Electronic Journal of Statistics
Vol. 3 (2009) 1436–1454
ISSN: 1935-7524
DOI: 10.1214/09-EJS396

CDF and survival function estimation with infinite-order kernels

Arthur Berg*

Biostatistics Division
Penn State College of Medicine
Hershey, PA
e-mail: berg@psu.edu

and

Dimitris Politis

Department of Mathematics
University of California, San Diego
La Jolla, CA
e-mail: dpolitis@ucsd.edu

Abstract: A reduced-bias nonparametric estimator of the cumulative distribution function (CDF) and the survival function is proposed using infinite-order kernels. Fourier transform theory on generalized functions is utilized to obtain the improved bias estimates. The new estimators are analyzed in terms of their relative deficiency to the empirical distribution function and Kaplan-Meier estimator, and even improvements in terms of asymptotic relative efficiency (ARE) are present under specified assumptions on the data. The deficiency analysis introduces a deficiency *rate* which provides a continuum between the classical deficiency analysis and an efficiency analysis. Additionally, an automatic bandwidth selection algorithm, specially tailored to the infinite-order kernels, is incorporated into the estimators. In small sample sizes these estimators can significantly improve the estimation of the CDF and survival function as is illustrated through the deficiency analysis and computer simulations.

AMS 2000 subject classifications: Primary 62G05, 62N02, 62N02; secondary 62P10.
Keywords and phrases: Bandwidth, cumulative distribution function, deficiency, infinite-order kernels, nonparametric estimation, survival function.

Received March 2009.

1. Introduction

We consider the problem of estimating the CDF in contexts of independently and identically distributed (iid) data and randomly right-censored data. Indeed, the seminal paper of Kaplan and Meier [12] solves this problem with the product-limit estimator—the nonparametric maximum likelihood estimator of

*Corresponding author.

the CDF—but there is still room for improvement, especially when the sample size is small.

The most obvious drawback of the Kaplan-Meier estimator, like the empirical distribution function (EDF), is its lack of smoothness. Kernel smoothing easily remedies this problem, but also introduces two new issues of choosing the best kernel and bandwidth. Kernel smoothing also improves the estimator mean square error (MSE) performance by decreasing its variance while introducing a slight bias resulting in an overall improvement of the MSE. The MSE improvement, however, is typically only a second-order improvement, since the original estimator's first-order MSE convergence rate already achieves the best-possible $\sqrt{n}$-rate. When the asymptotic relative efficiency (ARE) between the Kaplan-Meier estimator and its smoothed counterpart is one, as is typically the case, a distinction in performance can be measured by considering the asymptotic relative *deficiency*, or just simply the deficiency between the two estimators. The general notion of deficiency and subsequent calculations with the proposed estimators is provided in Section 3 which also illustrates that an actual increase in efficiency can be achieved with the new estimators under certain (rather strong) assumptions of the distribution function.

Higher-order MSE improvement is influenced by the kernel order—the higher the kernel order, the greater the improvement. Therefore the best kernel-based estimators, the ones with smallest asymptotic MSE, are the estimators that use infinite-order kernels. Current methods traditionally invoke second-order kernels [29] and more recent approaches include using a Bézier curve [13] and a hybrid kernel estimator [15], but infinite-order kernel methods allow for the greatest improvement in bias rates without affecting the rates of the variance. The main argument against the use of large-order kernels in *density* estimation is the concern that the estimator may be negative on some intervals when it is known that the true probability density is always nonnegative. This argument, however, is moot in the density estimation context (so also in the CDF estimation context) since the estimator can easily be truncated to zero when it goes negative then renormalized to have a total area of one without affecting the MSE convergence rate. General construction of the infinite-order kernel estimators are introduced in the following section and a compatible bandwidth selection algorithm that adapts to the infinite-order kernels is described in Section 4.

Another pitfall of all kernel estimators of the density is the lack of consistency at boundary points when the support of the density lies in an interval or half-interval. Simple reflection [31] solves this problem in the density estimation context and an analogous fix also exists for CDF estimators. Boundary correction and standardization methods specific to kernel-smoothed CDF estimators are discussed in Section 5.

Simulations with iid and censored data illustrate the effectiveness of the infinite-order kernel estimators coupled with the automatic bandwidth selection algorithm of Section 6. Uniform improvement in MSE over existing estimators is observed in the simulations. Since estimation of the CDF is so fundamental in standard statical analysis, there are many applications of the new estima-

tors beyond just estimating the underlying CDF. Some of these applications are included in the last section on Discussions and Conclusions.

2. Estimation with flat-top kernels

The analysis will be confined to independently and identically distributed (iid) data, but extensions to randomly right censored with possible left truncation can be more generally derived; cf. [3, 30].

Let $X_1, \ldots, X_n$ be independent[1] and identically distributed random vectors in $\mathbb{R}$ with absolutely continuous distribution function F and corresponding probability density function f. Estimation of f with infinite-order kernels was considered in [24] and [3]; here we consider the integration of those estimators in the construction of the CDF estimator.

The traditional estimator of the CDF is the empirical distribution function, or EDF, which is given by

$$\hat{F}(t) = \frac{1}{n}\sum_{j=1}^{n} I(X_i \leq t)$$

where $I(\cdot)$ represents the indicator function. The kernel estimator of the probability density, f, is then given by

$$\hat{f}_h(x) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{t - X_j}{h}\right) d\hat{F}(t) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right).$$

where K is a kernel that integrates to one (but not necessarily nonnegative!) and h is the bandwidth parameter. Specific regularity conditions on K are described in [30, 32]. Here, the kernels K of interest will be Fourier transforms of a "flat-top function" (described below) where all necessary regularity conditions are satisfied [24].

To insure consistency of $\hat{f}_h$, h should satisfy the condition $h \to 0$ as $n \to \infty$ but with $nh \to \infty$. Further conditions are imposed on the asymptotics of h in Corollary 1 to achieve minimal asymptotic mean square error.

The smoothed estimation of the CDF, $\hat{F}_h$, is constructed by integrating $\hat{f}_h$. That is,

$$\hat{F}_h(t) = \int_{-\infty}^{t} \hat{f}_h(x)\, dx = \frac{1}{n}\sum_{j=1}^{n} \bar{K}\left(\frac{t - X_j}{h}\right) \tag{1}$$

where $\bar{K}(t) = \int_{-\infty}^{t} K(x)\, dx$.

The estimator $\hat{F}_h(t)$ is equivalent to the EDF in terms of first-order asymptotic performance, but improvements are achieved in the higher-order terms. The estimator $\hat{F}_h(t)$ effectively smooths the EDF, decreasing its variance at the

[1]The independent assumption can be relaxed under certain stationarity and mixing conditions; see [18, 9].

cost of introducing a slight bias. The variance improvement is uniform across different kernels, affecting only the second-order constant and not the second-order *rate* (refer to equation (2) below); however the additional bias that gets introduced in the smoothing can be minimized significantly by using kernels of large order with infinite-order kernels providing the most benefit. The variance of $\hat{F}_h(t)$, as derived in [17], is given by

$$\operatorname{var}\left[\hat{F}_h(t)\right] = \frac{F(t)[1-F(t)]}{n} - 2f(t)\left(\int u\bar{K}(u)K(u)\,du\right)\frac{h}{n} + o\left(\frac{h}{n}\right). \quad (2)$$

The bandwidth parameter h only enters the variance expression through the second-order term which is negative. So the larger h is, the smaller the variance of $\hat{F}_h(t)$ becomes. However, we will see below in Theorem 1 that the smaller h is, the smaller the bias of $\hat{F}_h(t)$ becomes. Therefore there is an optimal h that strikes a compromise between the bias and variance terms which is presented in Corollary 1 below.

We now construct a family of infinite-order kernels, following [24], that are derived from "flat-top functions". We start with a continuous, real-valued function κ given by

$$\kappa(s) = \begin{cases} 1, & |s| \le c \\ g(|s|), & \text{otherwise} \end{cases} \quad (3)$$

where g is any continuous, square-integrable function that is bounded in absolute value by one and satisfies $g(|c|) = 1$. The region $|s| < c$ is referred to as the "flat-top neighborhood", but in some cases we may wish to relax the requirement to allow $g(s) \approx 1$ when s is close to c. This "effective flat-top neighborhood" is useful when using an infinitely smooth function $\kappa(s)$ as described in [22] and Section 6 below. In general, the choice of c and function $g(s)$ has little impact on the performance of the infinite-order kernel in comparison to utilizing the flat-top neighborhood. However, one would want to avoid the ill-performing discontinuous rectangular kernel due to its Fourier transform (the Sinc function) having large and slowly-decaying side lobes. The trapezoidal function [26] and a class of infinitely-differentiable or super-smooth flat-tops [19] have been shown to be effective in finite sample simulations. The Fourier transform of κ then produces the infinite-order kernel, K, of interest. Specifically,

$$K(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \kappa(s)e^{-isx}\,ds. \quad (4)$$

Equation (4) presents an infinite-order kernel that, like all kernels of order greater than two, takes on negative values, and corrections due to negativity of the kernel are discussed in Section 5 on boundary correction and standardization. Such infinite-order kernels do however retain the property that $\int_{-\infty}^{\infty} K(x)\,dx = 1$ since the integral is equivalent to $\kappa(0)$ via the inverse Fourier transform. Infinite-order kernels have been been utilized in a number of applications including density estimation [5, 25, 28], censored density estimation [3], time series analysis [26, 27], and nonparametric regression [19].

The MSE of $\hat{F}_h(t)$ with an infinite-order kernel K is now computed under various assumptions on the smoothness of the underlying density. Let $\phi(t)$ be the characteristic function corresponding to $f(x)$, i.e.

$$\phi(s) = \int_{-\infty}^{\infty} f(x)e^{isx}\,dx.$$

The following three assumptions quantifies the degree of smoothness of the density $f(x)$ by the rate of decay of its characteristic function.

Assumption $A(p)$**:** There is a $p > 0$ such that $\int_{-\infty}^{\infty} |t|^p\,|\phi(t)| < \infty$.
Assumption B**:** There are positive constants d and D such that $|\phi(t)| \leq De^{-d|t|}$.
Assumption C**:** There is a positive constant b such that $\phi(t) = 0$ when $|t| \geq b$.

Theorem 1. *Let $\hat{F}_h(t)$ be a kernel smoothed estimator of the CDF with an infinite-order kernel derived from a flat-top function.*

(i) Suppose assumption $A(p)$ holds, then

$$\sup_{t\in\mathbb{R}} \left|\text{bias}\left(\hat{F}_h(t)\right)\right| = o\left(h^{p+1}\right).$$

(ii) Suppose assumption B holds, then

$$\sup_{t\in\mathbb{R}} \left|\text{bias}\left(\hat{F}_h(t)\right)\right| = O\big(he^{-d/h}\big) = o\big(e^{-d/h}\big).$$

(iii) Suppose assumption C holds. When $h \leq 1/b$,

$$\sup_{t\in\mathbb{R}} \left|\text{bias}\left(\hat{F}_h(t)\right)\right| = 0.$$

To optimize the amount of smoothing under the MSE criterion—i.e., to optimize the bandwidth h—we choose the bandwidth that allows the squared bias rates to be comparable to the second-order variance rates. The optimal bandwidths are provided in the following corollary.

Corollary 1. *Let $\hat{F}_h(t)$ be as in Theorem 1.*

(i) Suppose assumption $A(p)$ holds. Letting $h \sim an^{-\beta}$ where a is any positive constant and $\beta = (2p+1)^{-1}$ optimizes the tradeoff between the bias and variance of $\hat{F}_h(t)$ and gives

$$\sup_{t\in\mathbb{R}} \left|\text{bias}\left(\hat{F}_h(t)\right)\right| = o\left(n^{-\frac{p+1}{2p+1}}\right).$$

(ii) Suppose assumption B holds. Letting $h \sim a/\log n$ where $a < 2d$ is a constant optimizes the tradeoff between the bias and variance of $\hat{F}_h(t)$ and gives

$$\sup_{t\in\mathbb{R}} \left|\text{bias}\left(\hat{F}_h(t)\right)\right| = o\left(\frac{1}{\sqrt{n}\log n}\right).$$

(iii) Suppose assumption C holds. Letting $h \leq 1/b$ be fixed guarantees zero bias and the best possible variance rate.

Estimation of the survival function with randomly right censored data can be similarly improved with the smoothing of the Kaplan-Meier estimator with infinite-order kernels. Density estimation of censored data with infinite-order kernels is analyzed in [3], and an estimator of the survival function can be similarly derived from this density estimator through integration as in (1). The same conclusions as Theorem 1 and Corollary 1 will also hold for the smoothed version of the Kaplan-Meier estimator with infinite-order kernels. This is detailed in the following theorem where the proof has been omitted as it follows naturally from the iid case above.

Define $\hat{S}_h(t)$ to be a smoothed estimator of the survival function, $S(t) = 1 - F(t)$, derived from smoothing the Kaplan-Meier estimator with an infinite-order kernel of the form given in (4); i.e.,

$$\hat{S}_h(t) = \sum s_j \bar{K}\left(\frac{t - X_j}{h}\right) \tag{5}$$

where s_j is the height of the jump of the Kaplan-Meier estimator at X_j (cf. [3] for more details). The following theorem is consistent with the results described in [16].

Theorem 2. *Let $\hat{S}_h(t)$ be a kernel smoothed estimator of the survival function as in (5) above. Suppose assumption $A(p)$ holds, then*

$$\sup_{t\in\mathbb{R}} \left|\text{bias}\left(\hat{S}_h(t)\right)\right| = o\left(h^{p+1}\right) = o\left(n^{-\frac{p+1}{2p+1}}\right)$$

when $h \sim an^{-\beta}$ where a is any positive constant and $\beta = (2p+1)^{-1}$.

The analysis under assumptions B and C of the above theorem are considerably more complex and have been omitted.

3. Deficiency

The notion of deficiency was introduced in the article "Deficiency" by Hodges and Lehmann [11] wherein several deficiency calculations are provided. Many articles followed suit using the deficiency concept to compare kernel-smoothed estimators, but many of the approaches used in calculating the deficiency strayed from the original and simple techniques employed by Hodges and Lehmann; c.f. [1, 6, 7, 8, 29, 33]. The simplicity of the original deficiency computations is maintained in the proof of Theorem 3 below.

The deficiency concept is described as follows. Given an estimator, S_m, based on a sample of size m and a more efficient estimator, T_n, based on a sample of size n with equivalent performance as S_m. The difference between the sample sizes, $d = m - n$, defines the relative deficiency between the two estimators. The original paper of Hodges and Lehmann mostly dealt with situations where

d approaches a finite limit as n goes to infinity in which case the two estimators have an asymptotic relative *efficiency* (ARE) of one. However, it is still possible for two estimators to have an ARE of one yet with a deficiency that approaches infinity. Therefore calculation of the *rate* in which d approaches infinity gives a generalization of the original deficiency concept.

In the following theorem, a formula is derived for computing the generalized deficiency between two estimators from their MSE performance which explicitly computes the rate at which d approaches infinity.

Theorem 3. *Suppose the mean squared errors of two estimators S_n and T_n are given as*

$$MSE(S_n) = \frac{c}{n^r} + \frac{a}{n^{r+\delta}} + o\left(\frac{1}{n^{r+\delta}}\right)$$
$$MSE(T_n) = \frac{c}{n^r} + \frac{b}{n^{r+\delta}} + o\left(\frac{1}{n^{r+\delta}}\right)$$

Define $m = m(n)$ to be the sample size for which $MSE(T_m)$ equals (up to a second order term) $MSE(S_n)$. Then the asymptotic deficiency of T_n relative to S_n is $d = m - n$ and satisfies

$$\frac{d}{n^{1-\delta}} \longrightarrow \frac{b-a}{cr}$$

In the next theorem, the deficiency of two estimators is calculated when the second-order term in the MSE expansion decreases at the rate $n^r \log n$ which is very close to the leading term of n^r. Therefore the deficient index, d, will approach infinity at a faster rate indicating a larger discrepancy in the performance of the two estimators.

Theorem 4. *Suppose the mean squared errors of two estimators S_n and T_n are given as*

$$MSE(S_n) = \frac{c}{n^r} + \frac{a}{n^r \log n} + o\left(\frac{1}{n^r \log n}\right)$$
$$MSE(T_n) = \frac{c}{n^r} + \frac{b}{n^r \log n} + o\left(\frac{1}{n^r \log n}\right)$$

Define $m = m(n)$ to be the sample size for which $MSE(T_m)$ equals (up to a second order term) $MSE(S_n)$. Then the asymptotic expected deficiency of T_n relative to S_n is $d = m - n$ and satisfies

$$d\left(\frac{\log n}{n}\right) \longrightarrow \frac{b-a}{cr}$$

These formulas, combined with the results of Corollary 1 and equation (2), are used to derive the deficiency of infinite-order kernel estimators to the unsmoothed EDF under the assumptions $A(p)$, B, and C. In the case of assumption C, the improvement in MSE performance is first-order, and therefore improvement in terms of efficiency, or ARE, is present.

Corollary 2. *Let $\hat{F}_h(t)$ be as in Theorem 1 and $\hat{F}(t)$ be the empirical distribution function estimator. Assume $F(t)\,(1-F(t)) \neq 0$.*

(i) Suppose assumption $A(p)$ holds. When $h \sim an^{-\beta}$ where $a > 0$ is constant and $\beta = (2p+1)^{-1}$, the deficiency of $\hat{F}_h(t)$ relative to $\hat{F}(t)$ is

$$\left(\frac{2af(t)\left(\int u\bar{K}(u)K(u)\,du\right)}{F(t)\,(1-F(t))}\right) n^{\frac{2p}{2p+1}}$$

(ii) Suppose assumption B holds. When $h \sim a/\log n$ where $a < 2d$ is a constant, the deficiency of $\hat{F}_h(t)$ relative to $\hat{F}(t)$ is

$$\left(\frac{2af(t)\left(\int u\bar{K}(u)K(u)\,du\right)}{F(t)\,(1-F(t))}\right) \frac{n}{\log n}$$

(iii) Suppose assumption C holds. When $h \leq 1/b$ is constant, the deficiency of $\hat{F}_h(t)$ relative to $\hat{F}(t)$ is

$$\left(\frac{2f(t)\left(\int u\bar{K}(u)K(u)\,du\right)}{F(t)\,(1-F(t))}\right) n.$$

These deficiency rates indicate, asymptotically, the effective increase in sample size when smoothing $\hat{F}(t)$ with infinite-order kernels. In particular, under the very strong assumption C that requires the characteristic function of the underlying distribution to be compactly supported, there is a deficiency rate of order n indicating an improvement in actual efficiency. This is because utilizing infinite-order kernels can retain retain zero bias under such strong assumptions of the underlying density [28, 3] while the variance of the estimate improves. Of course, such an assumption of the underlying density is unreasonably strong, but near near-efficiency improvements, or a deficiency rate of $n/\log n$, is seen under assumption B that only requires the characteristic function to decay at an exponential rate. A number of parametric densities satisfy this assumption indicating a strong advantage in utilizing infinite-order kernels when smoothing the empirical distribution function.

4. Bandwidth selection

We now present a simple bandwidth selection algorithm that requires very minimal computation and adapts to the specialized family of infinite-order kernels that is utilized in this paper. The methods suggested in [21] for iid data and in [3] for censored data present an algorithm that automatically selects the optimal bandwidth in *density* estimation. Remarkably, these same algorithms can also be used to select the best bandwidth in CDF estimation. Although the bias in estimating the CDF is smaller than the bias of the density estimators, the variance of the CDF estimator is also smaller than the variance of the density estimator. This algorithm automatically adapts to the appropriate assumption

$A(p)$, B, or C and generates a bandwidth that is consistent for the ideal bandwidth given by Corollary 1. The algorithm is also computationally light as well as being simple to describe, and we now proceed to describe it.

Let $\hat{\phi}$ be the natural estimate of the characteristic function given by

$$\hat{\phi}(t) = \int_{-\infty}^{\infty} e^{itx}\, d\hat{F}(x) = \frac{1}{n}\sum_{j=1}^{n} e^{itX_j}.$$

In the context of censored data, $\hat{F}(x)$ in the above expression is replaced with the Kaplan-Meier estimator of the CDF. The main key to the algorithm is finding when $\phi(t) \approx 0$; more specifically, determining the smallest value t^* such that $\phi(t) \approx 0$ for all $t \in (t^*, t^* + \varepsilon)$ for some pre-specified ε. Then the estimate of the bandwidth is given by $\hat{h} = 1/t^*$. The formal algorithm is presented below.

Bandwidth Selection Algorithm
Let $C > 0$ be a fixed constant, and ε_n be a nondecreasing sequence of positive real numbers tending to infinity such that $\varepsilon_n = o(\log n)$. Let t^* be the smallest number such that

$$|\hat{\phi}(t)| < C\sqrt{\frac{\log_{10} n}{n}} \qquad \text{for all } t \in (t^*, t^* + \varepsilon_n) \tag{6}$$

Then let $\hat{h} = c/t^*$ where c is the "flat-top radius" depicted in equation (3).

The positive constant C is irrelevant in the asymptotic theory, but is relevant for finite-sample calculations. The central idea in this algorithm is determining the smallest t such that $\phi(t) \approx 0$. In most cases this can be visually seen without explicitly computing the threshold in (6).

5. Boundary correction and standardization

Vanilla versions of the kernel estimators for *density* estimation break down when the support of the density is restricted to a subset of the real line. For instance, in estimating the probability density function of data taken from an exponential distribution, most kernel estimators give substantial area to negative values even when it is known that the support of the density is nonnegative. It is not too difficult to see that simple kernel estimators of the density will not be consistent at the boundary of the density's support; cf. [31]. However, a simple remedy by reflection works well when the support is not too complex. For instance when the support of the density is $[a, \infty)$, then the estimator

$$\hat{\hat{f}}_h(x) = \left(\hat{f}_h(x) + \hat{f}_h(2a - x)\right) 1_{[a,\infty)(x)} \tag{7}$$

is consistent at the boundary point a ([31]).

This problem, therefore, also carries over to the situation of estimating the CDF. Indeed the EDF and Kaplan-Meier estimators do not suffer from this drawback, but the kernel smoothed versions do. By integrating (7), we deduce a boundary-corrected version of the kernel-smoothed CDF estimator with the

same formulation as (7). For $t \in [a, \infty)$,

$$\begin{aligned}\hat{\hat{F}}_h(t) &= \int_{-\infty}^{t} \hat{\hat{f}}_h(x)\,dx \\ &= \int_{a}^{t} \left(\hat{f}_h(x) + \hat{f}_h(2a-x)\right)\,dx \\ &= \hat{F}_h(t) - \hat{F}_h(a) + \int_{2a-t}^{a} \hat{f}_h(x)\,dx \\ &= \hat{F}_h(t) - \hat{F}_h(a) + \hat{F}_h(a) - \hat{F}_h(2a-t) \\ &= \hat{F}_h(t) - \hat{F}_h(2a-t)\end{aligned}$$

and $\hat{F}_h(t) = 0$ when $t < a$. In the special case $a = 0$, we have the simple formula

$$\hat{\hat{F}}_h(t) = \left(\hat{F}_h(t) - \hat{F}_h(-t)\right) 1_{[0,\infty)(t)}$$

There is an additional issue that only affects higher-order kernel estimators and not second-order estimators. Specifically, higher-order kernel estimators of the density are not necessarily nonnegative, which means higher-order kernels estimators of the CDF are not necessarily contained within the range $[0, 1]$ or forced to be nondecreasing. The natural remedy for these density estimators is to truncate negative estimates to zero and then renormalize the area to one. When this is performed, the corresponding CDF estimator will be a valid CDF. However this approach causes the kernel estimator of the CDF to lose its simplistic representation that is given in the right-hand side of (1), so instead, a simple alternative standardization technique is suggested. To insure the estimator is nondecreasing, $\hat{F}_h(t)$ is replaced by $\sup_{(-\infty,t)} \hat{F}_h(t)$, and to insure the range is between 0 and 1, $\max(\hat{F}_h(t), 1)$ and $\min(\hat{F}_h(t), 0)$ are invoked.

Replacing $\hat{F}_h(t)$ with $\sup_{(-\infty,t)} \hat{F}_h(t)$ is equivalent to replacing the estimator of the density $\hat{f}_h(x)$ with the truncated version $\hat{f}_h^+(x) = \max(\hat{f}_h(x), 0)$ and then integrating the truncated density estimator from $-\infty$ to t. Since $\hat{f}_h^+(x)$ has better MSE performance than the nontruncated counterpart $\hat{f}_h(x)$ [23], it follows that the nondecreasing estimator $\sup_{(-\infty,t)} \hat{F}_h(t)$ has better MSE performance than the original $\hat{F}_h(t)$. Similarly, the MSE of the range restricted estimator produced from $\max(\hat{F}_h(t), 1)$ and $\min(\hat{F}_h(t), 0)$ will also be improved since it is known the CDF has a range bounded in [0,1]. This is formalized in the following corollary.

Corollary 3. *Let $\hat{F}_h(t)$ be as in Theorem 1. A modified estimator is defined as*

$$\tilde{F}_h(t) = \min\left(\max\left(\sup_{(-\infty,t]} \hat{F}_h(t), 0\right), 1\right).$$

Then it follows that

$$\mathrm{MSE}\left(\tilde{F}_h(t)\right) \le \mathrm{MSE}\left(\hat{F}_h(t)\right)$$

and $\tilde{F}_h(t)$ satisfies the necessary properties of a CDF.

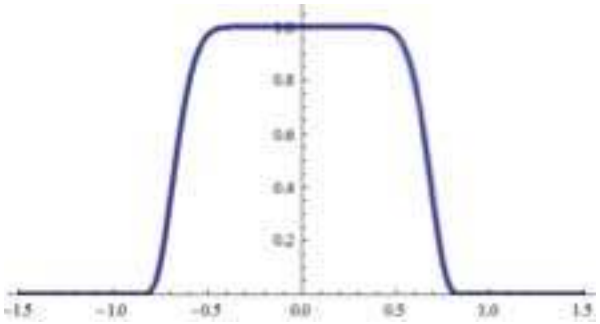

FIG 1. *Infinitely differentiable flat-top function* (8) *with parameters* $b = 1$ *and* $c = .05$.

6. Simulations

We evaluate the performance of the proposed infinite-order kernel estimators with the more traditional second-order kernel estimators and the EDF/Kaplan-Meier estimator. Boundary correction, as described in Section 5, is applied to the estimators when appropriate. As any choice of function $g(x)$ in (3) will insure the ideal asymptotics of an infinite-order kernel, the selection of infinite-order kernels is quite large. An easy choice for the function $g(x)$ is the straight line truncated at zero, i.e. $g(x) = (\frac{1-x}{1-c})^+, x \in [c, 1]$, yielding a trapezoidal shape for κ. The simulations below invoke this trapezoidal function, κ, with parameter $c = .75$.

By making the flat-top function $\kappa(x)$ infinitely smooth, the resulting kernel via the Fourier transform will have tails that decay exponentially. Therefore in situations in estimating the density with boundary conditions, the kernel derived from the infinitely smooth flat-top function is more close to having the desirable quality of being compactly supported than the kernel which is derived from the trapezoidal function. One example of an infinitely smooth $\kappa(x)$ is [19]

$$\kappa(s) = \begin{cases} 1 & \text{if } |s| < c \\ \exp\left(\frac{-b\exp\left(\frac{-b}{(|x|-c)^2}\right)}{(|x|-1)^2}\right) & \text{if } c < |x| < 1 \\ 0 & \text{if } |x| \geq 1 \end{cases} \tag{8}$$

which resembles and infinitely smooth trapezoid and is controlled by the two parameters b and c. In the simulations, we also used this function κ for comparisons with the parameters $b = 1$ and $c = .05$. A plot of this κ is given below.

This function is perfectly flat only from 0 to .05, but it is "effectively" flat from 0 to about .5. Therefore the effective flat-top radius is taken to be .5, and it is this value that is used in the bandwidth selection algorithm described above in Section 4.

A slightly modified bandwidth selection algorithm was invoked that retains the function of the bandwidth algorithm described above. The key in the bandwidth algorithm is to find the smallest value of t^* so that $\hat{\phi}(t^*) \approx 0$. To automate

TABLE 1
*Comparison of the EDF with a Gaussian kernel estimator and two infinite-order kernel estimators (trapezoid and smoothed trapezoid) on iid normal data**

	$t=-1.5$		$t=0$		$t=1.5$	
n	15	30	15	30	15	30
MSE_{EDF}	4.30	2.09	16.29	8.73	4.42	2.14
$\text{MSE}_{\text{Gauss}}$	3.50	1.75	13.02	7.20	3.67	1.82
MSE_{trap}	**2.85**	**1.48**	**11.72**	**6.49**	**2.93**	**1.63**
$\text{MSE}_{\text{smooth}}$	2.95	1.55	12.01	6.71	3.06	1.69

*MSE values are blown up by 10^3 for easier comparison.

TABLE 2
*Comparison of the EDF with a Gaussian kernel estimator and two infinite-order kernel estimators (trapezoid and smoothed trapezoid) on censored Weibull data**

	$t=.75$		$t=1.25$		$t=1.75$	
n	15	30	15	30	15	30
MSE_{EDF}	6.47	3.51	17.0	7.75	12.0	5.62
$\text{MSE}_{\text{Gauss}}$	5.45	2.84	10.1	5.27	8.56	4.11
MSE_{trap}	5.83	2.70	**8.68**	**4.28**	**9.32**	**4.06**
$\text{MSE}_{\text{smooth}}$	**5.04**	**2.36**	9.81	4.85	8.84	5.62

*MSE values are blown up by 10^3 for easier comparison.

this procedure, the value t^* was chosen to be the first value for which $\hat{\phi}(t^*)$ starts to level off.

A Gaussian kernel is used in the second-order kernel estimator, and cross validation, as suggested in [4], is used to select the bandwidth for this estimator. Estimates were simulated over 1000 realizations.

The first simulation study considers the estimation of a $N(0,1)$ CDF from iid data. One may imagine the second-order Gaussian kernel estimator to do quite well in this context, but in fact the infinite-order kernel performs consistently better over the selected points. MSE estimates are provided at three points $(t=-1.5,0,1.5)$ and under two different sample sizes $(n=15,30)$.

The second simulation study considers the estimation of a Weibull distribution with censored data. Lifetime variables, the variables of interest, are simulated from a Weibull distribution with shape parameter 3 and scale parameter 1.5 and the censoring variables are independently drawn from a Weibull distribution with shape parameter 4 and scale parameter 3. Since the support of the lifetime density is on the positive real line, the boundary correction of Section 7 is implemented. MSE estimates are provided at three points $(t=.75,1.5,1.5)$ and under two different sample sizes $(n=15,30)$. Here again the infinite-order kernels consistently outperform the second-order kernel estimator and the Kaplain-Meier estimator in term of MSE performance.

7. Discussion and conclusions

The proposed estimators have implications far beyond just providing a more accurate estimators of the CDF and survival function. For instance, it is standard practice to compare the effects of two drugs based on their respected survival

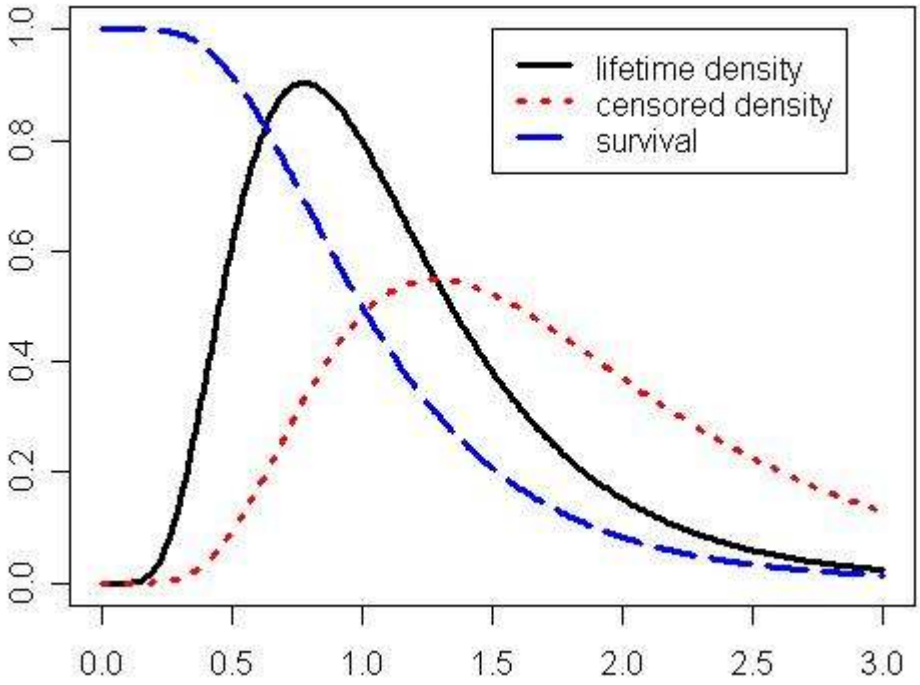

FIG 2. *Lifetime and censored Weibull densities considered in the simulations with a plot of the survival function also included.*

functions, but the cost of running clinical trials limits the sample size of the available data. From the deficiency calculations of Section 3, we see that the proposed estimators can produce the same results as the traditional Kaplan-Meier estimator yet with a significantly smaller sample size.

Another very standard use of the EDF is found in the bootstrap method. In the smoothed bootstrap, data is drawn from a smoothed EDF, and when the estimator of the smoothed EDF is improved, the smoothed bootstrap is also improved to give more accurate inferences [10, 20]. The bootstrap method is particularly beneficial when sample sizes are small, and therefore invoking infinite-order kernel estimators in this situation is often very natural in which having an improved CDF estimator may be crucial.

Hazard function estimation on small samples can also be significantly be improved. Hazard estimators, constructed from dividing a smoothed density estimate by a smoothed survival function, as in [14], have performance that is typically dictated by the convergence of the density estimator [3]. However in small sample sizes, accurate estimation of the survival function is just as crucial as accurate estimation of the density.

The new infinite-order kernel estimators of the CDF and survival function is shown through analysis and demonstrated through simulations to be more accu-

rate than the EDF and Kaplain-Meier estimators with significant improvements seen in small sample sizes and data from a distribution that has a rapidly decaying characteristic function. Significant improvements in terms of an increase in efficiency is also produced by the new estimators when the characteristic function of the data is identically zero after some finite value. Additionally, the bandwidth selection algorithm that accompanies the new estimator is computationally simpler with faster convergence rates than the cross-validation bandwidth selection algorithms used with finite-order kernels.

Appendix A: Technical proofs

Proof of Theorem 1.

From the following computation

$$\mathrm{E}\left[\hat{F}_h(t)\right] = \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}\left[\bar{K}\left(\frac{t-X_i}{h}\right)\right],$$

computing the bias of $\hat{F}_h(t)$ amounts to computing the bias of $\bar{K}\left(\frac{t-X_i}{h}\right)$. Starting with its expectation, we have

$$\begin{aligned}
\mathrm{E}\left[\bar{K}\left(\frac{t-X_i}{h}\right)\right] &= \int_{-\infty}^{\infty} \bar{K}\left(\frac{t-x}{h}\right) f(x)\,dx \\
&= \int_{-\infty}^{\infty} \bar{K}\left(\frac{t-x}{h}\right) dF(x) \\
&= \underbrace{\bar{K}\left(\frac{t-x}{h}\right) F(x)\Big|_{x=-\infty}^{x=\infty}}_{=0} + \frac{1}{h}\int F(x) K\left(\frac{t-x}{h}\right) dx \\
&= \frac{1}{h}\int F(x) K\left(\frac{t-x}{h}\right) dx.
\end{aligned}$$

If we define $K_h(t) = \frac{1}{h}K\left(\frac{t}{h}\right)$, then the expectation above can be written in very simply as

$$\mathrm{E}\left[\bar{K}\left(\frac{t-X_i}{h}\right)\right] = F \star K_h(t)$$

where $\star$ denotes convolution.

To proceed, we will employ Fourier transform theory on (mathematical) distributions, otherwise known as generalized functions. By invoking generalized functions, we can compute the Fourier transform of not just the standard class of integrable functions, but also many non-integrable functions like constants and cumulative distribution functions. This theory, in general, is very technical and readers unfamiliar with the subject are referred to [2] for a nice treatment of the subject.

As K is the Fourier transform of κ, κ is therefore the inverse Fourier transform of K. Through a simple change of variables, we have

$$\mathcal{F}^{-1}\left(K_h(t)\right) = \kappa(th)$$

where the notation $\mathcal{F}$ and $\mathcal{F}^{-1}$ will represent the Fourier transform and its inverse.

Next we wish to derive the Fourier transform of the CDF $F(t)$. This is the first generalized function that we encounter and its Fourier transform involves the Dirac delta function, $\delta(s)$. Using the Heaviside step function $H(x)$ given by $H(x) = 1(x > 0)$, we rewrite $F(t)$ as

$$F(t) = \int_{-\infty}^{t} f(x)\,dx = \int_{-\infty}^{\infty} f(x)H(t-x)\,dx = f \star H(t)$$

Therefore the Fourier transform of $F(t)$ reduces to the product of the Fourier transforms of $f(x)$ and $H(x)$; i.e.

$$\begin{aligned}
\mathcal{F}\left(F(t)\right) &= \phi(s)\left(\pi\delta(s) + \frac{1}{is}\right) \\
&= \pi\phi(0)\delta(s) + \frac{\phi(s)}{is} \\
&= \pi\delta(s) + \frac{\phi(s)}{is}.
\end{aligned}$$

We will now proceed with estimating the bias of $\hat{F}_h(t)$.

$$\begin{aligned}
\text{bias}\left(\hat{F}_h(t)\right) &= K_h \star F(t) - F(t) \\
&= \mathcal{F}\left(\mathcal{F}^{-1}\left(K_h \star F(t) - F(t)\right)\right) \\
&= \mathcal{F}\left(\mathcal{F}^{-1}\left(K_h\right)\cdot\mathcal{F}^{-1}\left(F\right) - \mathcal{F}^{-1}\left(F\right)\right) \\
&= \mathcal{F}\left(\left(\mathcal{F}^{-1}\left(K_h\right) - 1\right)\mathcal{F}^{-1}\left(F\right)\right) \\
&= \mathcal{F}\left(\left(\kappa(sh) - 1\right)\left(\pi\delta(s) + \frac{\phi(s)}{is}\right)\right) \\
&= \mathcal{F}\left(\left(\kappa(sh) - 1\right)\frac{\phi(s)}{is}\right) - \pi\mathcal{F}\left(\left(\kappa(sh) - 1\right)\delta(s)\right) \\
&= \mathcal{F}\left(\left(\kappa(sh) - 1\right)\frac{\phi(s)}{is}\right) - \underbrace{\pi\mathcal{F}\left(\left(\kappa(sh) - 1\right)\Big|_{s=0}\right)}_{=0} \\
&= \frac{1}{2\pi}\int_{|s|>1/h}\left(\kappa(sh) - 1\right)\frac{\phi(s)}{is}\,ds.
\end{aligned}$$

The last equality comes from the flat-top property of κ function; specifically, $\kappa(sh) = 1$ for $|sh| \leq 1$ implies $\kappa(sh) - 1 = 0$ for $|s| \leq 1/h$. Since κ is bounded by one, we have the following bound on the bias of $\hat{F}_h(t)$,

$$\left|\text{bias}\left(\hat{F}_h(t)\right)\right| \leq \frac{2}{2\pi}\int_{|s|>1/h}\frac{|\phi(s)|}{|s|}\,ds.$$

We now bound the bias under the three assumptions $A(p)$, B, and C. Under assumption $A(p)$, we have

$$\begin{aligned}\int_{|s|>1/h} \frac{|\phi(s)|}{|s|}\,ds &= \int_{|s|>1/h} \frac{|s|^p|\phi(s)|}{|s|^{p+1}}\,ds \\ &\le h^{p+1}\int_{|s|>1/h} |s|^p|\phi(s)|\,ds \\ &= o(h^{p+1}).\end{aligned} \tag{9}$$

Under assumption B,

$$\begin{aligned}\int_{|s|>1/h} \frac{|\phi(s)|}{|s|}\,ds &\le h\int_{|s|>1/h} |\phi(s)|\,ds \\ &\le h\int_{|s|>1/h} De^{-d|s|}\,ds \\ &\le \frac{Dh}{e^{d/h}}\int_{|s|>1/h} e^{d(1/h-|s|)}\,ds \\ &= O\left(he^{-d/h}\right).\end{aligned} \tag{10}$$

And under assumption C,

$$\int_{|s|>1/h} \frac{|\phi(s)|}{|s|}\,ds = 0 \tag{11}$$

when $h \le 1$. Therefore parts (i) through (iii) are proven from equations (9), (10), and (11) respectively. □

Proof of Theorem 3.

If the mean square errors are equal, up to a fraction of the sample size, then we have

$$\frac{c}{n^r} + \frac{a}{n^{r+\delta}} + o\left(\frac{1}{n^{r+\delta}}\right) = \frac{c}{m^r} + \frac{b}{m^{r+\delta}} + o\left(\frac{1}{m^{r+\delta}}\right)$$

which implies

$$\frac{1}{n^r}\left[c + \frac{a+o(1)}{n^\delta}\right] = \frac{1}{m^r}\left[c + \frac{b+o(1)}{m^\delta}\right].$$

Dividing through by c and solving for $\frac{m}{n}$ gives

$$\frac{m}{n} = \left[1 + \frac{b+o(1)}{cn^\delta}\right]^{1/r}\left[1 + \frac{a+o(1)}{cm^\delta}\right]^{-1/r}.$$

From the above expression, we see that $m/n \to 1$ and therefore $o(1/n) = o(1/m)$. Using the approximation $(1+x)^s = 1 + sx + O(x^2)$ gives

$$\frac{m}{n} = 1 + \frac{b}{crn^\delta} - \frac{a}{crm^\delta} + o\left(\frac{1}{n^\delta}\right)$$

Recalling $m = n + d$, we have

$$\frac{d}{n} = \frac{b}{crn^{\delta}} - \frac{a}{crm^{\delta}} + o\left(\frac{1}{n^{\delta}}\right).$$

Multiplying both sides of the above equation by n^{δ} gives

$$\frac{d}{n^{1-\delta}} = \frac{b}{cr} - \frac{a}{cr}\left(\frac{n}{m}\right)^{\delta} + o(1) \longrightarrow \frac{b-a}{cr}.$$

□

Proof of Theorem 4.

The proof of Theorem 4 follows the same lines as the proof of Theorem 3 with n^{δ} replaced with $\log n$. □

References

[1] Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika* **68** 326–328. MR614972 (82f:62094)

[2] Beerends, R. J., ter Morsche, H. G., van den Berg, J. C. and van de Vrie, E. M. (2003). *Fourier and Laplace transforms*. Cambridge University Press, Cambridge. Translated from the 1992 Dutch edition by Beerends. MR2001192 (2004g:42001)

[3] Berg, A. and Politis, D. N. (2009). Density Estimation of Censored Data with Infinite-Order Kernels. *Submitted for publication*.

[4] Bowman, A., Hall, P. and Prvan, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika* **85** 799.

[5] Devroye, L. (1992). A note on the usefulness of superkernels in density estimation. *The Annals of Statistics* **20** 2037–2056.

[6] Falk, M. (1983). Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statist. Neerlandica* **37** 73–83. MR711744 (85d:62040)

[7] Ghorai, J. K. (1989). Deficiency of the MLE of a smooth survival function under the proportional hazard model. *Comm. Statist. Theory Methods* **18** 3047–3056. MR1033148 (91f:62160)

[8] Ghorai, J. K. and Rejtő, L. (1990). Relative deficiency of kernel type estimators of quantiles based on right censored data. *Comm. Statist. Theory Methods* **19** 1653–1670. MR1075495 (91k:62037)

[9] Gyorfi, L., Hardle, W., Sarda, P. and Vieu, P. (1989). Nonparametric Curve Estimation from Time Series. *Lecture Notes in Statistics* **60**.

[10] Hall, P., DiCiccio, T. and Romano, J. (1989). On Smoothing and the Bootstrap. *The Annals of Statistics* **17** 692–704.

[11] Hodges Jr, J. and Lehmann, E. (1970). Deficiency. *The Annals of Mathematical Statistics* **41** 783–801.

[12] Kaplan, E. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53** 457–481.
[13] Kim, C., Park, B., Kim, W. and Lim, C. (2003). Bezier curve smoothing of the Kaplan-Meier estimator. *Annals of the Institute of Statistical Mathematics* **55** 359–367.
[14] Kim, C., Bae, W., Choi, H. and Park, B. U. (2005). Non-parametric hazard function estimation using the Kaplan-Meier estimator. *J. Nonparametr. Stat.* **17** 937–948. MR2192167 (2006g:62033)
[15] Kim, C., Kim, S., Park, M. and Lee, H. (2006). A bias reducing technique in kernel distribution function estimation. *Computational Statistics* **21** 589-601.
[16] Kulasekera, K., Williams, C., Coffin, M. and Manatunga, A. (2001). Smooth estimation of the reliability function. *Lifetime Data Analysis* **7** 415–433.
[17] Li, Q. and Racine, J. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.
[18] Liu, R. and Yang, L. (2007). Kernel estimation of multivariate cumulative distribution function.
[19] McMurry, T. and Politis, D. (2004). Nonparametric regression with infinite order flat-top kernels. *Journal of Nonparametric Statistics* **16** 549–562.
[20] Polansky, A., Schucany, W. and et al. (1997). Kernel Smoothing to Improve Bootstrap Confidence Intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59** 821–838.
[21] Politis, D. N. (2003). Adaptive bandwidth choice. *Journal of Nonparametric Statistics* **15** 517–533.
[22] Politis, D. N. (2005). Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices. *UCSD Economics Discussion Papers*.
[23] Politis, D. N. and Romano, J. P. (1995). Bias-corrected nonparametric spectral estimation. *J. Time Ser. Anal.* **16** 67–103. MR1323618 (95k:62256)
[24] Politis, D. N. and Romano, J. P. (1999). Multivariate density estimation with general flat-top kernels of infinite order. *J. Multivariate Anal.* **68** 1–25. MR1668848 (2000d:62057)
[25] Politis, D. and Romano, J. (1993). On a family of smoothing kernels of infinite order. In Computing Science and Statistics In *Proceedings of the 25th Symposium on the Interface (M. Tarter and M. Lock, Eds.), The Interface Foundation of North America* 141–145.
[26] Politis, D. and Romano, J. (1995). Bias Corrected Nonparametric Spectral Density Estimator. *Journal of Time Series Analysis* **16** 67–103.
[27] Politis, D. and Romano, J. (1996). On flat-top kernel spectral density estimators for homogeneous random fields. *Journal of Statistical Planning and Inference* **51** 41–53.

[28] Politis, D. and Romano, J. (1999). Multivariate density estimation with general flat-top kernels of infinite order. *Journal of Multivariate Analysis* **68** 1–25.
[29] Reiss, R.-D. (1981). Nonparametric estimation of smooth distribution functions. *Scand. J. Statist.* **8** 116–119. MR623587 (82k:62080)
[30] Sánchez-Sellero, C., González-Manteiga, W. and Cao, R. (1999). Bandwidth selection in density estimation with truncated and censored data. *Ann. Inst. Statist. Math.* **51** 51–70. MR1704646 (2000k:62066)
[31] Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC.
[32] Wang, S. (1991). Nonparametric estimation of distribution functions. *Metrika* **38** 259–267.
[33] Xiang, X. (1995). Deficiency of the sample quantile estimator with respect to kernel quantile estimators for censored data. *Ann. Statist.* **23** 836–854. MR1345203 (97c:62092)