Reduced bias nonparametric lifetime density and hazard estimation

Arthur Berg, Dimitris Politis, Kagba Suaray & Hui Zeng

TEST
An Official Journal of the Spanish Society of Statistics and Operations Research

ISSN 1133-0686
Volume 29
Number 3

TEST (2020) 29:704-727
DOI 10.1007/s11749-019-00677-z

Springer

TEST (2020) 29:704–727
https://doi.org/10.1007/s11749-019-00677-z

ORIGINAL PAPER

Reduced bias nonparametric lifetime density and hazard estimation

Arthur Berg[1] · **Dimitris Politis**[2] · **Kagba Suaray**[3] · **Hui Zeng**[1]

Received: 16 October 2018 / Accepted: 7 August 2019 / Published online: 17 August 2019

Abstract
Kernel-based nonparametric hazard rate estimation is considered with a special class of infinite-order kernels that achieves favorable bias and mean square error properties. A fully automatic and adaptive implementation of a density and hazard rate estimator is proposed for randomly right censored data. Careful selection of the bandwidth in the proposed estimators yields estimates that are more efficient in terms of overall mean square error performance, and in some cases, a nearly parametric convergence rate is achieved. Additionally, rapidly converging bandwidth estimates are presented for use in second-order kernels to supplement such kernel-based methods in hazard rate estimation. Simulations illustrate the improved accuracy of the proposed estimator against other nonparametric estimators of the density and hazard function. A real data application is also presented on survival data from 13,166 breast carcinoma patients.

Keywords Bandwidth estimation · Density estimation · Fourier transform · Hazard function estimation · Infinite-order kernels · Nonparametric estimation · Survival analysis

Mathematics Subject Classification 62G07 · 62G20 · 62N99

1 Introduction

Hazard rate estimation has been extensively studied in the literature as it encompasses fundamental characteristics of time-to-event data with applications spanning medicine, engineering and economics. The first kernel-based nonparametric estimator of the hazard function with non-censored data appeared in Watson and Leadbetter (1964).

✉ Arthur Berg
berg@psu.edu

1 Division of Biostatistics and Bioinformatics, Penn State College of Medicine, Hershey, PA, USA

2 Department of Mathematics, University of California, La Jolla, CA, USA

3 Department of Mathematics and Statistics, California State University Long Beach, Long Beach, CA, USA

For censored data, density estimation approaches are described in Földes et al. (1981) and Padgett and McNichols (1984), and an empirical hazard approach is described in Tanner and Wong (1983). Kernel-based estimation of the hazard function under censoring was studied by Yandell (1983), Ramlau-Hansen (1983), Tanner and Wong (1984) and Müller and Wang (1994), among others. However, all of these kernel-based approaches capitalized on traditional theory of second-order kernels when constructing their kernel-based estimates. Through the use of infinite-order kernels, we demonstrate that considerable asymptotic improvements are attainable.

The benefit of using infinite-order kernels, also called superkernels, in estimating the probability density function under iid data is well known; cf. Devroye (1992). More recently, Politis and others have investigated a class of infinite-order kernels that are constructed by taking the Fourier transform of flat-top functions—functions that are flat in a neighborhood of the origin (Berg and Politis 2009, 2010; McMurry and Politis 2004; Politis and Romano 1993, 1999). These estimators, under a correctly specified bandwidth, attain mean square error (MSE) properties superior to their second-order analogs and also perform well in small sample simulation studies. These same properties translate nicely to the context of density estimation under random right censoring, as investigated here. Improved MSE convergence rates in nonparametric estimation of the hazard function and derivatives of the density follow as corollaries to the density estimation theory.

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 density estimators from censored data to become essentially negligible in certain situations. Section 3 completes the proposed estimator by providing a bandwidth selection algorithm that automatically adapts to the unknown density at hand. A second use of the infinite-order estimators is realized in Sect. 4 by providing rapidly converging bandwidths for use in second-order kernels. In Sect. 5, we give practical suggestions for implementing the proposed estimator and provide simulations exhibiting improved performance in estimating the lifetime density and hazard function when compared with other nonparametric estimators including the `muhaz` and `pehaz` estimators of Hess and Gentleman (2014), Müller and Wang (1994) and the `presmooth` estimator of Lopez de Ullibarri and Jácome (2013). In Sect. 6, the proposed hazard function estimator and the previously mentioned estimators are simultaneously compared on breast carcinoma survival data involving 13,166 women.

2 Estimation with infinite-order kernels

We lay out the notation under the context of random right censorship. [This can be generalized to allow for left truncation; see, for example, Sánchez-Sellero et al. (1999).] Let $X_1^0, \ldots, X_n^0$ be iid lifetime variables with density f and cdf F, and independently, let $U_1, \ldots, U_n$ be iid censoring variables with density g and cdf G. We observe the data Z_i and Δ_i where

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

for $i = 1, \ldots, n$. (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

ith-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

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

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

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

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

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

See Földes et al. (1981) and Marron and Padgett (1987) for background and properties of this estimator. Many authors require K to be of compact support for ease of analysis, but this is unnecessary; see, for example, Tanner and Wong (1983). Therefore, we only assume K is an even function that integrates to one.

It will be assumed that sufficient conditions are satisfied so that

$$\mathrm{var}\left(\hat{f}(x) \right) = O\left(\frac{1}{nh} \right). \tag{2}$$

This typically requires the lifetime density f to be continuously differentiable at x and the censored distribution to be of compact support. Under these conditions, precise variance expressions are provided in Gijbels and Wang (1993). See Wasserman (2006) for simpler derivations of these results for iid data.

Following Politis (2001), we now describe a class of infinite-order kernels constructed from the Fourier transform of a flat-top function. We start in the Fourier domain with a function κ given by

$$\kappa(t) = \begin{cases} 1, & |t| \le c \\ q(|t|), & \text{otherwise}, \end{cases} \tag{3}$$

where c is any positive constant and q is any continuous, square-integrable function that is bounded in absolute value by one and satisfies $q(|c|) = 1$. Then the infinite-order kernel corresponding to κ is the Fourier transform of κ, specifically,

$$K(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \kappa(t)\mathrm{e}^{-itx}\,\mathrm{d}t, \tag{4}$$

or equivalently,

$$\frac{1}{h}K(x/h) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \kappa(th)\mathrm{e}^{-itx}\,\mathrm{d}t. \tag{5}$$

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

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

A natural estimator of the characteristic function is

$$\hat{\phi}(t) = \int_{-\infty}^{\infty} \mathrm{e}^{itx}\,\mathrm{d}\hat{F}(x) = \sum_{j=1}^{n} s_j\,\mathrm{e}^{itX_j}. \tag{6}$$

In the context of non-censored data, $\hat{\phi}$ is an unbiased estimator of ϕ, but in the presence of censoring, bias is present. We assume the bias of $\hat{\phi}(x)$ is $O(\frac{1}{n})$, which is justified by the following lemma.

Lemma 1 *Suppose $g(x)$ (the censored density) is compactly supported and contains the support of $f(x)$, then*

$$bias(\hat{\phi}(x)) = O\left(\frac{1}{n}\right). \tag{7}$$

The assumption in Lemma 1 that the support of the censored distribution contains the support of the lifetime distribution can be found in other papers [see, e.g., de Uña-Álvarez and Meira-Machado (2015)]. This assumption may be limiting in certain applications, but often it suffices to estimate a truncated lifetime distribution.

The bias of $\hat{f}(x)$ is smaller when $f(x)$ is smoother (has more derivatives), and the smoother the $f(x)$ is, the faster its characteristic function decays to zero. The following assumptions classify the smoothness of $f(x)$ into one of the three categories.

- *Assumption A(r)* There is an $r > 0$ such that $\int_{|t|>\delta} |t|^r\,|\phi(t)|\,\mathrm{d}t < \infty$ for any $\delta > 0$.
- *Assumption B* There are positive constants d and D such that $|\phi(t)| \le D\mathrm{e}^{-d|t|}$.
- *Assumption C* There is a positive constant b such that $|\phi(t)| = 0$ when $|t| \ge b$.

The following theorem provides rates of the bias and MSE under each of the assumptions above.

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) and the bias assumption in (7).*

(i) *Suppose assumption* $A(r)$ *holds. Let* $h \sim an^{-\beta}$ *(for any* $a > 0$*) with* $\beta = (2r+1)^{-1}$*, then*

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

(ii) *Suppose assumption B holds. Let* $h \sim 1/(a\log n)$ *with* $a > 1/(2d)$*, then*

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

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

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

This theorem illustrates the mean square error of $\hat{f}(x)$ is just as good as second-order kernel density estimation when $f(x)$ is only twice differentiable ($r = 2$), but considerable improvements are gained when more smoothness of $f(x)$ is present. Even a parametric convergence rate is possible when assumption C is satisfied. Parametric convergence rates in non-censored data have also demonstrated by Davis (1977), Ibragimov and Khasminskii (1983) and Chacón et al. (2007), such as for data following a Vallé–Poussin density given by

$$f(x) = \frac{1-\cos x}{\pi x^2}, \quad x \in \mathbb{R}$$

with finitely supported characteristic function given by

$$\phi(t) = (1-|t|)I_{[-1,1]}(t).$$

Corollary 1 *The hazard function* $\lambda(x) = f(x)/S(x)$ *is naturally estimated by* $\hat{\lambda}(x) = \hat{f}(x)/\hat{S}(x)$*, 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* $A(r)$, $\text{MSE}(\hat{\lambda}(x)) = O\left(n^{\frac{-2r}{2r+1}}\right)$;

(ii) *Under assumption B,* $\text{MSE}(\hat{\lambda}(x)) = O\left(\frac{\log n}{n}\right)$;

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

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

$$\hat{f}_p(x) = \frac{1}{h^{p+1}} \sum_{j=1}^{n} s_j K^{(p)}\left(\frac{x-X_j}{h}\right) \tag{8}$$

is an estimate of the pth derivative of f (Singh 1977). Similarly, under sufficient conditions on f, the variance of this estimator is

$$\operatorname{var}\left(\hat{f}_p(x)\right) = O\left(\frac{1}{n\,h^{p+1}}\right). \tag{9}$$

The previous theorem is now 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 (8) where K is an infinite-order kernel, and assume (7) and (9) hold.*

(i) *Suppose assumption $A(r+p)$ holds. Let $h \sim an^{-\beta}$ (for any $a > 0$) with $\beta = (2r+p+1)^{-1}$, then*

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

(ii) *Suppose assumption B holds. Let $h \sim 1/(a\log n)$ with $a > 1/(2d)$, then*

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

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

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

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.

The bias properties stated in Theorem 2.1 and Theorem 2.3 are consistent with the properties of other kernel-based estimators with infinite-order kernels used in different contexts; see, e.g., Berg and Politis (2009) and Politis and Romano (1999).

3 Bandwidth selection algorithm

Theorem 1 in the previous section assumes one is handed a bandwidth h that is precisely molded to the underlying smoothness of the density of interest $f(x)$. In general, however, one does not necessarily know the level of smoothness of the underlying density. This section presents a simple algorithm, adapted from Politis (2003), that automatically adjusts to the unknown smoothness of $f(x)$. This bandwidth estimation procedure is consistent for the optimal bandwidth h under assumptions B and C, and under assumption A(r), the bandwidth algorithm still adapts to the underlying smoothness, but consistency does not hold.

Let $\hat{\phi}(t)$ be the estimate of the characteristic function as given in (6). The algorithm, in essence, searches for the smallest value t^* such that $\hat{\phi}(t^*) \approx 0$ in which case the bandwidth estimate is taken to be $\hat{h} \approx 1/t^*$. The specific details are provided in the following algorithm.

BANDWIDTH SELECTION ALGORITHM
Let $C > 0$ be a fixed constant and ε_n be a non-decreasing sequence of positive real numbers tending to infinity such that $\varepsilon_n = o(\log n)$. Let t^* be the smallest number such that

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

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

Remark 1 The positive constant C and the choice of sequence ε_n are irrelevant in the asymptotic theory, but certainly relevant for finite sample calculations. The main idea behind the algorithm is to determine the smallest t such that $\phi(t) \approx 0$, and in most cases, this can be visually seen without explicitly providing the quantities C and ε_n in (10). For a more automated procedure, a recommendation to take $C = 2$ and $\varepsilon_n = 5$ is suggested in Remark 2.3 of Politis (2003). The R package `iosmooth` McMurry and Politis (2017) provides an implementation of the bandwidth selection algorithm for iid data.

Remark 2 If $q(t)$ in (3) is very close to one, in a neighborhood of the type $[c, c + \eta]$, then the "flat-top radius" is effectively increased to $c + \eta$. In this case, we would let $\hat{h} = (c + \eta)/t^*$ in the bandwidth selection algorithm. This is particularly relevant when considering infinitely smooth flat-top functions (McMurry and Politis 2004).

Theorem 3 *Assume the following two assumptions on* $\hat{\phi}(t)$*:*

$$\max_{s\in(0,1)} |\hat{\phi}(t+s) - \phi(t+s)| = O_P(1/\sqrt{n}) \tag{11}$$

and

$$\max_{s\in(0,n)} |\hat{\phi}(t+s) - \phi(t+s)| = O_P\left(\frac{\log n}{\sqrt{n}}\right) \tag{12}$$

uniformly in t*.*

(i) *If* $|\phi(t)| \sim A|t|^{-d}$ *for some positive constants* A *and* d*, then*

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

here, $A \overset{P}{\sim} B$ *means* $A/B \to 1$ *in probability.*

(ii) *If* $|\phi(t)| \sim A\xi^{|t|}$ *for some* $\xi \in (0, 1)$ *and* $A > 0$*, then*

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

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

(iii) *If* $|\phi(t)| = 0$ *when* $|t| \geq b$*, then* $\hat{h} \overset{P}{\sim} 1/b$.

Remark 3 The two assumptions (11) and (12) are typical assumptions invoked in this type of an algorithm [see, e.g., Politis (2003)], and verification of these assumptions, particularly with censored data, can be difficult and is not pursued here.

Remark 4 Note that the polynomial decay assumption of $\phi(t)$ in part (i) implies assumption $A(d - 1 - \varepsilon)$ (as considered in Theorem 1) for any $\varepsilon > 0$. The decay assumptions in parts (ii) and (iii) correspond to assumptions B and C, respectively, in Theorem 1.

Theorem 3 shows that the proposed bandwidth selection algorithm adapts to the underlying degree of smoothness of the density, yielding bandwidth estimates that largely match 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 second-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. Although Theorem 1 demonstrates asymptotic superiority of using infinite-order kernels over second-order kernels, the choice of bandwidth in estimation may be more critical than the choice of kernel. This hybrid approach provides improved (rapidly converging) bandwidth estimates for kernel density estimators using second-order kernels.

We begin with expressions for the MSE and the mean integrated square error (MISE) of $\hat{f}(x)$ with a symmetric second-order kernel Λ and standard assumptions on f and G. The MISE below is slightly generalized to incorporate a nonnegative weight function $\omega(x)$ to control the influence of error in the tails of the estimated density. The MSE calculations will assume the following conditions:

(i) $f(x)$, $G(x)$ and $\omega(x)$ are twice differentiable with bounded third derivative in a neighborhood of x.

(ii) $\omega(x)$ is compactly supported whose support is contained inside the support of the censoring distribution.

(iii) Λ is three times continuously differentiable, its first derivative is integrable, and

$$\lim_{|x| \to \infty} x^j \Lambda^{(j)}(x) = 0 \quad (j = 0, 1, 2, 3).$$

MSE and MISE expressions are now presented; further details of the derivations can be found in Sánchez-Sellero et al. (1999) and Suaray (2008).

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

where

$$c_\Lambda = \int_{-\infty}^{\infty} x^2 \Lambda(x)\,\mathrm{d}x \quad \text{and} \quad d_\Lambda = \int_{-\infty}^{\infty} \Lambda^2(x)\,\mathrm{d}x$$

and

$$\mathrm{MISE}(\hat{f}) = \int_{-\infty}^{\infty} \mathrm{MSE}(\hat{f}(x))\mathrm{d}x.$$

Minimizing the asymptotically dominant terms in the above expressions with respect to h yields pointwise and globally optimal bandwidths, respectively, given by

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

These optimal bandwidths involve values of the unknown values $f(x)$, $f''(x)$ and $G(x)$. Therefore, to estimate the respective bandwidths, we replace these unknown quantities with pilot estimates; $f(x)$ and $f''(x)$ are replaced with the infinite-order kernel estimates $\hat{f}(x)$ and $\hat{f}_2(x)$, respectively, and $1-G(x)$, the survival function of the *censored* random variables, is estimated using the Kaplan–Meier estimator with Δ_i replaced with $1-\Delta_i$. The bandwidth used in estimating $\hat{f}_2(x)$, and in general for $\hat{f}_p(x)$, is the same as that derived from the bandwidth selection algorithm above. If $f(x)$ is sufficiently smooth (for instance, when assumption B or C holds), then this bandwidth choice is optimal. Let $\hat{h}_{\mathrm{MSE}}$ and $\hat{h}_{\mathrm{MISE}}$ 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 plug-in infinite-order kernel estimators, as detailed in the following theorem.

Theorem 4 *Assume the conditions of Theorem 3, and assume conditions strong enough to ensure (9) holds for $p = 2$. Let $\hat{h}_{\mathrm{M}}$ be either $\hat{h}_{\mathrm{MSE}}$ or $\hat{h}_{\mathrm{MISE}}$ with h_{M} being the corresponding h_{MSE} or h_{MISE}.*

(i) *If $|\phi(t)| \sim A|t|^{-d}$ for some positive constants A and $d > 3$, then*

$$\hat{h}_{\mathrm{M}} = h_{\mathrm{M}} \left(1 + O_p \left(\frac{\log n}{n} \right)^{\frac{\lceil d-4 \rceil}{2d}} \right).$$

(ii) *If $|\phi(t)| \sim A\xi^{|t|}$ for some $\xi \in (0, 1)$ and $A > 0$, then*

$$\hat{h}_{\mathrm{M}} = h_{\mathrm{M}} \left(1 + O_p \left(\frac{\log n}{n} \right)^{\frac{1}{2}} \right).$$

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

$$\hat{h}_{\mathrm{M}} = h_{\mathrm{M}} \left(1 + O_p \left(\frac{1}{\sqrt{n}} \right) \right).$$

Cross-validation is suggested in Marron and Padgett (1987) as a means of minimizing the integrated square error (ISE), but the approach of minimizing ISE was shown in Hall and Marron (1991) 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 Marron and Padgett (1987) are $n^{-1/10}$, 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 slower than the $n^{-1/2}$ rate afforded by the proposed procedure under a sufficiently smooth density $f(x)$ (i.e., when $\phi(t)$ has a rapid decay to zero) as Theorem 4 demonstrates.

5 Simulations

Many different choices of "flat-top" functions (3) can be used to construct an infinite-order kernel, although highly non-smooth shapes like the rectangle, which gives rise to the sinc kernel, should be avoided due to its large and slowly decaying side lobes. The trapezoidal window, as suggested in Politis and Romano (1995), can be viewed as smoothening the rectangular window and has more rapidly decaying side lobes. Another possibility is the infinitely smooth trapezoidal flat-top shape McMurry and Politis (2004) which has side lobes that decay exponentially fast. The simulations in this article invoke a simple trapezoidal shape defined as

$$\kappa(t) = \begin{cases} 1, & |t| \leq \frac{1}{2} \\ -2|t| + 2, & \frac{1}{2} \leq |t| \leq 1 \\ 0, & \text{else.} \end{cases} \tag{13}$$

Taking the Fourier transform of this function gives the infinite-order kernel of interest:

$$K(x) = \frac{2\left(\cos(x/2) - \cos(x)\right)}{\pi x^2}. \tag{14}$$

We demonstrate the performance of using this infinite-order kernel for randomly right censored density and hazard function estimation in finite sample simulations. Reproducible code for all of the simulations is provided as supplementary materials.

5.1 Normal kernel versus infinite-order kernel with normal data

In this simulation, we simply compare the performance of a normal kernel against the infinite-order kernel (14) and remove the complicating issue of bandwidth selection. Specifically, we determine the MSE performance for each estimator under their respective optimal bandwidth. Lifetime and censoring data are simulated independently following a standard normal distribution, thus yielding a censoring rate of 50% on average. The characteristic function of the standard normal distribution is $\phi(t) = \exp(-\frac{t^2}{2})$, which implies Assumption B is valid and the infinite-order kernel is asymptotically more efficient. Estimates of the normal density at three points ($x = 0$, 1 and 2) are considered along with two different sample sizes ($n = 50$ and 500). Results are provided for 999 realizations, which is sufficiently large to yield very small confidence intervals of the estimates. The results of the simulation study (Table 1) shows improved MSE performance when using an infinite-order kernel, particularly with the larger sample size.

Table 1 Comparison of the infinite-order kernel to the normal kernel with their respective optimal bandwidths

	$x = 0$	$x = 1$	$x = 2$
$n = 50$			
$\text{MSE}^{\text{a}}_{\text{infinite}}$	$\mathbf{3.96}_{.40}$	$\mathbf{1.98}_{.70}$	$1.78_{.50}$
$\text{MSE}^{\text{a}}_{\text{normal}}$	$5.90_{.50}$	$3.93_{.90}$	$\mathbf{1.33}_{.90}$
$n = 500$			
$\text{MSE}^{\text{a}}_{\text{infinite}}$	$\mathbf{.54}_{.30}$	$\mathbf{.28}_{.50}$	$\mathbf{.47}_{.40}$
$\text{MSE}^{\text{a}}_{\text{normal}}$	$1.14_{.30}$	$.60_{.50}$	$.61_{.50}$

Values in bold indicate the smaller of the two MSE values in each comparison

[a]MSE values are multiplied by 10^3 for easier comparison and subscripted values correspond to the optimal bandwidth

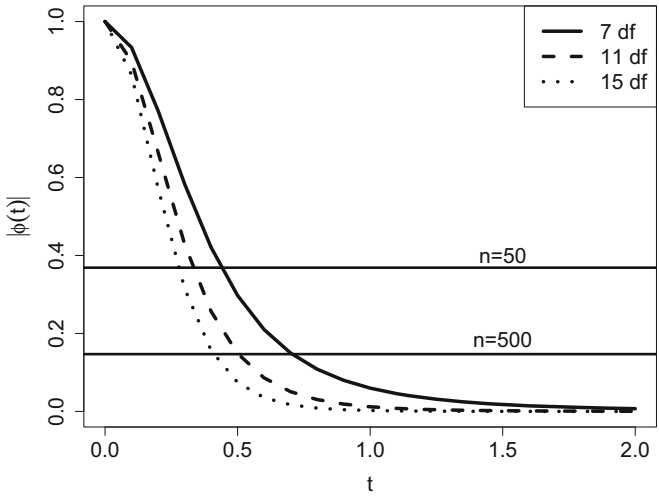

Fig. 1 True characteristic function for each of the three Chi-square densities with horizontal lines corresponding to the threshold in (10)

5.2 Hazard function estimation with Chi-square data

In this simulation, we evaluate the performance of kernel density estimation on χ^2_ν data using three different degrees of freedom: $\nu = 7$, $\nu = 11$ and $\nu = 15$. The characteristic function for the χ^2_ν distribution is $(1-2it)^{-\nu/2}$, which implies assumption $A(r)$ holds for $r = 2$, 4 and 6, respectively. We first demonstrate the performance of the adaptive bandwidth selection algorithm discussed in Sect. 3.

5.2.1 Bandwidth selection algorithm

The true characteristic function for each of the three densities is shown in Fig. 1. The two horizontal lines correspond to the thresholds given in Eq. (10) for $C = 2$ and $n = 50$ and 500, respectively. Following the bandwidth selection algorithm, we let t^* be the value of t corresponding to the point where $|\phi(t)|$ crosses the horizontal line, and then, we set $h = 1/(2t^*)$.

Figure 2 shows the distribution of bandwidths for estimating the density at $x = 10$ as determined by the bandwidth selection algorithm. This fully automated procedure consistently identified the bandwidths in a narrow range, and its adaptive nature is observed as it produces increasingly larger bandwidths as the smoothness of the underlying density increases. It also adapts to the sample size by producing smaller bandwidths with larger sample sizes.

5.2.2 Comparison of hazard estimators

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

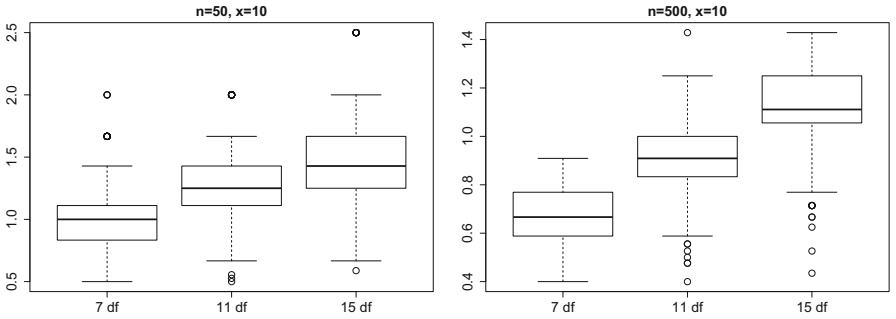

Fig. 2 Distribution of bandwidths from the bandwidth selection algorithm for estimating the density at $x = 10$ when $n = 50$ (left) and $n = 500$ (right)

estimators are not consistent near the boundary points. However, a number of fixes for this boundary issue are available [see Jones (1993); Karunamuni and Alberts (2005) for a survey of several methods], and we adopt the simple reflection principle to resolve boundary problems in our estimator. Specifically, as described in Marron and Ruppert (1994), when the density is known to have its support on $[0, \infty)$, we use the estimator $\hat{\hat{f}}(x) = \hat{f}(x) + \hat{f}(-x)$ near the boundary ($0 \le x \le h$) to ensure consistency near the boundary point $x = 0$; see Schuster (1985) and Silverman (1986) for discussions of this method with non-censored data.

In Table 2, we compare various estimators of the hazard function on the Chi-square data. The infinite-order kernel estimator of the hazard function is $\hat{f}(x)/\tilde{S}(x)$ where $\hat{f}(x)$ is the usual infinite-order density estimator and $\tilde{S}(x)$ is a smoothed Kaplan–Meier estimator. (The R function `ksmooth` was applied to $\hat{S}$ to produce $\tilde{S}(x)$.) The other estimators considered are derived from the R packages `muhaz` and `survPresmooth`. The `muhaz` estimator is based on the paper (Müller and Wang 1994) with local bandwidth selection (denoted muhaz-l). The `presmooth` estimator is based on the paper (Tanner and Wong 1983) and uses the plug-in method for bandwidth selection (Cao and López de Ullibarri 2007).

For the parameters in (10) of the bandwidth selection algorithm, we set $C = 2$. In these simulations, $|\hat{\phi}(t)|$ is still decreasing when it hits the bound in (10), so in order to simplify and easily automate the bandwidth selection procedure for these simulations, we simply fixed $\varepsilon_n = 0$. Also, since `presmooth` would fail in some simulations when estimating $x = 10$, we focused the simulations on estimating $x = 7$. Finally, to provide comparison across a wide range of sample sizes and censoring rates, we consider three different sample sizes—$n = 50$, 250 and 500—and three different censorship rates—25%, 50% and 75%. Censorship of 50% is obtained by setting the censoring distribution equal to the lifetime distribution. Censorship of 25% is obtained by simple translation of the lifetime distribution to the right to achieve the desired censorship rate. Similarly, censorship of 75% is obtained by translating the censored distribution to the right to achieve the desired censorship rate. Simulations are conducted over 999 realizations and the standard errors are negligible.

Indeed we see the mean square error generally improves with sample size and worsens with increased censoring. The infinite-order kernel approach is generally expected to perform better with higher degrees of freedom of the Chi-square distribution, which

Table 2 Mean square error performance of the proposed method (`infinite`) to two other hazard estimators over a range of censorship rates, sample sizes and lifetime distributions

% Cens	n	χ^2_d	`infinite`	`muhaz`	`presmooth`
25	50	7 df	8.86	3.62	**2.36**
		11 df	0.24	0.18	**0.09**
		15 df	**0.37**	0.98	0.39
	250	7 df	**0.16**	0.17	0.37
		11 df	0.08	**0.04**	0.05
		15 df	**0.02**	0.09	0.04
	500	7 df	**0.02**	0.06	0.10
		11 df	**0.01**	0.03	0.02
		15 df	**0.01**	0.04	0.03
50	50	7 df	11.95	15.70	**8.71**
		11 df	0.69	**0.58**	0.81
		15 df	**0.10**	0.21	0.20
	250	7 df	1.52	**1.41**	1.57
		11 df	0.14	**0.11**	0.14
		15 df	**0.03**	0.06	0.09
	500	7 df	0.85	**0.73**	0.91
		11 df	0.09	**0.06**	0.08
		15 df	**0.02**	0.05	0.07
75	50	7 df	**10.81**	14.47	14.59
		11 df	7.68	6.75	**6.17**
		15 df	0.68	0.66	**0.49**
	250	7 df	13.59	13.08	**9.66**
		11 df	7.71	**6.31**	7.18
		15 df	0.76	0.69	**0.37**
	500	7 df	9.32	10.35	**8.02**
		11 df	7.86	**6.71**	7.17
		15 df	0.84	**0.62**	0.83

The above MSE values are multiplied by 10^3 for easier comparison, and within each row, the smallest MSE value is bolded

is observed with these results. In this simulation, none of the three automated hazard density estimators dominates the others in terms of mean square error performance: `infinite` tended to perform best under low censorship and high degrees of freedom, `muhaz` tended to perform best under larger sample sizes, and `presmooth` tended to perform best under high censorship.

6 Breast carcinoma survival data

Breast carcinoma survival data, originally analyzed in Yakovlev et al. (2000), involve 13,166 breast carcinoma patients identified through the Utah Cancer Registry. In the

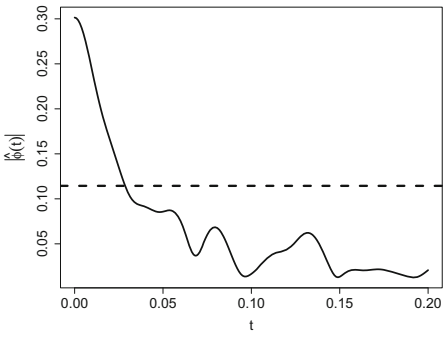

Fig. 3 $|\hat{\phi}(t)|$ for the breast cancer dataset (age < 46, local) along with the threshold used to determine the bandwidth

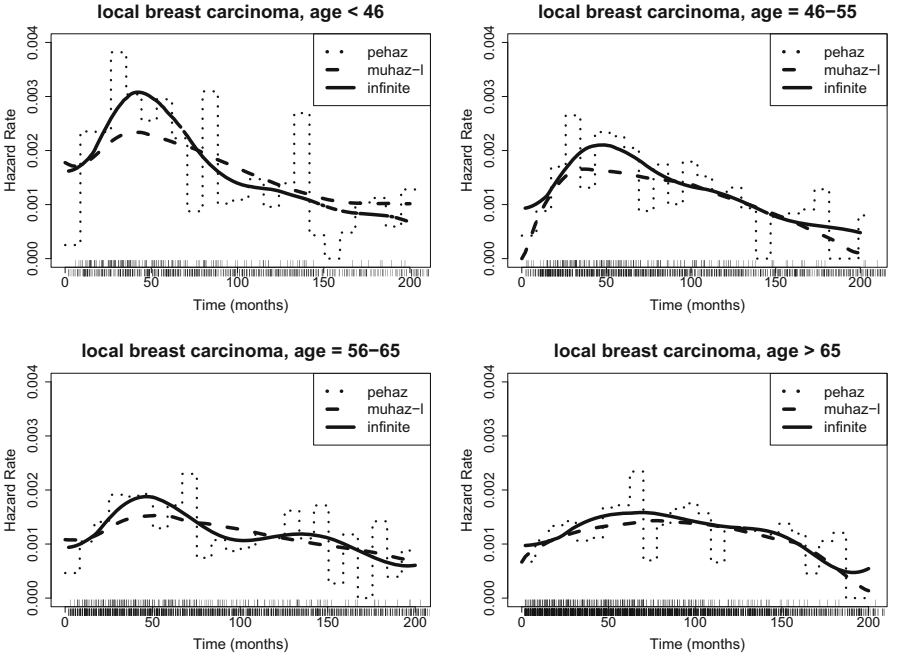

Fig. 4 Result of the three different hazard function estimators for localized breast cancer survival data for four different age ranges

original study, a piecewise hazard function and a kernel-based method were used to estimate the hazard functions of the individuals across different strata based on age and whether carcinoma was localized or not. This dataset is reanalyzed with the proposed hazard function estimator along with the `muhaz` and `pehaz` estimators of Hess and Gentleman (2014) and Müller and Wang (1994).

Figure 3 shows a graph of $|\hat{\phi}(t)|$ along with the same threshold as used in the simulations. We can observe $\hat{\phi}(t)$ smoothly decays toward zero in this real dataset. This allows one to easily determine a reasonable range for the bandwidths to accompany the kernel density estimator.

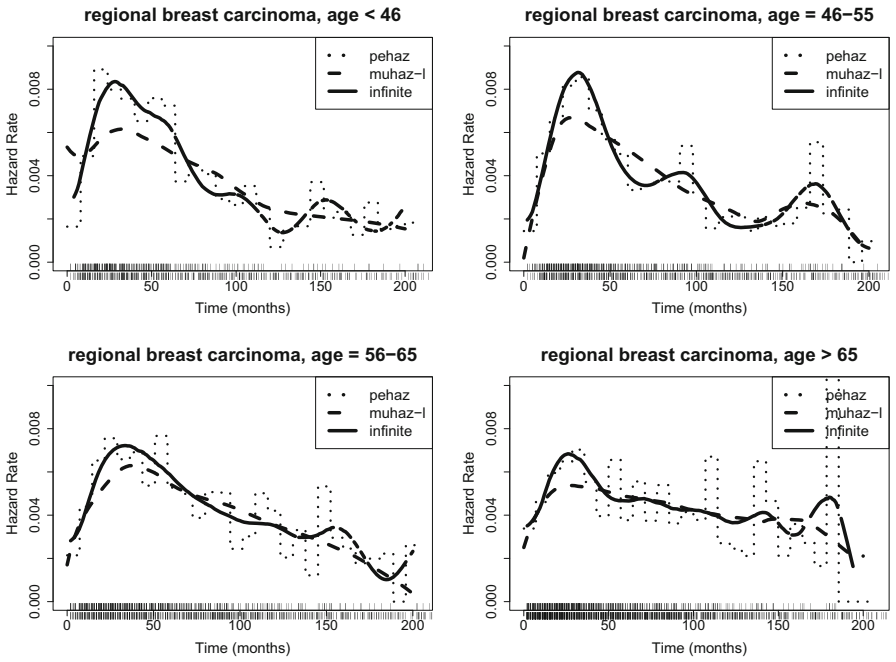

Fig. 5 Result of the three different hazard function estimators for non-local (regional) breast cancer survival data for four different age ranges

Figures 4 and 5 present the results of the different hazard function estimators on the breast cancer dataset. It is consistently depicted among all of the estimators that as the severity of the disease increases, so does the hazard rate. There is little difference in the estimated hazard rates for the different age-groups. The muhaz and infinite-order kernel estimator with adaptive bandwidth choice perform similarly on this dataset.

7 Conclusions

The proposed infinite-order estimator, when used with its tailored bandwidth selection algorithm, produces a nearly $\sqrt{n}$-convergent nonparametric estimator when the underlying density is sufficiently smooth, which corresponds to a rapidly decaying characteristic function. Even in the least ideal situation of a slow decay of the characteristic function to zero (i.e., when the density is not very smooth), the estimator maintains the same performance as traditional kernel density estimators of censored data. The same kernel was used 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. Additionally, the proposed estimator is robust to sample size since no parameter estimation is involved and it can succeed in estimating the hazard function and density in small sample sizes where competing estimators may fail to produce an estimate. Finally, the proposed

estimator demonstrated reliable performance on the simulated data as well as on a actual datasets.

Acknowledgements We appreciate the efforts of Professors Alex Tsodikov and Richard Keeber who helped us to obtain the Utah Cancer Registry data that were analyzed in this manuscript. We also appreciate the meticulous efforts of the anonymous reviewers, which have led to substantial improvements of the manuscript.

A Technical proofs

A.1 Proof of Lemma 1

Theorem 2.1 in Zhou (1988) provides the following result: If $\theta(t)$ is a continuous nonnegative measurable function with $\mathrm{E}[\theta(X_1)] < \infty$, then

$$0 \le \int_{-\infty}^{\infty} \theta(t)\,\mathrm{d}F(t) - \mathrm{E}\left(\int_{-\infty}^{\infty} \theta(t)\,\mathrm{d}\hat{F}(t)\right) \le \int_{-\infty}^{\infty} P(Z_1 \le t)^n \theta(t)\,\mathrm{d}F(t).$$

By linearity of the integral and since $\theta(t) = \theta^+(t) - \theta^-(t)$ where $\theta^+(t) = \max(\theta(t), 0)$ and $\theta^-(t) = \max(-\theta(t), 0)$, we have the following result for general $\theta(t)$

$$\left|\int_{-\infty}^{\infty} \theta(t)\,\mathrm{d}F(t) - \mathrm{E}\left(\int_{-\infty}^{\infty} \theta(t)\,\mathrm{d}\hat{F}(t)\right)\right| \le \int_{-\infty}^{\infty} P(Z_1 \le t)^n (\theta^+(t)+\theta^-(t))\,\mathrm{d}F(t).$$

In particular, for $\theta(t) = \mathrm{e}^{itx} = \cos(tx) + i\sin(tx)$, it follows that

$$\begin{aligned}
\left|\mathrm{bias}\left(\hat{\phi}(x)\right)\right| &= \left|\int_{-\infty}^{\infty} \theta(t)\,\mathrm{d}F(t) - \mathrm{E}\left(\int_{-\infty}^{\infty} \theta(t)\,\mathrm{d}\hat{F}(t)\right)\right| \\
&\le \left|\int_{-\infty}^{\infty} \cos(tx)\,\mathrm{d}F(t) - \mathrm{E}\left(\int_{-\infty}^{\infty} \cos(tx)\,\mathrm{d}\hat{F}(t)\right)\right| \\
&\quad + \left|\int_{-\infty}^{\infty} \sin(tx)\,\mathrm{d}F(t) - \mathrm{E}\left(\int_{-\infty}^{\infty} \sin(tx)\,\mathrm{d}\hat{F}(t)\right)\right| \\
&\le 2\int_{-\infty}^{\infty} P(Z_1 \le t)^n\,\mathrm{d}F(t). \\
&\le 2\max\left(\int_{-\infty}^{\infty} F(t)^n\,\mathrm{d}F(t), \int_{-\infty}^{\infty} G(t)^n\,\mathrm{d}F(t)\right).
\end{aligned}$$

Note that

$$\int_{-\infty}^{\infty} F(t)^n\,\mathrm{d}F(t) = \frac{F(t)^{n+1}}{n+1}\Bigg|_{-\infty}^{\infty} = O\left(\frac{1}{n}\right).$$

From the assumptions of the lemma, we have $f(x)/g(x) \leq M$ for some $M > 0$, which gives

$$\begin{aligned}\int_{-\infty}^{\infty} G(t)^n \, \mathrm{d}F(t) &= \int_{\{t:f(t)\neq 0\}} G(t)^n f(t) \, \mathrm{d}t \\ &= \int_{\{t:f(t)\neq 0\}} G(t)^n \frac{f(t)}{g(t)} g(t) \, \mathrm{d}t \\ &\leq M \int_{\{t:f(t)\neq 0\}} G(t)^n \, \mathrm{d}G(t) \\ &= M \frac{G(t)^{n+1}}{n+1}\Big|_{-\infty}^{\infty} \\ &= O\left(\frac{1}{n}\right).\end{aligned}$$

This establishes the bias of $\hat{\phi}(t)$ is $O(1/n)$ under the assumptions of Lemma 1.

A.2 Proof of Theorem 1

Proof In order to evaluate the bias of $\hat{f}(x)$, we reformulate $\hat{f}(x)$ in terms of $\hat{\phi}(x)$ as follows

$$\begin{aligned}\hat{f}(x) &= \frac{1}{h}\sum_{j=1}^{n} s_j K\left(\frac{x - X_j}{h}\right) \\ &= \frac{1}{h}\sum_{j=1}^{n} s_j \frac{h}{2\pi}\int_{-\infty}^{\infty} \kappa(th)\mathrm{e}^{-it(x-X_j)} \, \mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\sum_{j=1}^{n} s_j \mathrm{e}^{itX_j}\right)\kappa(th)\mathrm{e}^{-itx} \, \mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{\phi}(t)\kappa(th)\mathrm{e}^{-itx} \, \mathrm{d}t.\end{aligned} \tag{15}$$

From the representation in (15), the expectation of $\hat{f}(x)$ is

$$\begin{aligned}\mathrm{E}[\hat{f}(x)] &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{E}\left[\hat{\phi}(t)\right]\kappa(th)\mathrm{e}^{-itx} \, \mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi(t)\kappa(th)\mathrm{e}^{-itx} \, \mathrm{d}t + O\left(\frac{1}{n}\right).\end{aligned}$$

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

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi(t)\mathrm{e}^{-itx} \, \mathrm{d}t. \tag{16}$$

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

$$\operatorname{bias}(\hat{f}(x)) = \mathrm{E}[\hat{f}(x)] - f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}(\kappa(th)-1)\phi(t)\mathrm{e}^{-itx}\,\mathrm{d}t + O\left(\frac{1}{n}\right).$$

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

$$\operatorname{bias}(\hat{f}(x)) = \frac{1}{2\pi}\int_{|t|>1/h}(\kappa(th)-1)\phi(t)\mathrm{e}^{-itx}\,\mathrm{d}t + O\left(\frac{1}{n}\right).$$

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

$$|\operatorname{bias}(\hat{f}(x))| \le \frac{2}{2\pi}\int_{|t|>1/h}|\phi(t)|\,\mathrm{d}t + O\left(\frac{1}{n}\right).$$

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

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

If the bias is $o(h^r) + O\left(\frac{1}{n}\right)$ 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

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

This proves part (i).

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

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

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

$$O\left(\mathrm{e}^{\frac{-2d}{h}}\right) + O\left(\frac{1}{n^2}\right) = O\left(\mathrm{e}^{-2da \log n}\right) + O\left(\frac{1}{n^2}\right) = O\left(n^{-2da}\right) + O\left(\frac{1}{n^2}\right)$$

and a variance of

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

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

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

This proves part (ii).

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

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

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

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

which completes the proof of the theorem.

A.3 Proof of Theorem 2

Proof By taking the pth derivative on both sides of the identity (5), we have

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

By taking the pth derivative on both sides of the identity (16), we have

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

Following the steps in (15), we have

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

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

$$\text{bias}(\hat{f}_p(x)) = \frac{1}{2\pi}\int_{-\infty}^{\infty}(-it)^p(\kappa(th)-1)\phi(t)\mathrm{e}^{-itx}\,\mathrm{d}t + O\left(\frac{1}{n}\right).$$

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

$$|\,\text{bias}(\hat{f}_p(x))| \le \frac{2}{2\pi}\int_{|t|>1/h}|t|^p|\phi(t)|\,\mathrm{d}t + O\left(\frac{1}{n}\right).$$

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

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

If the bias is $o(h^r)+O\left(\frac{1}{n}\right)$ 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

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

Under assumption B,

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

Under assumption C,

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

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

A.4 Proof of Theorem 3

Proof The proof follows the proof of Theorem 3 in Berg and Politis (2010) with little modification.

A.5 Proof of Theorem 4

Proof Parts (ii) and (iii) follow from Theorems 1 and 2 and the δ-method. The convergence of $\hat{h}_{\mathrm{M}}$ in part (i) is dictated by the slowly converging $\widehat{f''}(x)$. However, the convergence rate of $\hat{h}_{\mathrm{M}}$ 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 Bühlmann (1996)] then Theorem 1 is still valid. If $|\phi(t)| \sim A|t|^{-d}$, then by Theorem 3,

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

From Theorem 2, part (i), if

$$\int_{-\infty}^{\infty} |t|^{r+2} |\phi(t)| < \infty, \tag{17}$$

then the bias of $\widehat{f''}(x)$ is $o(h^r)$. In order for (17) 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 4 is now proved.

References

Berg A, Politis DN (2009) Higher-order accurate polyspectral estimation with flat-top lag-windows. Ann Inst Stat Math 61(2):477–498

Berg A, Politis D (2010) CDF and survival function estimation with infinite-order kernels. Electron J Stat 3:1436–1454

Bühlmann P (1996) Locally adaptive lag-window spectral estimation. J Time Ser Anal 17(3):247–270

Cao R, López de Ullibarri I (2007) Product-type and presmoothed hazard rate estimators with censored data. Test 16(2):355–382

Chacón JE, Montanero J, Nogales AG (2007) A note on kernel density estimation at a parametric rate. J Nonparametr Stat 19:13–21

Davis KB (1977) Mean integrated square error properties of density estimates. Ann Stat 5(3):530–535

de Uña-Álvarez J, Meira-Machado L (2015) Nonparametric estimation of transition probabilities in the non-Markov illness-death model: a comparative study. Biometrics 71(2):364–375

Devroye L (1992) A note on the usefulness of superkernels in density estimation. Ann Stat 20(4):2037–2056

Földes A, Rejtő L, Winter BB (1981) Strong consistency properties of nonparametric estimators for randomly censored data, II: estimation of density and failure rate. Period Math Hung 12(1):15–29
Gijbels I, Wang JL (1993) Strong representations of the survival function estimator for truncated and censored data with applications. J Multivar Anal 47(2):210–229
Hall P, Marron JS (1991) Lower bounds for bandwidth selection in density estimation. Probab Theory Relat Fields 90(2):149–173
Hess K, Gentleman R (2014) muhaz: hazard function estimation in survival analysis. http://CRAN.R-project.org/package=muhaz, version 1.2.6 edition
Ibragimov IA, Khasminskii RZ (1983) Estimation of distribution density belonging to a class of entire functions. Theory Probab Appl 27(3):551–562
Jones MC (1993) Simple boundary correction for density estimation. Stat Comput 3:135–146
Karunamuni RJ, Alberts T (2005) On boundary correction in kernel density estimation. Stat Methodol 2(3):191–212
Lopez de Ullibarri I, Jácome MA (2013) survPresmooth: an R package for presmoothed estimation in survival analysis. J Stat Softw 54(11):1–26
Marron JS, Padgett WJ (1987) Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. Ann Stat 15(4):1520–1535
Marron JS, Ruppert D (1994) Transformations to reduce boundary bias in kernel density estimation. J R Stat Soc Ser B (Methodol) 56(4):653–671
McMurry TL, Politis DN (2004) Nonparametric regression with infinite order flat-top kernels. J Nonparametr Stat 16(3–4):549–562
McMurry TL, Politis DN (2017) iosmooth: functions for smoothing with infinite order flat-top kernels. https://CRAN.R-project.org/package=iosmooth, R package version 0.94 edition
Müller HG, Wang JL (1994) Hazard rate estimation under random censoring with varying kernels and bandwidths. Biometrics 50(1):61–76
Padgett WJ, McNichols DT (1984) Nonparametric density estimation from censored data. Commun Stat Theory Methods 13(13):1581–1611
Politis DN (2001) On nonparametric function estimation with infinite-order flat-top kernels. In: Charalambides Ch et al (eds) Probability and statistical models with applications. Chapman and Hall/CRC, Boca Raton, pp 469–483
Politis DN (2003) Adaptive bandwidth choice. J Nonparametr Stat 15(4–5):517–533
Politis DN, Romano JP (1993) On a family of smoothing kernels of infinite order. In: Tarter M, Lock M (eds) Computing science and statistics. Proceedings of the 25th symposium on the interface. The Interface Foundation of North America, Fairfax Station, pp 141–145
Politis DN, Romano JP (1995) Bias-corrected nonparametric spectral estimation. J Time Ser Anal 16(1):67–103
Politis DN, Romano JP (1999) Multivariate density estimation with general flat-top kernels of infinite order. J Multivar Anal 68(1):1–25
Ramlau-Hansen H (1983) Smoothing counting process intensities by means of kernel functions. Ann Stat 11(2):453–466
Sánchez-Sellero C, González-Manteiga W, Cao R (1999) Bandwidth selection in density estimation with truncated and censored data. Ann Inst Stat Math 51(1):51–70
Schuster EF (1985) Incorporating support constraints into nonparametric estimators of densities. Commun Stat A Theory Methods 14(5):1123–1136
Silverman BW (1986) Density estimation for statistics and data analysis. Monographs on statistics and applied probability. Chapman & Hall, London
Singh RS (1977) Improvement on some known nonparametric uniformly consistent estimators of derivatives of a density. Ann Stat 5(2):394–399
Suaray KN (2008) An alternate mean squared error computation for a censored data kernel density estimator. J Appl Probab Stat 3(2):287–297
Tanner MA, Wong WH (1983) The estimation of the hazard function from randomly censored data by the kernel method. Ann Stat 11(3):989–993
Tanner MA, Wong WH (1984) Data-based nonparametric estimation of the hazard function with applications to model diagnostics and exploratory analysis. J Am Stat Assoc 79(385):174–182
Wasserman L (2006) All of nonparametric statistics. Springer, New York
Watson GS, Leadbetter MR (1964) Hazard analysis I. Biometrika 51(1–2):175–184

Yakovlev AY, Tsodikov AD, Boucher K, Kerber R (2000) The shape of the hazard function in breast carcinoma. Cancer 85(8):1789–1798
Yandell BS (1983) Nonparametric inference for rates with censored survival data. Ann Stat 11(4):1119–1135
Zhou M (1988) Two-sided bias bound of the Kaplan–Meier estimator. Probab Theory Relat Fields 79(2):165–173