

Density Estimation of Censored Data with Infinite-Order Kernels

Arthur Berg Dimitris N. Politis

University of California, San Diego

aberg@ucsd.edu dpolitis@ucsd.edu

Abstract

Higher-order accurate density estimation under random right censorship is achieved using kernel estimators from a family of infinite-order kernels. A compatible bandwidth selection procedure is also proposed that automatically adapts to level of smoothness of the underlying lifetime density. The combination of infinite-order kernels with the new bandwidth selection procedure produces a considerably improved estimate of the lifetime density and hazard function surpassing the performance of competing estimators. Infinite-order estimators are also utilized in a secondary manner as pilot estimators in the plug-in approach for bandwidth choice in second-order kernels. Simulations illustrate the improved accuracy of the proposed estimator against other nonparametric estimators of the density and hazard function.

KEYWORDS: Density estimation, hazard function estimation, infinite-order kernels, nonparametric estimation, survival analysis

1 Introduction

In kernel-based density estimation of uncensored iid data, the usefulness of using infinite-order kernels, or “superkernels”, is well known; cf. [3, 17, 16]. In general, using superkernels reduces bias by orders of magnitude without increasing the order of magnitude of the variance thus producing estimates with better mean square error (MSE) properties. Therefore one would imagine higher-order, if not infinite-order, kernels to be much more popular than second-order kernels; this, however, has not been the case mainly for three reasons that affect both iid density estimation and density estimation of censored data.

The foremost concern with using high-order kernels is the possibility of the estimate being negative at some places when it is known that densities are nonnegative everywhere. The simple fix to this issue by truncating all negative values to zero works well but produces a secondary issue of having an estimate of the pdf that integrates to a value that is less than one. Indeed, the area can be renormalized back to one and still

have the same MSE order. Therefore the complaint that higher-order kernels produce density estimates that are not densities themselves is inconsequential since any lack of nonnegativity can easily be remedied.

Synonymous with the problem of choosing an appropriate kernel in kernel density estimation is the problem of choosing the correct bandwidth. Because second-order kernels have been so popular, several bandwidth selection procedures have been proposed and analyzed for these kernels; refer to [6, 9] for a review of several methods. However, most of the bandwidth procedures for second-order kernels do not carry over to infinite-order kernels, and the methods that do carry over, like cross validation, are known to have poor performance properties [6]. Yet in 2001, Politis [13] showed how a very simple and intuitive bandwidth selection algorithm works well with infinite-order kernels.

Another concern about using infinite-order kernels is not the asymptotic performance which is guaranteed, but rather their finite sample performance. Specifically, in using the high-order kernels, the bias improves at a cost to increasing the variance by some constant factor independent of the sample size. Indeed, there are many poor choices of infinite-order kernels, just as there are many poor choices of second-order kernels. One of the simplest and most popular infinite-order kernels is the sinc function which is a very poor choice due to its large and slowly decaying side lobes. However, a class of favorably performing infinite-order kernels has been proposed in [14, 16] and has been shown to outperform mainstream second-order kernels in finite sample simulations. In addition to being infinite-order, it is also advisable to have a kernel with tails that die off fast; this was also noted by Devroye in [3]. Requiring the kernel to have tails that die off quickly is equivalent to the kernels Fourier transform being very smooth. So the reason the sinc function is such a poor choice of kernel is because its Fourier transform is a rectangle—the most unsmooth flat-top shape. Improvements on the rectangle shape include the trapezoid and the infinitely differentiable flat-top function of McMurry and Politis [11]. This paper adopts the infinite-order kernel that is derived from the trapezoidal shape as it is a simple choice of kernels that works well in practice.

As a simple example to illustrate the effectiveness of the proposed density estimator, we present the results of a simple simulation with uncensored iid data. In the simulation, we estimate the pdf of a $\mathcal{N}(0, 1)$ distribution with datasets of sizes $n = 50$ and $n = 500$ and two estimators—the infinite-order estimator with bandwidth selection procedure described in this paper (see Section 4 for the exact estimator used) and the default density estimator used in R version 4.2.1 (`density`) with a Gaussian kernel and its built-in bandwidth selection procedure. After repeating the simulations over 10,000 realizations, we compute the mean square error at three points ($x = 0, 1, 2$) and on an equally spaced grid of 41 points in the interval $[-2, 2]$. Here are the results:

We see that even using a Gaussian kernel to estimate a Gaussian density is not as good as using the infinite-order kernels with accompanying bandwidth selection procedures that is proposed in Section 3.

In the next section, we define the general class of flat-top infinite-order kernels and, through Theorem 1, describe how using these kernels can cause the bias of the kernel density estimators of censored data to become essentially negligible in certain situa-

n	$x = 0$		$x = 1$		$x = 2$		avg on $[-2,2]$	
	50	500	50	500	50	500	50	500
$\text{MSE}_{\text{infinite}}^*$	2.42	.346	1.86	.299	.973	.115	1.83	.262
$\text{MSE}_{\text{density}}^*$	4.37	.755	2.48	.420	.956	.151	2.58	.439

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

Table 1: Comparison of the proposed infinite-order kernel density estimator with the Gaussian kernel density estimator on iid $\mathcal{N}(0,1)$ data.

tions. Section 3 completes the proposed estimator by providing a bandwidth selection algorithm that automatically adapts to the unknown density at hand. In Section 4 we give practical suggestions for implementing the proposed estimator and provide several simulations exhibiting optimal performance in estimating the lifetime density and hazard function when compared with other nonparametric estimators including the `muhaz` estimator [12] and the logspline estimator [7].

2 The Flat-Top Estimators

We lay out the notation under the context of random right censorship which can be generalized to also allow for left truncation; see for example [18]. Let X_1^0, \dots, X_n^0 be iid lifetime variables with density f and cdf F , and independently, let U_1, \dots, U_n be iid censoring variables with cdf G . We observe the data Z_i and Δ_i where

$$Z_i = \min\{X_i^0, U_i\} \quad \text{and} \quad \Delta_i = 1_{[X_i^0 \leq U_i]} \in \{0, 1\}$$

for $i = 1, \dots, n$ (here $1_{[\cdot]}$ represents the indicator function). We order the pairs (Z_i, Δ_i) according to the Z_i 's and relabel them as (X_i, δ_i) where $X_i = Z_{(i)}$, the i^{th} order statistics of the Z 's, and δ_i is the indicator variable that accompanies X_i , i.e. the concomitant of X_i . The Kaplan-Meier estimator is the nonparametric maximum likelihood estimate of the survival function $S(t) = 1 - F(t)$ given by

$$\hat{S}(t) = \begin{cases} 1, & 0 \leq t \leq X_1 \\ \prod_{j=1}^{k-1} \left(\frac{n-j}{n-j+1} \right)^{\delta_j}, & X_{k-1} < t \leq X_k, \quad k = 2, \dots, n \\ 0, & t > X_n \end{cases}$$

where the height of the jump of \hat{S} at X_j is

$$s_j = \begin{cases} \hat{S}(X_j) - \hat{S}(X_{j+1}), & j = 1, \dots, n-1 \\ \hat{S}(X_n), & j = n. \end{cases}$$

The kernel estimate of f is constructed through the convolution of $\hat{F} = 1 - \hat{S}$ with a smooth kernel K , i.e.

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

Many authors require K to be of compact support for ease of analysis, but this is unnecessary; see for example [22]. Therefore we only assume K is an even function that integrates to one.

It will be assumed that sufficient conditions on the the density f or kernel K are satisfied so that

$$\text{var}\left(\hat{f}(x)\right) = O\left(\frac{1}{nh}\right); \quad (2)$$

some sufficient conditions for (2) are provided in [22].

Following [14], we now describe a class of infinite-order kernels constructed from the Fourier transform of flat-top function. We start in the Fourier domain with a function κ given by

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

where g is any continuous, square-integrable function that is bounded in absolute value by 1 and satisfies $g(\pm 1) = 1$ (g will typically be compactly supported, but this is not required). Then the infinite-order kernel corresponding to κ is the Fourier transform of κ , specifically,

$$K(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \kappa(t) e^{-itx} dt. \quad (4)$$

Let $\phi(t)$ be the characteristic function corresponding to $f(x)$, i.e. $\phi(t)$ is the inverse Fourier transform of $f(x)$ given by

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx.$$

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(r): There is an $r > 0$ such that $\int_{-\infty}^{\infty} |t|^r |\phi(t)| < \infty$.

Assumption B: There are positive constants d and D such that $|\phi(t)| \leq D e^{-d|t|}$.

Assumption C: There is a positive constant b such that $\phi(t) = 0$ when $|t| \geq b$.

Theorem 1. Suppose $\hat{f}(x)$ is the kernel estimator as defined in (1) with infinite-order kernel given by (4) and assume the variance assumption in (2) holds.

(i) Suppose assumption A(r) holds. Let $h \sim an^{-\beta}$ with $\beta = (2r + 1)^{-1}$, then

$$\sup_{x \in \mathbb{R}} \left| \text{bias} \left\{ \hat{f}(x) \right\} \right| = o\left(n^{\frac{-r}{2r+1}}\right) \quad \text{and} \quad \text{MSE} \left\{ \hat{f}(x) \right\} = O\left(n^{\frac{-2r}{2r+1}}\right).$$

(ii) Suppose assumption B holds. Let $h \sim 1/(a \log n)$ where a is a constant such that $a > 1/(2d)$, then

$$\sup_{x \in \mathbb{R}} \left| \text{bias} \left\{ \hat{f}(x) \right\} \right| = O \left(\frac{1}{\sqrt{n}} \right) \quad \text{and} \quad \text{MSE} \left\{ \hat{f}(x) \right\} = O \left(\frac{1}{\sqrt{n}} \right).$$

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

$$\sup_{x \in \mathbb{R}} \left| \text{bias} \left\{ \hat{f}(x) \right\} \right| = 0 \quad \text{and} \quad \text{MSE} \left\{ \hat{f}(x) \right\} = O \left(\frac{1}{n} \right).$$

Corollary 1. The hazard function $H(x) = f(x)/S(x)$ is easily estimated by $\hat{H}(x) = \hat{f}(x)/\hat{S}(x)$, and since \hat{S} is a \sqrt{n} -convergent estimator of S , this estimate of the hazard function has the same MSE convergence rates as \hat{f} in the above theorem. Specifically:

(i) Under assumption A(r), $\text{MSE}(\hat{H}(x)) = O \left(n^{\frac{-2r}{2r+1}} \right)$;

(ii) Under assumption B, $\text{MSE}(\hat{H}(x)) = O \left(\frac{1}{\sqrt{n}} \right)$;

(iii) Under assumption C, $\text{MSE}(\hat{H}(x)) = O \left(\frac{1}{n} \right)$.

The p^{th} derivative of f can be estimated by the the p^{th} derivative of $\hat{f}(x)$; i.e. if $K^{(p)}(x)$ is the p^{th} derivative of $K(x)$, then

$$\hat{f}_p(x) = \frac{1}{h^{p+1}} \sum_{j=1}^n s_j K^{(p)} \left(\frac{x - X_j}{h} \right) \quad (5)$$

is an estimate of the p^{th} derivative of f . It can be shown, under sufficient conditions on f , that the variance of this estimator is

$$\text{var} \left(\hat{f}_p(x) \right) = O \left(\frac{1}{n h^{p+1}} \right). \quad (6)$$

The previous theorem is now be generalized in the following theorem to give asymptotic bias and MSE rates of $\hat{f}_p(x)$ with infinite-order kernels.

Theorem 2. Suppose $\hat{f}_p(x)$ is the kernel estimator as defined in (5) where K is an infinite-order kernel, and assume the variance assumption in (6) holds.

(i) Suppose assumption A($r + p$) holds. Let $h \sim an^{-\beta}$ with $\beta = (2r + p + 1)^{-1}$, then

$$\sup_{x \in \mathbb{R}} \left| \text{bias} \left\{ \hat{f}_p(x) \right\} \right| = o \left(n^{\frac{-r}{2r+p+1}} \right) \quad \text{and} \quad \text{MSE} \left\{ \hat{f}_p(x) \right\} = O \left(n^{\frac{-2r}{2r+p+1}} \right).$$

(ii) Suppose assumption B holds. Let $h \sim 1/(a \log n)$ where a is a constant such that $a > 1/(2d)$, then

$$\sup_{x \in \mathbb{R}} \left| \text{bias} \left\{ \hat{f}_p(x) \right\} \right| = O \left(\frac{1}{\sqrt{n}} \right) \quad \text{and} \quad \text{MSE} \left\{ \hat{f}_p(x) \right\} = O \left(\frac{1}{\sqrt{n}} \right).$$

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

$$\sup_{x \in \mathbb{R}} \left| \text{bias} \left\{ \hat{f}_p(x) \right\} \right| = 0 \quad \text{and} \quad \text{MSE} \left\{ \hat{f}_p(x) \right\} = O \left(\frac{1}{n} \right).$$

In particular, we see that if the underlying density is infinitely smooth (as in the case of assumptions B and C), then the same asymptotic MSE rates of $\hat{f}_p(x)$ hold for every p .

3 Bandwidth Selection for Flat-Top Estimators

Let $\hat{\phi}(t)$ be an estimate of the characteristic function $\phi(t)$ given by

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

We now follow the general recipe for bandwidth selection using flat-top kernels that is detailed in [13]. Specifically, we determine the smallest value t^* such that $\hat{\phi}(t) \approx 0$ for all $t \in (t^*, t^* + \varepsilon)$ for some pre-specified ε , then the estimate of the bandwidth is $\hat{h} = 1/t^*$. The details are provided in the following algorithm.

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

$$|\hat{\phi}(t)| < C \sqrt{\frac{\log_{10} n}{n}} \quad \text{for all } t \in (t^*, t^* + \varepsilon_n) \quad (7)$$

Then let $\hat{h} = 1/t^*$.

Remark 1. The positive constant C is irrelevant in the asymptotic theory, but is relevant for finite-sample calculations. The main idea behind the algorithm is to determine the smallest t such that $\phi(t) \approx 0$; in most cases this can be visually seen without explicitly computing the threshold in (7).

Remark 2. If $g(t)$ in (3) is identically one, or very close to one, in a neighborhood of the type $[1, 1 + \eta]$, then the “flat-top radius” is effectively increased to some value $1 + \eta$. In this case, we would let $\hat{h} = (1 + \eta)/t^*$ in the bandwidth selection algorithm.

Theorem 3. Assume the following two natural assumptions:

$$\max_{s \in (0,1)} |\hat{\phi}(t+s) - \phi(t+s)| = O_P(1/\sqrt{n})$$

uniformly in t , and

$$\max_{s \in (0,n)} |\hat{\phi}(t+s) - \phi(t+s)| = O_P \left(\frac{\log n}{\sqrt{n}} \right)$$

(i) Assume $\phi(t) \sim A|t|^{-d}$ for some positive constants A and d . Then

$$\hat{h} \stackrel{P}{\sim} \tilde{A} \left(\frac{\log n}{n} \right)^{\frac{1}{2d}};$$

here $A \stackrel{P}{\sim} B$ means $A/B \rightarrow 1$ in probability.

(ii) Assume $\phi(t) \sim A\xi^{|t|}$ for some $\xi \in (0, 1)$ and $A > 0$. Then

$$\hat{h} \stackrel{P}{\sim} 1/(\tilde{A} \log n).$$

where $\tilde{A} = -1/\log \xi$.

(iii) Assume $|\phi(t)| = 0$ when $t \geq b$. Then $\hat{h} \stackrel{P}{\sim} 1/b$.

Theorem 3 shows how the proposed bandwidth selection algorithm adapts to the underlying degree of smoothness of the density matching nearly identically to the ideal bandwidths in Theorem 1. When there is only polynomial decay of the characteristic function, as in part (i) of the above theorem, the bandwidth selection algorithm produces a slightly smaller bandwidth than the theoretically optimal bandwidth given in Theorem 1, but the discrepancy diminishes with faster decay.

4 Bandwidth Selection for 2nd-Order Kernels

We now propose a bandwidth selection procedure for use with *second-order* kernels, based on using the infinite-order estimators as pilots in the plug-in approach to bandwidth selection. Indeed, Theorem 1 points out the superiority of using the infinite-order kernels over second order kernels, but as a stepping stone to using infinite-order kernels directly in estimation, we now introduce the infinite-order kernels in pilot estimation. The result is a bandwidth selection procedure that converges very fast (comparable to the rates of Theorem 1).

We begin with the MSE and mean integrated square error (MISE) of $\hat{f}(x)$ with a second order kernel Λ and standard assumptions on Λ [18, 21].

$$\begin{aligned} \text{MSE}(\hat{f}) &= h^4 \cdot \left(\frac{f''(x)}{2} \int_{-\infty}^{\infty} x^2 \Lambda(x) dx \right)^2 \\ &+ \frac{1}{nh} \cdot \frac{f(x)}{1-G(x)} \int_{-\infty}^{\infty} \Lambda^2(x) dx \\ &+ \frac{1}{n} \cdot f(x)^2 \left[\int_{-\infty}^x \frac{f(r)}{1-G(r)} dr - \frac{1}{(1-F(x))(1-G(x))} \right] \\ &+ O(h^6) + O\left(\frac{h}{n}\right) + o\left(\frac{1}{nh}\right) \end{aligned}$$

$$\begin{aligned}
\text{MISE}(\hat{f}) &= h^4 \cdot \int_{-\infty}^{\infty} f''(x)^2 \omega(x) dx \left(\frac{1}{2} \int_{-\infty}^{\infty} x^2 \Lambda(x) dx \right)^2 \\
&+ \frac{1}{nh} \cdot \int_{-\infty}^{\infty} \frac{f(x)}{1-G(x)} \omega(x) dx \int_{-\infty}^{\infty} \Lambda^2(x) dx \\
&+ \frac{1}{n} \cdot \int_{-\infty}^{\infty} \left[\int_{-\infty}^x \frac{f(r)}{1-G(r)} dr - \frac{1}{(1-F(x))(1-G(x))} \right] f(x)^2 \omega(x) dx \\
&+ O(h^6) + O\left(\frac{h}{n}\right) + o\left(\frac{1}{nh}\right)
\end{aligned}$$

The MISE above has been generalized slightly to incorporate a nonnegative weight function $\omega(x)$ to control the influence of error in the tails the estimated density. If we minimize the above MSE with respect to h , we arrive at the optimal bandwidth for estimating the density *at a given point*. And if we minimize the MISE with respect to h , then we arrive at an optimal *global* bandwidth. The optimal bandwidths in each situation will involve values of the unknown underlying density that we wish to estimate, so we are forced to use some initial-or pilot-estimate of these values. Minimizing the above MSE and MISE values leads to the optimal bandwidths h_{MSE} and h_{MISE} , respectively, given below.

$$\begin{aligned}
h_{\text{MSE}} &= \left(\frac{\frac{f(x)}{1-G(x)} \int_{-\infty}^{\infty} \Lambda^2(x) dx}{\left(f''(x) \int_{-\infty}^{\infty} x^2 \Lambda(x) dx \right)^2} \right)^{1/5} n^{-1/5} \\
h_{\text{MISE}} &= \left(\frac{\int_{-\infty}^{\infty} \frac{f(x)}{1-G(x)} \omega(x) dx \int_{-\infty}^{\infty} \Lambda^2(x) dx}{\int_{-\infty}^{\infty} f''(x)^2 \omega(x) dx \left(\int_{-\infty}^{\infty} x^2 \Lambda(x) dx \right)^2} \right)^{1/5} n^{-1/5}
\end{aligned}$$

In the above expression, we shall replace $f(x)$ and $f''(x)$ with infinite-order estimators $\hat{f}(x)$ and $\hat{f}_2(x)$ respectively. The bandwidth used in estimating $\hat{f}_2(x)$, and in general for $\hat{f}_p(x)$, is the same bandwidth derived from the bandwidth selection algorithm above. The function $1 - G(x)$ is the survival function of the *censored* random variables, therefore by replacing Δ_i with $1 - \Delta_i$, the Kaplan-Meier estimator will give a \sqrt{n} -consistent estimate of $1 - G(x)$. Let \hat{h}_{MSE} and \hat{h}_{MISE} refer to the plug-in estimates corresponding to h_{MSE} and h_{MISE} respectively. These estimators have rapid convergence rates due to the ultra-fast convergence of the pilot flat-top estimators, and this is revealed in the following theorem.

Theorem 4. *Assume the conditions of Theorem 3, and assume conditions strong enough to ensure (6) holds for $p = 2$. Let \hat{h}_M be either \hat{h}_{MSE} or \hat{h}_{MISE} with h_M being the corresponding h_{MSE} or h_{MISE} .*

(i) *Assume $\phi(t) \sim A|t|^{-d}$ for some positive constants A and $d > 3$. Then*

$$\hat{h}_M = h_M \left(1 + O_p \left(\frac{\log n}{n} \right)^{\frac{[d-4]}{2d}} \right).$$

(ii) Assume $\phi(t) \sim A\xi^{|t|}$ for some $\xi \in (0, 1)$ and $A > 0$. Then

$$\hat{h}_M = h_M \left(1 + O_p \left(\frac{\log n}{n} \right)^{\frac{1}{2}} \right).$$

(iii) Assume $|\phi(t)| = 0$ when $t \geq b$, then

$$\hat{h}_M = h_M \left(1 + O_p \left(\frac{1}{\sqrt{n}} \right) \right).$$

Marron and Padgett [10] suggest cross-validation as a means of minimizing the integrated square error (ISE), but this approach of minimizing ISE was shown in [4] to be less optimal than minimizing the MISE. In particular, the relative convergence rates (as in the above theorem) of the cross-validation approach in [10] is $n^{-1/10}$, regardless of the degree of smoothness of $f(x)$. If one uses the plug-in approach that we have adopted above but with pilots consisting of second-order kernels, then the relative convergence rates are at best $n^{-2/5}$, again, regardless of the degree of smoothness of $f(x)$. All of these rates are considerably smaller than the $n^{-1/2}$ rate afforded by the proposed procedure under a sufficiently smooth density $f(x)$ (i.e. when $\phi(x)$ has a rapid decay to zero) as Theorem 4 demonstrates.

5 Simulations

We constructed our infinite-order kernel from a “flat-top” function κ in (3) with any choice of continuous, square-integrable function g ; an easy choice is $g(x) = (1 + c - c|x|)^+$, which gives κ a trapezoidal shape. Other possibilities for the function g are considered in [15]. For the following simulations, we focus on a trapezoidal shape for κ ; we are still left to determine the parameter c controlling the slopes on the sides of the trapezoid. The parameter $c = 4$ seemed to work generally well and was used throughout all of the simulations, but there is certainly some flexibility in choice of c ; see [16] for further discussion on choosing the parameter c .

In many situations, particularly involving censored data, the support is known to lie in a half-line or some compact interval, and unaltered versions of kernel density estimators are not even consistent near the boundary points. However there have been many fixes to this boundary issue (see [5] for a survey of several methods), and we adopt the simple reflection principle to resolve boundary problems in our estimator. Specifically, when the density is known to have its support on $[0, \infty)$, we use the estimator $\hat{f}(x) = \hat{f}(x) + \hat{f}(-x)$ to ensure consistency near the boundary point $x = 0$; see [19] and [20] for discussions of this method in the uncensored iid context.

We mimic the simulation presented in the introduction but with censoring. Independent and identically distributed lifetime variables are drawn from a $\mathcal{N}(0, 1)$ distribution and, independently, the censoring variables are drawn from the same distribution. Therefore we would expect to see about 50% censoring on average. The challenger to our infinite-order density estimator is (1) with Gaussian kernel K . The cross-validation criterion is suggested in [10] for selecting the bandwidth for the Gaussian kernel, but one of its drawbacks is the computational time required to compute it

which is greatly magnified over several thousand realizations. So instead of computing the cross-validations, we gave the Gaussian kernel a distinct advantage by choosing the bandwidth in which it performs the best (these were determined by finite-sample simulation). These optimal bandwidths are underscored next to their corresponding MSE values in the table below. For comparison, we have also included the MSE values for the infinite-order estimator with its best-choice bandwidth.

n	$x = 0$		$x = 1$		$x = 2$		avg on $[-2,2]$	
	50	500	50	500	50	500	50	500
$\text{MSE}_{\text{infinite}}^*$	6.40	.642	6.24	.795	2.76	1.01	4.60	.622
$\text{MSE}_{\text{infinite}}^{*\dagger}$	3.18 _{.60}	.470 _{.50}	1.85 _{1.00}	.139 _{1.00}	1.78 _{.75}	.394 _{.65}	2.92 _{.65}	.425 _{.55}
$\text{MSE}_{\text{Gaussian}}^{*\dagger}$	5.64 _{.50}	1.15 _{.30}	5.18 _{1.00}	.448 _{.65}	1.63 _{.80}	.620 _{.55}	5.46 _{.65}	.779 _{.40}

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

† Optimal bandwidths were used whose values are subscripted.

Table 2: Comparison of the infinite-order kernel with the Gaussian kernel on censored data from lifetime and censoring variables that are iid $\mathcal{N}(0,1)$.

Comparing the two kernels with their respective optimal bandwidths, the infinite-order estimator is clearly the better choice, and even when the bandwidth selection algorithm is used, the infinite-order estimator outperforms the Gaussian estimator with optimal bandwidth at the origin and on the interval $[-2,2]$. So given pretty much any bandwidth selection rule for the Gaussian kernel, the infinite-order estimator is bound to be more accurate over each criterion.

The next simulation uses the same data, but this time we wish to estimate the hazard function. Our infinite-order estimate of the hazard function is $\hat{f}(x)/\hat{S}(x)$ where $\hat{f}(x)$ is the usual infinite-order density estimator and $\hat{S}(x)$ is a smoothed Kaplan-Meier estimator (the R function `ksmooth` was applied to \hat{S} to produce $\hat{S}(x)$). The other two estimators are from the R packages `muhaz` and `logspline`. The `muhaz` estimator is based on the paper [12], and for this simulation the boundary correction is turned off and both global and local bandwidths are invoked (denoted `muhaz-g` and `muhaz-l` respectively). These estimators behave somewhat erratically with small N , so we have changed the sample sizes to 100 and 1000, and we have limited the range of values to $[-1,1]$. The `logspline` estimator (based on [7]) uses splines to estimate the density, and the result is then divided by the smoothed Kaplan-Meier estimate to give an estimate of the hazard function.

	$x = 0$		$x = 1$		avg on $[-1,1]$		
	n	100	1000	100	1000	100	1000
MSE_{infinite}		.0177	.00243	.342	.0334	.0350	.00469
$MSE_{\text{muhaz-g}}$.0478	.0137	.979	.261	.106	.0439
$MSE_{\text{muhaz-l}}$.0239	.00293	.407	.0718	.0557	.00750
$MSE_{\text{logspline}}$.0174	.00284	.204	.119	.0354	.0200

Table 3: Comparison of the infinite-order kernel estimator of the hazard function with the muhaz estimator (with global and local bandwidth selection) and logspline estimator. The lifetime and censoring variables are both iid $\mathcal{N}(0,1)$.

With a large enough sample size, the infinite-order estimator is expected to outperform its competitors; this is witnessed in the above simulation as the infinite-order has the smallest MSE in each category at $n = 1000$ with 50% censoring.

The previous simulation may be considered more of a theoretical comparison since in most applications the censored data is nonnegative. Therefore in the next simulation we use lifetime and censored variables drawn from a lognormal distribution with means 0 and .5 and standard deviations .5 and .5 respectively (values are given on the log scale). Again, due to limitations of the other estimators, we consider datasets of size 100 and 1000, and we only consider the estimates on the interval $[0,1.5]$. Although the lifetime distribution has support on the positive reals, its density takes the value 0 at the origin, so a boundary correction is not necessary for this simulation.

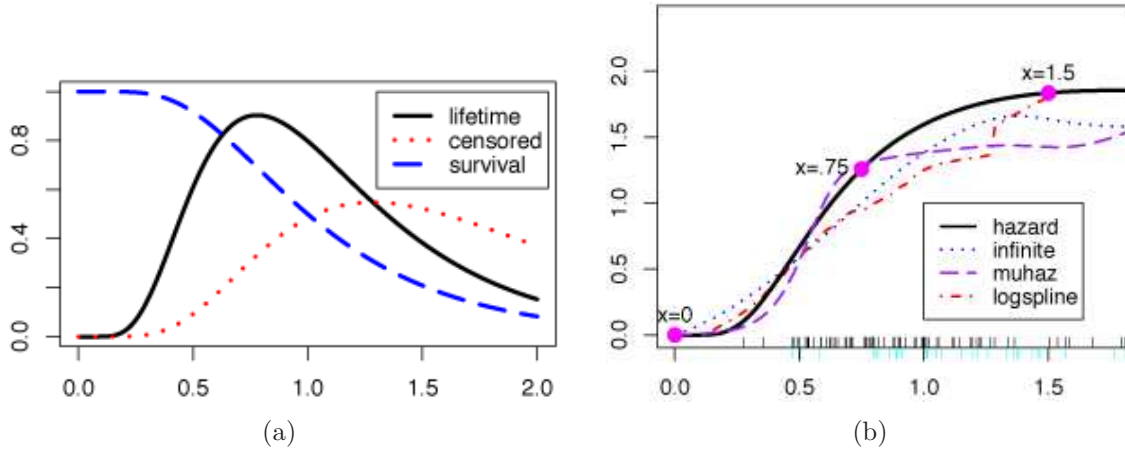


Figure 1: (a) Plot of the lifetime and censored lognormal densities with the survival function. (b) Plot of the hazard function and one realization of the three estimators with $N=100$.

n	$x = 0$		$x = .75$		$x = 1.5$		avg on $[0,1.5]$	
	100	1000	100	1000	100	1000	100	1000
$\text{MSE}_{\text{infinite}}$.000280	2.26e-06	.0527	.00555	.173	.0115	.0502	.0115
$\text{MSE}_{\text{muhaz-g}}$.00462	.000147	.0445	.0111	.429	.0998	.0883	.0203
$\text{MSE}_{\text{muhaz-l}}$.118	.0320	.0478	.0206	.1451	.0585	.0809	.0239
$\text{MSE}_{\text{logspline}}$.000169	4.51e-06	.0757	.00873	.134	.0303	.0661	.0123

Table 4: Comparison of the infinite-order kernel estimator of the hazard function with the muhaz estimator and logspline estimator on lognormal data.

Once again, with the larger data size, the infinite-order estimator in this example outperforms the other estimators in terms of MSE performance. The muhaz estimator is particularly suited for dealing with boundary effects and adapting its bandwidth appropriately as is shown in the next simulation. In the following simulation, the lifetime variables have an exponential distribution with mean one and the censoring variables have an exponential distribution with mean four. Therefore the hazard function in this model is constant with value one.

n	$x = 0$		$x = .75$		$x = 1.5$		avg on $[0,1.5]$	
	100	1000	100	1000	100	1000	100	1000
$\text{MSE}_{\text{infinite}}$.0318	.00419	.0216	.00361	.0676	.00829	.0316	.00769
$\text{MSE}_{\text{muhaz-g}}$.0425	.00514	.0430	.00425	.419	.0125	.0835	.00560
$\text{MSE}_{\text{muhaz-l}}$.0356	.00299	.0251	.00176	.209	.00382	.0562	.00235
$\text{MSE}_{\text{logspline}}$.518	.503	.0319	.00427	.0436	.00610	.0648	.0218

Table 5: Comparison of the infinite-order kernel estimator of the hazard function with the muhaz estimator and logspline estimator on exponential data.

Here we see the infinite-order estimator doing best with the smaller sample size, and it improves with n , but its performance is not as good as the muhaz estimator with local bandwidth selection when $n = 1000$. We expected this behavior, and we describe two reasons that account for asymptotic performance of the infinite-order estimator. The first reason follows directly from Theorem 1—since the characteristic function of the exponential distribution behaves like $|\phi(t)| \sim 1/t$, Theorem 1 indicates there is no benefit to using the infinite-order kernel. In the previous examples, the normal and lognormal distributions have characteristic functions that behave like $|\phi(t)| \sim e^{-t^2}$ and $|\phi(t)| \sim 1/t^{\log t}$ respectively¹, and in both cases Theorem 1 implies a significant MSE improvement with large n . The second reason is due to a lack of a local bandwidth selection procedure for the infinite-order estimator. Simulations computed the optimal

¹Refer to [8] for the derivation of the characteristic function a lognormal distribution

bandwidths in each of the four scenarios in Table 5 for the infinite-order estimator, and the optimal bandwidths were found to be .1, 1.1, .5, and 1.5 which vary widely (compare with optimal bandwidths in Table 2). Therefore a localized bandwidth procedure would be particularly ideal in this situation.

6 Conclusions

The proposed infinite-order estimator together with its tailored bandwidth selection algorithm produce a nearly \sqrt{n} -convergent nonparametric estimator in many standard situations. Even in the least ideal situation of a slow decay of the characteristic function to zero (i.e., when the pdf is not very smooth), the estimator still holds up and can outperform existing methods in many situations. One of the nicest qualities of this estimator is its simplicity—we used the exact same kernel throughout all of the simulations, so no parameter estimation was involved in choosing the kernel, and the accompanying bandwidth selection algorithm requires very little computation to implement. Finally, the proposed estimator is very robust, and since no parameter estimation is involved, it succeeds in estimating the hazard function and density in small sample sizes where competing estimators like `muhaz` and `logspline` fail to even produce an estimate.

A Technical Proofs

PROOF OF THEOREM 1.

Proof. Using the identity

$$\frac{1}{h}K(x/h) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \kappa(th)e^{-itx} dt, \quad (8)$$

and the sample characteristic function given by

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

we rewrite $\hat{f}(x)$ as follows

$$\begin{aligned} \hat{f}(x) &= \frac{1}{h} \sum_{j=1}^n s_j K\left(\frac{x - X_j}{h}\right) \\ &= \frac{1}{h} \sum_{j=1}^n s_j \frac{h}{2\pi} \int_{-\infty}^{\infty} \kappa(th)e^{-it(x-X_j)} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\sum_{j=1}^n s_j e^{itX_j} \right) \kappa(th)e^{-itx} dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\phi}(t)\kappa(th)e^{-itx} dt. \end{aligned} \quad (9)$$

Since $\sum s_j = 1$, we have $E\hat{\phi}(t) = \phi(t)$, and from the representation in (9), the expectation of $\hat{f}(x)$ is

$$E[\hat{f}(x)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(t)\kappa(th)e^{itx} dt.$$

Since $\phi(t)$ is the inverse Fourier transform of $f(x)$, $f(x)$ is therefore the Fourier transform of $\phi(t)$; i.e.

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(t)e^{-itx} dt. \quad (10)$$

Therefore the bias of $\hat{f}(x)$ is

$$\text{bias}(\hat{f}(x)) = E[\hat{f}(x)] - f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (\kappa(th) - 1)\phi(t)e^{-itx} dt.$$

But since $\kappa(th) = 1$ for $|t| \leq 1/h$, we have

$$\text{bias}(\hat{f}(x)) = \frac{1}{2\pi} \int_{|t|>1/h} (\kappa(th) - 1)\phi(t)e^{-itx} dt.$$

Since $|\kappa(t)| \leq 1$ for all t , $|\kappa(th) - 1| \leq 2$ for all h and t . We can then bound the bias by

$$|\text{bias}(\hat{f}(x))| \leq \frac{2}{2\pi} \int_{|t|>1/h} |\phi(t)| dt.$$

Under the assumption $\int |t|^r |\phi(t)| dt < \infty$ in (i), we have

$$\begin{aligned} \int_{|t|>1/h} |\phi(t)| dt &= \int_{|t|>1/h} \frac{|t|^r |\phi(t)|}{|t|^r} dt \\ &\leq h^r \int_{|t|>1/h} |t|^r |\phi(t)| dt \\ &= o(h^r). \end{aligned}$$

If the bias is $o(h^r)$ and the variance is $O\left(\frac{1}{nh}\right)$, then we wish to choose h such that $h^{2r} \sim \frac{1}{nh}$ which occurs if $h \sim an^{-\beta}$ with $\beta = (2r+1)^{-1}$. With this choice of h , we have

$$\sup_{x \in \mathbb{R}} \left| \text{bias} \left\{ \hat{f}(x) \right\} \right| = o\left(n^{\frac{-r}{2r+1}}\right) \quad \text{and} \quad \text{MSE} \left\{ \hat{f}(x) \right\} = O\left(n^{\frac{-2r}{2r+1}}\right).$$

This proves part (i).

Under the assumption $|\phi(t)| \leq De^{-d|t|}$ for some positive constants d and D , we have

$$\begin{aligned} \int_{|t|>1/h} |\phi(t)| dt &\leq D \int_{|t|>1/h} e^{-d|t|} dt \\ &= \frac{D}{e^{d/h}} \int_{|t|>1/h} e^{d(1/h-|t|)} dt \\ &= O\left(e^{-d/h}\right) \end{aligned}$$

So the bias is $O(e^{-d/h})$, and by letting $h \sim 1/(a \log n)$ gives a squared-bias of

$$O\left(e^{-\frac{2d}{h}}\right) = O\left(e^{-2da \log n}\right) = O\left(n^{-2da}\right)$$

and a variance of

$$O\left(\frac{1}{nh}\right) = O\left(\frac{a \log n}{n}\right).$$

Therefore if $a > 1/(2d)$, then

$$\sup_{x \in \mathbb{R}} \left| \text{bias} \left\{ \hat{f}(x) \right\} \right| = O\left(\frac{1}{\sqrt{n}}\right) \quad \text{and} \quad \text{MSE} \left\{ \hat{f}(x) \right\} = O\left(\frac{1}{\sqrt{n}}\right)$$

This proves part (ii).

Under the assumption $\phi(t) = 0$ when $|t| \geq b$, we have

$$\int_{|t| > 1/h} |\phi(t)| dt = 0$$

when $h \leq 1/b$. So by letting $h \leq 1/b$, we have

$$\sup_{x \in \mathbb{R}} \left| \text{bias} \left\{ \hat{f}(x) \right\} \right| = 0 \quad \text{and} \quad \text{MSE} \left\{ \hat{f}(x) \right\} = O\left(\frac{1}{n}\right)$$

which completes the proof of the theorem. \square

Proof. PROOF OF THEOREM 2. By taking the p^{th} derivative on both sides of the identity (8), we have

$$\frac{1}{h^{p+1}} K^{(p)}\left(\frac{x}{h}\right) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (-it)^p \kappa(th) e^{-itx} dt.$$

By taking the p^{th} derivative on both sides of the identity (10), we have

$$f^{(p)}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (-it)^p \hat{\phi}(t) \kappa(th) e^{-itx} dt.$$

Following the steps in (9), we have

$$\hat{f}_p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\phi}(t) \kappa(th) e^{-itx} dt.$$

and we can now compute the bias of $\hat{f}_p(x)$ to be

$$\text{bias}(\hat{f}_p(x)) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (-it)^p (\kappa(th) - 1) \phi(t) e^{-itx} dt.$$

Proceeding as in the proof of Theorem 1, this bias is bounded as

$$|\text{bias}(\hat{f}_p(x))| \leq \frac{2}{2\pi} \int_{|t| > 1/h} |t|^p |\phi(t)| dt.$$

Under assumption $A(r + p)$, we have

$$\begin{aligned} \int_{|t|>1/h} |t|^p |\phi(t)| dt &= \int_{|t|>1/h} \frac{|t|^{r+p} \phi(t)}{|t|^r} dt \\ &\leq h^r \int_{|t|>1/h} |t|^{r+p} |\phi(t)| dt \\ &= o(h^r). \end{aligned}$$

If the bias is $o(h^r)$ and the variance is $O\left(\frac{1}{nh^{p+1}}\right)$, then we wish to choose h such that $h^{2r} \sim \frac{1}{nh^{p+1}}$ which occurs if $h \sim an^{-\beta}$ with $\beta = (2r + p + 1)^{-1}$. With this choice of h , we have

$$\sup_{x \in \mathbb{R}} \left| \text{bias} \left\{ \hat{f}(x) \right\} \right| = o\left(n^{\frac{-r}{2r+p+1}}\right) \quad \text{and} \quad \text{MSE} \left\{ \hat{f}(x) \right\} = O\left(n^{\frac{-2r}{2r+p+1}}\right).$$

Under assumption B ,

$$\begin{aligned} \int_{|t|>1/h} |t|^p |\phi(t)| dt &\leq D \int_{|t|>1/h} |t|^p e^{-d|t|} dt \\ &= \frac{D}{e^{d/h}} \int_{|t|>1/h} |t|^p e^{d(1/h-|t|)} dt \\ &= O\left(e^{-d/h}\right). \end{aligned}$$

Under assumption C ,

$$\int_{|t|>1/h} |t|^p |\phi(t)| dt = 0$$

when $h \leq 1/b$. Finally, the bias and MSE results for parts (ii) and (iii) now follow along the same lines as Theorem 1. \square

PROOF OF THEOREM 3.

Proof. The proof is very similar to the proof of Theorem 3 in [1] with little modification. \square

PROOF OF THEOREM 4.

Proof. Parts (ii) and (iii) follow from Theorems 1 and 2 and the δ -method. The convergence of \hat{h}_M in part (i) is dictated by the slowly converging $\widehat{f}''(x)$. However, the convergence rate of \hat{h}_M is unhampered by the convergence rate of \hat{h} ; for instance, if h is replaced with the random quantity $h(1 + o_p(1))$ (refer to the proof of Lemma 2 in [2]) then Theorem 1 is still valid. If $\phi(t) \sim A|t|^{-d}$, then by Theorem 3,

$$\hat{h} \stackrel{P}{\sim} \tilde{A} \left(\frac{\log n}{n} \right)^{\frac{1}{2d}}$$

From Theorem 2, part (i), if

$$\int_{-\infty}^{\infty} |t|^{r+2} |\phi(t)| < \infty, \tag{11}$$

then the bias of $\widehat{f}''(x)$ is $o(h^r)$. In order for (11) to be satisfied, r must be less than $d - 3$, so we let $r = \lceil d - 4 \rceil$. Therefore the bias of $\widehat{f}''(x)$ (which dominates the MSE of $\widehat{f}''(x)$) is

$$o\left(\frac{\log n}{n}\right)^{\frac{\lceil d-4 \rceil}{2d}},$$

and coupled with the δ -method, part (i) of Theorem 1 is now proved. \square

References

- [1] A. Berg and D. Politis. Higher-order polyspectral estimation with flat-top lag-windows. *Submitted for publication*.
- [2] Peter Bühlmann. Locally adaptive lag-window spectral estimation. *J. Time Ser. Anal.*, 17(3):247–270, 1996. ISSN 0143-9782.
- [3] Luc Devroye. A note on the usefulness of superkernels in density estimation. *Ann. Statist.*, 20(4):2037–2056, 1992. ISSN 0090-5364.
- [4] Peter Hall and J. S. Marron. Lower bounds for bandwidth selection in density estimation. *Probab. Theory Related Fields*, 90(2):149–173, 1991. ISSN 0178-8051.
- [5] M. C. Jones. Simple boundary correction for density estimation. *Statist. Comput.*, 3:135–146, 1993.
- [6] M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.*, 91(433):401–407, 1996. ISSN 0162-1459.
- [7] Charles Kooperberg and Charles J. Stone. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*, 1(4):301–628, 1992.
- [8] Roy B. Leipnik. On lognormal random variables. I. The characteristic function. *J. Austral. Math. Soc. Ser. B*, 32(3):327–347, 1991. ISSN 0334-2700.
- [9] Clive R. Loader. Bandwidth selection: classical or plug-in? *Ann. Statist.*, 27(2):415–438, 1999. ISSN 0090-5364.
- [10] J. S. Marron and W. J. Padgett. Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *Ann. Statist.*, 15(4):1520–1535, 1987. ISSN 0090-5364.
- [11] Timothy L. McMurry and Dimitris N. Politis. Nonparametric regression with infinite order flat-top kernels. *J. Nonparametr. Stat.*, 16(3-4):549–562, 2004. ISSN 1048-5252.
- [12] Hans-Georg Müller and Jane-Ling Wang. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, 50(1):61–76, 1994. ISSN 0006-341X.

- [13] Dimitris N. Politis. Adaptive bandwidth choice. *J. Nonparametr. Stat.*, 15(4-5): 517–533, 2003. ISSN 1048-5252.
- [14] Dimitris N. Politis. On nonparametric function estimation with infinite-order flat-top kernels. In Ch. Charalambides et al., editor, *Probability and Statistical Models with applications*, pages 469–483. Chapman and Hall/CRC, Boca Raton, 2001.
- [15] Dimitris N. Politis. Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices. *Working paper*, 2007.
- [16] Dimitris N. Politis and Joseph P. Romano. Multivariate density estimation with general flat-top kernels of infinite order. *J. Multivariate Anal.*, 68(1):1–25, 1999. ISSN 0047-259X.
- [17] Dimitris N. Politis and Joseph P. Romano. On a family of smoothing kernels of infinite order. In M. Tarter and M. Lock, editors, *Computing Science and Statistics, Proceedings of the 25th Symposium on the Interface*, pages 141–145. The Interface Foundation of North America, 1993.
- [18] C. Sánchez-Sellero, W. González-Manteiga, and R. Cao. Bandwidth selection in density estimation with truncated and censored data. *Ann. Inst. Statist. Math.*, 51(1):51–70, 1999. ISSN 0020-3157.
- [19] Eugene F. Schuster. Incorporating support constraints into nonparametric estimators of densities. *Comm. Statist. A—Theory Methods*, 14(5):1123–1136, 1985. ISSN 0361-0926.
- [20] B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986. ISBN 0-412-24620-1.
- [21] Kagba N. Suaray. *On Kernel Density Estimation for Censored Data*. PhD thesis, Department of Mathematics, University of California, San Diego, 2004.
- [22] Martin A. Tanner and Wing Hung Wong. The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.*, 11(3):989–993, 1983. ISSN 0090-5364.