)w_k(y)\mathbb{E}w_i(y)+w_i(x)w_j(x)\mathbb{E}w_i(y)\mathbb{E}w_k(y)\\
&\qquad-w_i(x)w_i(y)w_k(y)\mathbb{E}w_j(x)+w_i(x)w_i(y)\mathbb{E}w_j(x)\mathbb{E}w_k(y)\\
&\qquad-w_i(x)\mathbb{E}w_j(x)w_k(y)\mathbb{E}w_i(y)-w_i(x)\mathbb{E}w_j(x)\mathbb{E}w_i(y)\mathbb{E}w_k(y)\\
&\qquad+w_i(y)w_k(y)\mathbb{E}w_i(x)\mathbb{E}w_j(x)-w_i(y)\mathbb{E}w_k(y)\mathbb{E}w_i(x)\mathbb{E}w_j(x)\\
&\qquad-w_k(y)\mathbb{E}w_i(y)\mathbb{E}w_i(x)\mathbb{E}w_j(x)+\mathbb{E}w_i(y)\mathbb{E}w_k(y)\mathbb{E}w_i(x)\mathbb{E}w_j(x))\,dx\,dy
\end{aligned}$$

Now, the identity

$$\mathbb{E}\prod_{i=1}^{4}X_i=\prod_{i=1}^{4}\mathbb{E}X_i$$

for independent X_i's is a standard result. Since $i\neq j\neq k$ and since the $Y_i, i=1,\dots,n$ are i.i.d., applying the identity in the integrand of the above double integral gives the result. Verification of (49) and (50) is entirely similar (although longer). □

Acknowledgements

The authors would like to thank the Associate Editor and the reviewers for their helpful comments and suggestions which leaded to improving this research.

Supplementary Material

Supplement to "A goodness of fit test for the survival function under random right censoring"
(doi: 10.1214/13-EJS853SUPP; .zip).

References

[1] Adamidis, I. and Loukas, S. (1998). A lifetime distribution with decreasing failure rate. *Statist Probab. Letters* **39** 35–42. MR1649319

[2] Bagkavos, D. and Ioannides, D. (2012). Smooth confidence intervals for the survival function under random right censoring. *Electron. J. Stat.* **6** 843–860. MR2988431

[3] Bagkavos, D., Ioannides, D. and Kalamatianou, A. (2013). Supplement to "A goodness of fit test for the survival function under random right censoring". DOI: 10.1214/13-EJS853SUPP.

[4] Berg, A. and Politis, D. (2009). Cdf and survival function estimation with infinite-order kernels. *EElectron. J. Stat.* **3** 1436–1454. MR2578832

[5] Cordeiro, G. M. and Lemonte, A. J. (2011). The β–Birnbaum–Saunders distribution: An improved distribution for fatigue life modeling. *Comp. Statist. and Data Analys.* **55** 1445–1461. MR2741426

[6] Chacon, J. E., Duong, T. and Wand, M. P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica* **21** 807–840. MR2829857

[7] Dahiya, R. C. and Gurland, J. (1972). Goodness of fit tests for the gamma and exponential distributions. *Technometrics* **14** 791–801.

[8] Del Castillo, J. and Puig, P. (1999). Invariant exponential models applied to reliability theory and survival analysis. *J. Amer. Statist. Assoc.* **94** 522–528. MR1702322

[9] Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Mult. Anal.* **14** 1–16. MR0734096

[10] Hall, P. and Heyde, C. (1980). *Martingale limit theory and its applications*, Academic Press, New York. MR0624435

[11] Hyndman, R. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician* **50** 361–365.

[12] Ioannides, D. (1992). Integrated square error of nonparametric estimators of regression function: The fixed design case. *Statist. Probab. Letters* **15** 85–94. MR1219277

[13] Jones, C. (1993). Simple boundary correction for density estimation kernel. *Statist. Computing* **3** 135–146.

[14] Gleser, L. G. (1989). The gamma distribution as a mixture of exponential distributions. *Amer. Statist.* **43** 115–117. MR1007623

[15] Gulati, S. and Padgett, W. J. (1996). Families of smooth confidence bands for the survival function under the general random censorship model. *Lifetime Data Anal.* **2** 349–362.

[16] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481. MR0093867

[17] Keating, J. P., Glaser, R. E. and Ketchum, N. S. (1990). Testing hypotheses about the shape parameter of a gamma distribution. *Technometrics* **32** 67–82. MR1050281

[18] Kus, C. (2007). A new lifetime distribution. *Comp. Statist. and Data Analys.* **51** 4497–4509. MR2364461

[19] Kim, C., Bae, W., Choi, H. and Park, B. U. (2005). Non-parametric hazard function estimation using the Kaplan-Meier estimator. *J. Nonparam. Statist.* **17** 937–948. MR2192167

[20] Kraus, D. (2009). surv2sample: Two-sample tests for survival analysis. http://cran.r-project.org/src/contrib/Archive/surv2sample/.

[21] Marron, J. S. and Padgett, W. J. (1987). Asymptotically optimal bandwidth selection for kernel density estimators from randomly right censored samples. *Ann. Statist.* **15** 1520–1535. MR0913571

[22] Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics* **5** 375–383.

[23] Tahmasbi, R. and Rezaei, S. (2008). A two-parameter lifetime distribution with decreasing failure rate. *Comp Statist. and Data Analys.* **52** 3889–3901. MR2432214