Prob. Th. Rel. Fields 85, 523–538 (1990)

Probability Theory and Related Fields

Locally adaptive hazard smoothing

Hans-Georg Müller* and Jane-Ling Wang**
Division of Statistics, University of California, Davis, CA 95616, USA

Received August 8, 1988; in revised form December 4, 1989

Summary. We consider nonparametric estimation of hazard functions and their derivatives under random censorship, based on kernel smoothing of the Nelson (1972) estimator. One critically important ingredient for smoothing methods is the choice of an appropriate bandwidth. Since local variance of these estimates depends on the point where the hazard function is estimated and the bandwidth determines the trade-off between local variance and local bias, data-based local bandwidth choice is proposed. A general principle for obtaining asymptotically efficient data-based local bandwiths is obtained by means of weak convergence of a local bandwidth process to a Gaussian limit process. Several specific asymptotically efficient bandwidth estimators are discussed. We propose in particular an asymptotically efficient method derived from direct pilot estimators of the hazard function and of the local mean squared error. This bandwidth choice method has practical advantages and is also of interest in the uncensored case as well as for density estimation.

1. Introduction

We consider the problem of local bandwidth choice for the nonparametric kernel estimation of a hazard function and its derivatives under censoring. Local bandwidth choice is of specific interest here since the variance of such an estimator depends critically on the number of observations available above the point where the hazard function is to be estimated. The variance of the hazard estimate goes to infinity as the last observation is approached; on the other hand it is well known that for the kernel method the bandwidth regulates the tradeoff between bias and variance (see Rosenblatt, 1956, 1971, and Parzen, 1962, for density estimation). Therefore, when estimating the hazard function or its derivatives, bandwidths should be chosen locally, adapting to the local Mean Squared Error (MSE). The estimation of derivatives of the hazard function is of interest e.g. in order to track maxima or minima via zeros of a derivative. Bandwidth choice for derivatives is especially delicate, since there the variance depends even more critically on the local bandwidth. We will show (Theorems 1 and

* Research supported by UC Davis Faculty Research Grant and by Air Force grant AFOSR-89-0386
** Research supported by Air Force grant AFOSR-89-0386

2) that locally adaptive bandwidth choice is indeed feasible and propose asymptotically efficient methods to achieve it. In particular, a pilot estimation approach is proposed, where in a first step estimators of finite bias and variance are found based on pilot estimates of relevant quantities. Then the minimizer of the resulting estimate of finite mean squared error is the estimated optimal bandwidth. This new method of bandwidth choice is not only efficient, but also practically feasible and applies as well to the density estimation problem, both in censored and uncensored cases.

As kernel estimator for the hazard function or its derivatives, we consider the convolution of the Nelson (1972) estimator with a kernel function (cf. (2.3)). When the target of interest is the hazard function itself, i.e. when $v=0$, properties of estimator (2.3), like asymptotic normality, were investigated by Ramlau-Hausen (1983), Tanner and Wong (1983) and Yandell (1983), using different techniques. Our analysis makes use of an asymptotic representation (2.8) of estimator (2.3) and is a consequence of an asymptotic representation of the Nelson estimator (cf. (2.6)) as a sum of i.i.d. random variables due to Lo and Singh (1986). A similar representation for a different hazard function estimate was studied by Lo, Mack and Wang (1989). Another representation for the kernel density estimate under censoring is given in Diehl and Stute (1988), and is used to establish the exact rates of pointwise and uniform convergence as well as asymptotic distributions. This representation together with the oscillation behavior of empirical processes was also applied by Stute (1985) to obtain a result similar to Theorem 1 for the case $v=0$ and $k=2$. Our approach is different, since the representation (2.8) allows us to take moments on the remainder term and therefore to compute actual variance and bias of the estimator (2.3) as well as to establish weak convergence of a bandwidth process to a Gaussian limiting process in a straightforward way. In addition, for $v=0$, the remainder term of our representation is $O\left(\frac{\log n}{nb}\right)$ a.s., whereas the remainder term in the above paper is $O\left(\frac{\log\log n}{(nb)^{1/2}}+(b\log\log n)^{1/2}\right)$ a.s., b being the bandwidth of the kernel estimator.

Efficient local bandwidth choice based on weak convergence was discussed for kernel density estimates by Abramson (1982) and Krieger and Pickands (1981) and for nonparametric kernel regression in the random design case by Mack and Müller (1987). Global bandwidth choice for density estimation under censoring was considered by Marron and Padgett (1987). A review on nonparametric density estimation under censoring is given in Padgett and McNichols (1984).

The paper is organized as follows: In Sect. 2, notations and kernel estimator are introduced, and the asymptotic representation of this estimator is derived. The main results on weak convergence of a bandwidth process and feasibility of local bandwidth choice are stated in Sect. 3. Some further remarks and discussions are compiled in Sect. 4, and proofs and auxiliary results are given in Sect. 5.

2. Preliminary results and asymptotic representation

In order to write down the estimator we introduce some notation.

We asume that $T_1, \ldots, T_n$ are i.i.d. lifetimes with distribution function (d.f.) F, and that $C_1, \ldots, C_n$ are i.i.d. censoring times with d.f. G. The C_i, T_i are assumed to be independent, and the actual observations are (X_i, δ_i), $i=1\dots n$, where $X_i=\min(T_i, C_i)$ and $\delta_i=1_{\{X_i=T_i\}}$ is an indicator of the censoring status of X_i. Then X_i are i.i.d. with d.f. L which satisfies $\bar{L}=\bar{F}\bar{G}$, where for any d.f. E, we denote by $\bar{E}=1-E$ the corresponding survival function. Further let $H(x)=-\log(\bar{F}(x))$ be the cumulative hazard function.

Our aim is to estimate $h(x)=H'(x)=f(x)/\bar{F}(x)$, the hazard function (where $f=F'$ is assumed to exist), or more generally some derivative $h^{(\nu)}(x)$, $\nu\geqq 0$, on an interval $[0, T]$, such that $L(T)<1$. A basic assumption we make for all of the following is:

$$\text{For some } k\geqq\nu, \qquad h\in\mathscr{C}^k([0, T]), \tag{2.1}$$

i.e. h is k times continuously differentiable on $[0, T]$.

Denote by $L_1(x)=P(X_i\leqq x, \delta_i=1)$ the subdistribution function for the uncensored observations, and by

$$L_{1n}(x)=\frac{1}{n+1}\sum_{i=1}^{n} 1_{\{X_i\leqq x, \delta_i=1\}},$$

the corresponding modified empirical subdistribution function. Similarly, denote the modified empirical distribution function of $L(x)$ by

$$L_n(x)=\frac{1}{n+1}\sum_{i=1}^{n} 1_{\{X_i\leqq x\}}.$$

Using the fact that

$$dL_1(t)/\bar{L}(t)=h(t), \tag{2.2}$$

the Nelson (1972) estimator of $H(x)$ is

$$H_n(x)=\int_0^x [1-L_n(y)]^{-1}\, dL_{1n}(y).$$

As estimator of $h^{(\nu)}(x)$ we consider the following kernel estimate, which is a convolution of the Nelson estimator H_n with an appropriate kernel function K_ν:

$$\hat{h}^{(\nu)}(x)=\frac{1}{b^{\nu+1}}\int K_\nu\left(\frac{x-u}{b}\right) dH_n(u)=\frac{1}{b^{\nu+1}}\sum_{i=1}^{n} K_\nu\left(\frac{x-X_{(i)}}{b}\right)\delta_{(i)}/(n-i+1). \tag{2.3}$$

Here, $X_{(i)}$ are the order statistics of X_i, and $\delta_{(i)}$ is the concomitant of $X_{(i)}$. Further, $b=b(n)$ is a sequence of bandwidths for which we require

$$b\to 0, \quad nb^{2\nu+1}\to\infty, \quad \frac{nb}{(\log n)^2}\to\infty, \quad \text{as } n\to\infty, \tag{2.4}$$

and K_ν is a kernel function of bounded variation and is of order (ν, k) with support $[-1, 1]$, i.e. satisfying

$$K_\nu \in \mathcal{M}_{\nu,k} = \left\{ f \in L^2([-1, 1]) \colon \int f(x)\, x^j \, dx = \begin{cases} (-1)^\nu \nu! & j=\nu \\ 0 & 0 \leqq j < k,\ j \neq \nu \\ \neq 0 & j=k \end{cases} \right\}. \tag{2.5}$$

When estimating near 0 or T, there usually will occur boundary effects due to the fact that the "effective support" $[x-b, x+b]$ of the kernel is not contained in $[0, T]$. These boundary effects can be avoided in the finite sample situation by modifying estimators (2.3) in such a way that near these end-points kernels with asymmetric supports are used in complete analogy to nonparametric regression (see, e.g., Rice, 1984); we will not go into the details here. Asymptotically, since $L(T) < 1$, we will not encounter boundary effects at T since $b \to 0$, but there are still such effects when estimating at 0. Therefore, we assume that for local considerations, $x \in (0, T]$, and uniform and global considerations will always be made with respect to a compact subset C in $(0, T]$.

Let

$$g(x) = \int_0^x (\bar{L}(x))^{-2} \, dL_1(s),$$

and for positive reals z and x, and $\delta = 0$ or 1, let

$$\xi(z, \delta, x) = g(\min(z, x)) - 1_{\{z \leqq x, \delta = 1\}} / \bar{L}(x).$$

Observe that

$$E(\xi(X_i, \delta_i, x)) = 0$$

and

$$\operatorname{cov}(\xi(X_i, \delta_i, s), \xi(X_i, \delta_i, t)) = g(\min(s, t)).$$

It follows from the proof of Theorem 1 of Lo and Singh (1986), that

$$H_n(x) - H(x) = \frac{1}{n} \sum_{i=1}^{n} \xi(X_i, \delta_i, x) + r_n(x), \tag{2.6}$$

where

$$\sup_{0 \leqq x \leqq T} |r_n(x)| = O\{(\log n/n)^{3/4}\} \text{ almost surely.}$$

The remainder $r_n(x)$ is further improved by Lo et al. (1989) to the order of $O(\log n/n)$ a.s. More specifically, for any $\beta > 0$ there exists $\alpha > 0$ such that

$$P\{\sup_{0 \leqq x \leqq T} |r_n(x)| > \alpha(\log n/n)\} = O(n^{-\beta}). \tag{2.7}$$

The following asymptotic representation is the base of our considerations.

Asymptotic representation. *There exists a decomposition*

$$\hat{h}^{(\nu)}(x)=h^{(\nu)}(x)+\beta_n(x)+\sigma_n(x)+e_n(x) \text{ a.s.,} \tag{2.8}$$

where

$$\beta_n(x)=\frac{1}{b^{\nu+1}}\int H(x-by)\,dK_\nu(y)-h^{(\nu)}(x) \tag{2.9}$$

is the bias,

$$\sigma_n(x)=\frac{1}{nb^{\nu+1}}\sum_{i=1}^{n}\int \xi(X_i,\delta_i,x-by)\,dK_\nu(y) \tag{2.10}$$

is the stochastic component of $\hat{h}^{(\nu)}(x)$, *and* $e_n(x)$ *is the remainder of the approximation, satisfying*

$$\sup_{0\leqq x\leqq T}|e_n(x)|=O\left(\frac{\log n}{nb^{\nu+1}}\right) \text{ a.s.}$$

and

$$E(|e_n(x)|^r)=O\left(\left(\frac{\log n}{nb^{\nu+1}}\right)^r\right) \quad \text{for } r=1,2. \tag{2.11}$$

Proof. The asymptotic representation (2.8) is a result of (2.6), (2.7) and integration by parts. Observe that ξ's are bounded random variables on $[0, T]$, H is bounded on $[0, T]$ and from the definition of H_n, $H_n(x)\leqq 1+\log n$ for all x. Therefore, (2.6) implies that $r_n(x)=O(\log n)$ for all x in $[0, T]$. This fact and (2.7) thus imply (2.11). □

A further assumption we make is

$$L\in\mathscr{C}([0, T]), \tag{2.12}$$

i.e. L is continuous on $[0, T]$. Then, writing

$$B_{\nu,k}=((-1)^k/k!)\int K_\nu(u)\,u^k\,du \tag{2.13}$$

and

$$V_{\nu,k}=\int (K_\nu(u))^2\,du, \tag{2.14}$$

the decomposition (2.8) provides the following asymptotic limit distribution for any x in $(0, T]$ (see Sect. 5 for details on the proof):

Lemma 1. *Assume that* $\nu<k$ *and that*

$$d=\lim_{n\to\infty} nb^{2k+1}$$

exists for some $0 \leqq d < \infty$. *Then*

$$(nb^{2\nu+1})^{1/2}(\hat{h}^{(\nu)}(x) - h^{(\nu)}(x)) \xrightarrow{D} \mathcal{N}(d^{1/2}\, h^{(k)}(x)\, B_{\nu,k},\ V_{\nu,k}\, h(x)/\bar{L}(x)). \qquad (2.15)$$

From (2.15) one can derive the asymptotic MSE of $\hat{h}^{(\nu)}(x)$. We assume that $h^{(k)}(x) \neq 0$ in all of the following, and will discuss this assumption further in Sect. 4. Minimizing the asymptotic MSE with respect to b, yields what we define as optimal local bandwidth $b^*(x)$. It should be noted that in the proof of Lemma 1, (5.1) to (5.3), the actual variance and bias of $\hat{h}^{(\nu)}(x)$ are computed, so that $b^*(x)$ also minimizes the direct MSE defined via moments (for the distinction between asymptotic and actual MSE compare, e.g., Lehmann, 1986, Chapt. 6.1). We find

$$b^*(x) = s^*(x)\, n^{-1/(2k+1)},$$

where

$$s^*(x) = \left[\frac{2\nu+1}{2(k-\nu)}\, \frac{h(x)}{\bar{L}(x)}\, \frac{V_{\nu,k}}{(h^{(k)}(x)\, B_{\nu,k})^2}\right]^{1/(2k+1)}. \qquad (2.16)$$

An obvious problem for data-adaptive local bandwidth variation is that s^* depends on the unknowns h, $h^{(k)}$ and L. This problem will be addressed in the next section.

3. Main results

Denote the kernel estimate (2.3) with local bandwidth $b(x) = s(x)\, n^{-1/(2k+1)}$ by

$$\hat{h}^{(\nu)}(x, s) = \frac{1}{[sn^{-1/(2k+1)}]^{\nu+1}} \sum_{i=1}^{n} K_\nu\left(\frac{x - X_{(i)}}{sn^{-1/(2k+1)}}\right) \delta_{(i)}/(n-i+1). \qquad (3.1)$$

We show in Theorem 1 that any local bandwidth estimator $\hat{s}(x)$ satisfying $\hat{s}(x) \to s^*(x)$ in probability is asymptotically efficient in the sense of Theorem 1 below.

The proof of Theorem 1 relies on the weak convergence of a bandwidth process which is of interest in its own right and is given in the following lemma. Here we choose s_a, s_b such that $0 < s_a < s^*(x) < s_b < \infty$. The proof is in Sect. 5.

Lemma 2. *If* K_ν *is Lipschitz continuous,* $k > \nu$, *then the processes*

$$\eta_n(x, s) = n^{(k-\nu)/(2k+1)}(\hat{h}^{(\nu)}(x, s) - h^{(\nu)}(x)), \qquad s \in [s_a, s_b],$$

for all $x \in (0, T]$ *converge weakly on* $\mathscr{C}([s_a, s_b])$ *to a Gaussian process* $\eta(x, s)$ *with*

$$E(\eta(x, s)) = s^{k-\nu}\, h^{(k)}(x)\, \frac{(-1)^k}{k!} \int K_\nu(u)\, u^k\, du$$

and

$$\operatorname{cov}(\eta(x,s_1),\eta(x,s_2))=\frac{1}{s_1^{\nu+1}s_2^{\nu+1}}\,\frac{h(x)}{\bar{L}(x)}\int K_\nu\left(\frac{u}{s_1}\right)K_\nu\left(\frac{u}{s_2}\right)du.$$

Theorem 1. *Assume that K_ν is Lipschitz continuous and that $k>\nu$. For any local bandwidth estimator $\hat{s}(x)$ satisfying $\hat{s}(x)\xrightarrow{P}s^*(x)$ it then holds, that asymptotically,*

$$\mathscr{L}(n^{(k-\nu)/(2k+1)}(\hat{h}^{(\nu)}(x,\hat{s})-h^{(\nu)}(x)))\approx\mathscr{L}(n^{(k-\nu)/(2k+1)}(h^{(\nu)}(x,s^*)-h^{(\nu)}(x)))$$
$$\approx\mathcal{N}\left(s^{*\,k-\nu}h^{(k)}(x)\,B_{\nu,k},\ \frac{h(x)}{\bar{L}(x)\,s^{*\,2\nu+1}}\,V_{\nu,k}\right),$$

i.e., the hazard function estimator $\hat{h}^{(\nu)}(x,\hat{s}(x))$ has the same normal limit distribution as an estimator supplied with the optimal bandwidths, $\hat{h}^{(\nu)}(x,s^*(x))$.

Proof. According to Slutsky's theorem and Lemma 1 it remains to show that

$$n^{(k-\nu)/(2k+1)}(\hat{h}^{(\nu)}(x,\hat{s})-\hat{h}^{(\nu)}(x,s^*))\to 0 \text{ in probability.}$$

This follows by a standard argument from Lemma 2, since the limiting process is continuous. □

One way to construct a consistent estimate $\hat{s}$ of s^* is to estimate h, $h^{(k)}$ and L consistently in (2.16). Here the following result is helpful, which for $\nu<k$ is a consequence of Lemma 1.

Lemma 3. *Under the assumptions of Lemma* 1, *it holds for any $x\in(0,T]$*:

$$\textit{If } d>0 \textit{ and } \nu<k,\ \hat{h}^{(\nu)}(x)\xrightarrow{P}h^{(\nu)}(x). \tag{3.2}$$

$$\textit{If } nb^{2k+1}\to\infty, \text{and } \frac{nb^{k+1}}{\log n}\to\infty\ \ \hat{h}^{(k)}(x)\xrightarrow{P}h^{(k)}(x). \tag{3.3}$$

Proof. We observe that $d>0$ implies that $nb^{2\nu+1}\to\infty$, from which (3.2) follows via (5.1) and (5.3) in Sect. 5. In case that $\nu=k$, the variance of the estimator is $O(1/(nb^{2k+1}))$, and the bias (through continuity of $h^{(k)}$) is $o(1)$, which implies (3.3). □

Since obviously $\bar{L}_n(x)\to\bar{L}(x)$ in probability, we consequently have available a consistent estimator $\hat{s}(x)$ of $s^*(x)$ as

$$\hat{s}(x)=\left[\frac{2\nu+1}{2(k-\nu)}\,\frac{\hat{h}(x)}{\bar{L}_n(x)}\,\frac{V_{\nu,k}}{(\hat{h}^{(k)}(x)B_{\nu,k})^2}\right]^{1/(2k+1)}. \tag{3.4}$$

The data-adaptive procedure (3.4) is asymptotically efficient but might encounter some difficulties in practical application owing to the fact that $h^{(k)}(x)$ has to be estimated; estimators of $h^{(k)}(x)$ might be unstable for small sample sizes. Asymptotically, however, this would not cause any problem.

Another method which avoids estimation of derivatives is to estimate in a first step, for $\nu=0$, bias and variance directly, using a consistent pilot estimator of h and then in case that $v>0$ to apply in a second step the "factor method" described in Müller, Stadtmüller and Schmitt (1987), which relates optimal bandwidths for $v=0$ and for $v>0$ by a factor as explained below in (3.9). In order to estimate the bias (as a function of b) for $v=0$ (see (2.9)),

$$\beta(x, b)=\frac{1}{b}\int H(x-by)\,dK_0(y)-h(x),$$

where x is in a compact subset of $(0, T]$, we propose

$$\hat{\beta}(x, b)=\int \hat{h}(x-by)\,K_0(y)\,dy-\hat{h}(x). \tag{3.5}$$

For the variance we use the analogous estimate

$$\hat{v}(x, b)=\frac{1}{nb}\int K_0(y)^2\,\frac{\hat{h}(x-by)}{\bar{L}_n(x-by)}\,dy. \tag{3.6}$$

Note that $\hat{\beta}$, $\hat{v}$ can be readily evaluated numerically by discrete approximations to these convolution integrals, e.g., application of the fast Fourier transform would allow for fast computation. If in these estimators $\hat{h}$ and $\bar{L}_n$ are replaced by the true values, they yield closer approximations to the leading convolution integrals defining variance and bias than the asymptotically leading terms in the limit $n\to\infty$ as given in (2.15). Our proposed local bandwidth estimate for $v=0$ is then obtained as minimizer of

$$\mathrm{M\hat{S}E}(x, b)=\hat{v}(x, b)+\hat{\beta}^2(x, b) \tag{3.7}$$

with respect to b, which we denote by $\hat{b}_0$.

In order to obtain the local bandwidth estimate in case that $v>0$, we apply a kernel K_0 of order $(0, k)$ to estimate $h(x)$ and a kernel K_v of order (v, k) to estimate $h^{(v)}(x)$. For the corresponding optimal local bandwidths

$$b_j^*(x)=\left[\frac{2j+1}{2(k-j)}\,\frac{h(x)}{\bar{L}(x)}\,\frac{V_{v,k}}{(h^{(k)}(x)\,B_{v,k})^2\,n}\right]^{1/(2k+1)}, \quad j=0, v, \tag{3.8}$$

it holds according to (2.16) that:

$$b_v^*(x)=b_0^*(x)\left(\frac{k(2v+1)}{(k-v)}\,\frac{V_{v,k}}{V_{0,k}}\,\frac{B_{0,k}^2}{B_{v,k}^2}\right)^{1/(2k+1)}=b_0^*(x)\,d_{v,k}, \tag{3.9}$$

where $B_{0,k}$ and $V_{0,k}$ are defined as in (2.13) and (2.14).

The factor $d_{v,k}$ on the right hand side of (3.9) depends only on v, k and the kernel functions K_0, K_v and is therefore known. It follows, writing $b_v^*=s_v^*\,n^{-1/(2k+1)}$, $b_0^*=s_0^*\,n^{-1/(2k+1)}$, that $\hat{s}_0\to s_0^*$ in probability implies $\hat{s}_0\,d_{v,k}\to s_v^*$ in probability, if appropriate kernels are chosen. Therefore it satisfies for the current proposal to consider local bandwidth choice for $v=0$. Once this leads to an efficient procedure, an efficient procedure for $v>0$ will also be available. It remains to be shown that procedure (3.7) for $v=0$ indeed leads to efficient curve estimates in the sense of Theorem 1.

Theorem 2. *Let $\tilde{b}$ be a bandwidth sequence satisfying* (2.4) *and*

$$n\tilde{b}^{2k+2} \to \infty \tag{3.10}$$

and assume that $C \subset (0, T]$ is a compact set and that $x \in C$ is given. Assume that $\hat{h}$ in (3.5), (3.6) *employs bandwidth $\tilde{b}$ and a kernel of order $\mathcal{M}_{0,k}$ which is k times continuously differentiable, and that if $\nu > 0$, $\hat{h}^{(\nu)}$ is to be estimated using a kernel of $\mathcal{M}_{\nu,k}$. Let $\hat{b}_0(x)$ be the minimizer of* (3.7) *on an interval $[b^{(1)}, b^{(2)}]$, where $b^{(1)}, b^{(2)}$* satisfy (2.4), *and define* (*see* (3.9))

$$\hat{b}_\nu(x) = \hat{b}_0(x)\, d_{\nu,k} = \hat{s}_\nu(x)\, n^{-1/(2k+1)}. \tag{3.11}$$

Then it holds that

$$\hat{s}_\nu(x) \xrightarrow{P} s_\nu^*(x),$$

i.e., this method of local bandwidth choice for $\hat{h}^{(\nu)}$ leads to asymptotically efficient estimates.

The proof is postponed to Sect. 5. It is based on a uniform convergence result for kernel derivative estimates of hazard functions (Lemma 4), which is given in Sect. 5.

4. Discussion

1. It should be noted that although our proposed method (3.7), (3.11) of local bandwidth choice depends on pilot estimators $\hat{h}$ for which a preliminary bandwidth has to be employed, choice of this preliminary bandwidth is not critical as long as this bandwidth is somewhat oversmoothing according to requirement (3.10). Our method of local bandwidth choice can also be adapted for density estimation and nonparametric regression; for fixed design nonparametric regression, compare Müller (1985).

For instance, in the case of density estimation with or without censoring, we can consider local bandwidth choice for the estimator

$$\hat{f}(x) = \frac{1}{b} \int K\left(\frac{x-u}{b}\right) dF_n(u),$$

where F_n is the empirical distribution function in the uncensored case and the Kaplan-Meier estimator in case of censoring. Results analogous to Theorem 1 and 2 can be obtained for this situation, where our estimators for asymptotic bias and variance are

$$\tilde{\beta}(x, b) = \int \hat{f}(x - by) K(y)\, dy - \hat{f}(x), \qquad \tilde{v}(x, b) = \frac{1}{nb} \int K(y)^2 \hat{f}(x - by)\, dy.$$

Here again, $\hat{f}$ is a pilot estimator of f. The local bandwidth estimator is then defined as the minimizer of

$$\tilde{\mathrm{MSE}}(x, b) = \tilde{v}(x, b) + \tilde{\beta}^2(x, b),$$

as before, and this method will be efficient under analogous conditions as in Theorem 2.

2. In our analysis we assumed that at the point x where the hazard function h is to be estimated, it holds that $h^{(k)}(x) \neq 0$. This assumption is necessary in order to obtain the leading term for the bias and to obtain a proper minimizer of the mean squared error, and therefore the optimal rate of convergence. This is also reflected by the fact that $h^{(k)}(x)$ appears in the denominator of the optimal bandwidth s^* (2.16). In case that $h^{(k)}(x)=0$, according to (5.1), the bias $\beta_n(x) = o(b^{k-\nu})$, while the variance (5.3) remains unchanged (The same is true if a kernel of an order higher than k is used). Then asymptotic choice of $b \sim n^{-1/(2k+1)}$ yields $\mathrm{MSE} = o(n^{-2(k-\nu)/(2k+1)})$. Although Lemma 2 remains valid, the theorems do not and an asymptotic MSE minimizing bandwidth cannot be identified. However, it is easy to see that (5.9) still holds, so that consistent estimators of MSE will remain available, and these can be minimized to define finite sample bandwidth choice. Then, assuming b_0 is the (possibly non-unique) minimizer of MSE, and $\hat{b}$ the minimizer of $\widehat{\mathrm{MSE}}$, such that $b_0 \in [b^{(1)}, b^{(2)}]$, it follows from (5.9) that

$$\frac{\widehat{\mathrm{MSE}}(\hat{b})}{\mathrm{MSE}(b_0)} \leqq \frac{\widehat{\mathrm{MSE}}(b_0)}{\mathrm{MSE}(b_0)} = 1 + o_p(1),$$

which provides some asymptotic motivation for this procedure even in case that $h^{(k)}(x)=0$.

3. Among asymptotically efficient bandwidth estimators there could still be differences, in particular with respect to the finite sample behavior. For instance, an alternative to our proposed estimator $\hat{b}_0$ which was defined as minimizer of estimates of the finite MSE (3.7) is method (3.4), which is obtained by minimizing estimates of the leading terms of asymptotic MSE (2.15) and was also shown to be asymptotically efficient. If we only admit k continuous derivatives for h, the quotients [(estimated bandwidth)/(optimal bandwidth)] are $(1+o_p(1))$ for both methods, and the $o_p(1)$ term cannot be improved; the reason is that then $\hat{h}^{(k)}(x) \to h^{(k)}(x)$ in probability uniformly, without the possibility of assessing the rate over this class of functions. However, apart from these asymptotic considerations, it is clear that estimator (3.7) seems to be close to minimizing the finite MSE, since the approximation step involved in the asymptotic arguments leading to, e.g., (3.4) is avoided. Further, this method only requires estimators of h itself and not of $h^{(k)}$. Especially when higher order kernels ($k>2$) are used, this is a clear advantage.

4. The algorithmic implementation of our proposed method still requires a number of choices to be made. These concern $\tilde{b}$, the pilot bandwidth (see Theorem 2) and b^1, b^2, the bounds between which a minimizer is sought.

For the choice of $\tilde{b}$, a number of options are available, and it will depend on the specific situation which of these will be the most appropriate one. Once $\tilde{b}$ is found, it is recommended to choose $b^{(1)} = 0.5\tilde{b}$, $b^{(2)} = 2.0\tilde{b}$. This choice proved to be satisfactory in nonparametric regression. It is also advisable to smooth the chosen local bandwidths $\hat{b}(x)$ over x, if global curve estimates with local bandwidth choices are desired, since otherwise random fluctuations in $\hat{b}(x)$ will give rise to oscillations in estimated curves, with the possible effect of disturbing the aesthetic appeal of the estimated curve.

For the choice of $\tilde{b}$, we recommend the following options:

a) Application of a global bandwidth choice criterion like (modified) cross-validation; such a criterion was discussed for the estimation of densities under censoring by Marron and Padgett (1988).

b) Choose a parametric pilot by assuming initially that the survival times follow a common model, say exponential or Weibull. A maximum likelihood fit of such a model would then first be obtained. Then the fitted model acts as $\tilde{h}(\cdot)$ in (3.7). Comparison of this pilot estimator with the final kernel estimator can also be applied to check goodness-of-fit of the pilot model.

c) Another possibility is to find $\tilde{b}$ such that $\tilde{b}$ and the derived $\tilde{b}_0$ do not differ much. This would indicate that $\tilde{b}$ is already close to optimal. Such "fixed points" might be obtained by iterating the procedure.

It should be noted that according to Theorem 2, some oversmoothing in the pilot estimator is desirable, i.e., $\tilde{b}$ should rather be chosen larger than optimal. For instance, method b) above would achieve this, since parametric pilots in general would be smoother than the actual hazard functions. Theorem 2 also shows that efficiency can be achieved for a large class of pilot bandwidths $\tilde{b}$. This, of course, does not mean that choice of $\tilde{b}$ would not matter for the practical performance of the method. Which of the considered specific algorithms will prove most successful remains to be investigated. A limitation for bandwidth choice to keep in mind for practical purposes are boundary effects. The larger the chosen bandwidth is, the more boundary effects occur.

5. Auxiliary results and proofs

In this section we provide proofs for some of the results in Sects. 2, 3 and some auxiliary results which might be of independent interest.

Proof of Lemma 1. By a standard argument used repeatedly in curve estimation, compare, e.g., Rosenblatt (1971), we obtain, by a Taylor expansion, the bias (2.9) for $\nu<k$:

$$\beta_n(x)=b^{k-\nu}\{h^{(k)}(x)\,B_{\nu,k}+o(1)\}, \tag{5.1}$$

as long as $0<x\leqq T$. In the end point 0, (5.1) is not valid due to boundary effects, which have been discussed earlier.

Next we evaluate the variance of $\hat{h}^{(\nu)}(x)$. Observing $\operatorname{var}(\hat{h}^{(\nu)}(x))=\operatorname{var}(\sigma_n(x))+\operatorname{var}(e_n(x))+2\operatorname{cov}(\sigma_n(x),\,e_n(x))$, we obtain for the individual terms:

$$\begin{aligned}\operatorname{var}(\sigma_n(x))=E\{\sigma_n^2(x)\}&=\frac{1}{nb^{2(\nu+1)}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\min(x-ub,\,x-vb))\,dK_\nu(u)\,dK_\nu(v)\\&=\frac{1}{nb^{2(\nu+1)}}\int_{-\infty}^{\infty}\int_{-\infty}^{x-tb} K_\nu\left(\frac{x-u}{b}\right)dg(u)\,dK_\nu(v)\\&=\frac{1}{nb^{2(\nu+1)}}\int_{-\infty}^{\infty}\left\{K_\nu\left(\frac{x-u}{b}\right)\right\}^2 dg(u)\\&=\frac{1}{nb^{2(\nu+1)}}\int_{-\infty}^{\infty}(K_\nu(v))^2\{h(x-bv)/\bar{L}(x-bv)\}\,dv,\end{aligned}$$

where the last equality follows from the definition of $g(\cdot)$ and (2.2). Continuity of $h/\bar{L}$ at x implies that

$$\operatorname{var}(\sigma_n(x)) = \frac{1}{nb^{2\nu+1}}\left\{\frac{h(x)}{\bar{L}(x)} V_{\nu,k} + o(1)\right\}. \tag{5.2}$$

Further, $\operatorname{var}(e_n(x)) < E(e_n^2(x))$, which can be bounded via (2.11). The bandwidth condition then implies that $\operatorname{var}(e_n(x))$ is of smaller order than $\operatorname{var}(\sigma_n(x))$ and by the Cauchy-Schwarz inequality applied to the covariance term we have

$$\operatorname{var}(\hat{h}^{(\nu)}(x)) = \frac{1}{nb^{2\nu+1}}\left\{\frac{h(x)}{\bar{L}(x)} V_{\nu,k} + o(1)\right\}. \tag{5.3}$$

The asymptotic normality of $\hat{h}^{(\nu)}$ follows immediately from the representation (2.8) and the central limit theorem. The above results on the variance (5.3) and on the asymptotic bias (5.1) then imply Lemma 1. □

Proof of Lemma 2. We consider the stochastic processes

$$\zeta_n(s) = n^{(k-\nu)/(2k+1)}(\hat{h}^{(\nu)}(x, s) - E(\hat{h}^{(\nu)}(x, s))), \qquad s \in [s_a, s_b].$$

Observe that $\zeta_n \in \mathscr{C}([s_a, s_b])$.

Evaluating the covariance structure of these processes we obtain, writing $\sigma_n(x, s_i)$, $e_n(x, s_i)$ in obvious analogy to $\hat{h}^{(\nu)}(x, s_i)$, for $s_i \in [s_a, s_b]$, $i = 1, 2$:

$$\begin{aligned}
\operatorname{cov}(\zeta_n(s_1), \zeta_n(s_2)) &= n^{2(k-\nu)/(2k+1)}[\operatorname{cov}(\sigma_n(x, s_1), \sigma_n(x, s_2)) \\
&\quad + \operatorname{cov}(\sigma_n(x, s_1), e_n(x, s_2)) + \operatorname{cov}(e_n(x, s_1), \sigma_n(x, s_n)) \\
&\quad + \operatorname{cov}(e_n(x, s_1), e_n(x, s_2))] = n^{2(k-\nu)/(2k+1)}(\mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}).
\end{aligned}$$

We observe that, writing $b_i = s_i n^{-1/(2k+1)}$, $i = 1, 2$, and using the definition of g and (2.2),

$$\begin{aligned}
I &= E(\sigma_n(x, s_1)\, \sigma_n(x, s_2)) \\
&= \frac{1}{nb_1^{\nu+1} b_2^{\nu+1}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\min(x - ub_1, x - vb_2)\, dK_\nu(u)\, dK_\nu(v) \\
&= \frac{1}{nb_1^{\nu+1} b_2^{\nu+1}} \int_{-\infty}^{\infty} \int_{-\infty}^{x - vb_2} K_\nu\left(\frac{x-u}{b_1}\right) dg(u)\, dK_\nu(v) \\
&= \frac{1}{nb_1^{\nu+1} b_2^{\nu+1}} \int_{-\infty}^{\infty} K_\nu\left(\frac{v}{s_1}\right) K_\nu\left(\frac{v}{s_2}\right) \frac{h(x - vn^{-1/(2k+1)})}{\bar{L}(x - vn^{-1/(2k+1)})}\, dv \\
&= n^{-2(k-\nu)/(2k+1)}(s_1^{\nu+1} s_2^{\nu+1})^{-1}\left\{\frac{h(x)}{\bar{L}(x)} \int_{-\infty}^{\infty} K_\nu\left(\frac{u}{s_1}\right) K_\nu\left(\frac{u}{s_2}\right) du + o(1)\right\}.
\end{aligned} \tag{5.4}$$

Further, by (2.11),

$$\begin{aligned}\mathrm{II} &\leqq \{\operatorname{var}(\sigma_n(x, s_1))\, E|e_n(x, s_2)|^2\}^{1/2}\\ &= \left\{O(n^{-2(k-\nu)/(2k+1)})\, O\left(\left(\frac{\log n}{n b_2^{\nu+1}}\right)^2\right)\right\}^{1/2}\\ &= o(n^{-2(k-\nu)/(2k+1)}) \text{ if } k>\nu.\end{aligned}$$

The same bound holds for III, IV, and therefore

$$\operatorname{cov}(\zeta_n(s_1), \zeta_n(s_2)) = \frac{1}{s_1^{\nu+1} s_2^{\nu+1}} \frac{h(x)}{L(x)} \left\{ \int_{-\infty}^{\infty} K_\nu\left(\frac{u}{s_1}\right) K_\nu\left(\frac{u}{s_2}\right) du + o(1)\right\}. \tag{5.5}$$

By similar arguments as those leading to Lemma 1 and by the Cramer-Wold device, we obtain then for any $s_1, \ldots, s_m \in [s_a, s_b]$:
$[\zeta_n(s_1), \zeta_n(s_2), \ldots, \zeta_n(s_m)] \to \mathcal{N}(O, C)$ in distribution, where

$$C = (c_{ij})_{1 \leqq i, j \leqq m}, \qquad c_{ij} = \frac{h(x)}{L(x)\, s_i^{\nu+1} s_j^{\nu+1}} \int K_\nu\left(\frac{u}{s_i}\right) K_\nu\left(\frac{u}{s_j}\right) du. \tag{5.6}$$

Now turning to tightness of processes ζ_n, observe that, according to (2.8)–(2.11) (with obvious notation),

$$\zeta_n(s) = n^{(k-\nu)/(2k+1)}(\sigma_n(x, s) + e_n(x, s) - E e_n(x, s)),$$

and

$$\sup_{s_a \leqq s \leqq s_b} n^{(k-\nu)/(2k+1)}(|e_n(x, s)| + E e_n(x, s)) = o_p(1),$$

so that it remains to investigate tightness of $\psi_n(s) = n^{(k-\nu)/(2k+1)} \sigma_n(x, s)$. Now according to the derivation of (5.4), using the Lipschitz continuity of K_ν,

$$\begin{aligned}E(\psi_n(s_1) - \psi_n(s_2))^2 &\leqq c_1 \int \left[\frac{h(x - v n^{-1/(2k+1)})}{L(x - v n^{-1/(2k+1)})}\right]\left[\frac{1}{s_1^{\nu+1}} K_\nu\left(\frac{v}{s_1}\right) - \frac{1}{s_2^{\nu+1}} K_\nu\left(\frac{v}{s_2}\right)\right]^2 dv\\ &\leqq c_2 (s_1 - s_2)^2.\end{aligned}$$

Tightness of the processes ψ_n follows from Billingsley (1968), Theorems 6.2, 12.3 and formula (12.51). Lemma 2 thus follows by applying in addition (5.1) in order to evaluate the expectation. □

For the proof of Theorem 2, we observe that we only have to consider $\nu = 0$ and first derive a result on uniform convergence of hazard function estimates.

Lemma 4. *Under the assumptions of Theorem 2, using a kernel K of $\mathcal{M}_{0,k}$, which is k times continuously differentiable,*

$$\sup_{x \in C} |\hat{h}^{(j)}(x) - h^{(j)}(x)| = o_p(1), \qquad \textit{for } 0 \leqq j \leqq k.$$

Proof. We observe that on any compact set C in $(0, T]$, replacing ν by j, the bias β_n (5.1) can be uniformly bounded by $O(b^{k-j})$ under the assumptions made on bandwidth h and kernel K_j for the j-th derivative. Restriction to a compact subset of the interval $(0, T]$ is necessary due to the previously mentioned boundary effects. According to (2.8), it remains to investigate $\sup_{x\in C}|\sigma_n(x)+e_n(x)|$. From (2.8) and (2.9) we conclude that

$$\begin{aligned}\sup_{x\in C}|\sigma_n(x)+e_n(x)| &= \sup_{x\in C}\left|\hat{h}^{(j)}(x)-\frac{1}{b^{\nu+1}}\int H(x-by)\,dK_j(y)\right| \\ &= \sup_{x\in C}\left|\frac{1}{b^{j+1}}\int [H_n(u)-H(u)]\,dK_j\left(\frac{x-u}{b}\right)\right| \\ &< \frac{1}{b^{j+1}}\sup_{x\in C}|H_n(x)-H(x)|\,V(K_j)=\frac{1}{b^{j+1}}\,O_p(n^{-1/2}),\end{aligned}$$

according to Theorem 4 of Breslow and Crowley (1974), where $V(K_j)$ is the total variation of K_j. Lemma 4 now follows from the conditions imposed on the bandwidth. □

Proof of Theorem 2. Since by the Glivenko-Cantelli Lemma

$$\sup_{x\in C}|\bar{L}_n(x)-\bar{L}(x)|=o_p(1),$$

and by Lemma 4

$$\sup_{x\in C}|\hat{h}(x)-h(x)|=o_p(1),$$

we have (writing $\hat{h}(x, b)$ for the estimate employing local bandwidth b at x),

$$\frac{1}{nb}\int K_0^2(y)\left[\frac{\hat{h}(x-by)}{\bar{L}_n(x-by)}-\frac{h(x-by)}{\bar{L}(x-by)}\right]dy=o_p\left(\frac{1}{nb}\right),$$

and therefore by (5.2), (5.3) and (5.4),

$$\hat{v}(x, b)=\operatorname{var}(\hat{h}(x, b))(1+o_p(1)), \tag{5.7}$$

where the o_p-term is uniform on intervals $[b^{(1)}, b^{(2)}]$. Further,

$$\begin{aligned}&\sup_{x,y\in C}|(\hat{h}^{(k)}(y)-h^{(k)}(y))-(\hat{h}^{(k)}(x)-h^{(k)}(x))| \\ &\quad\leqq \sup_{y\in C}|\hat{h}^{(k)}(y)-h^{(k)}(y)|+\sup_{x\in C}|\hat{h}^{(k)}(x)-h^{(k)}(x)|=o_p(1),\end{aligned}$$

and therefore

$$\begin{aligned}&\int(\hat{h}(x-by)-h(x-by))K_0(y)\,dy-(\hat{h}(x)-h(x)) \\ &\quad=b^k\int K_0(y)\,y^k(\hat{h}^{(k)}(x-by)-h^{(k)}(x-by))\,dy \\ &\quad=b^k(\hat{h}^{(k)}(x)-h^{(k)}(x))\int K_0(y)\,y^k\,dy(1+o_p(1)) \\ &\quad=o_p(b^k),\ \text{uniformly in } b\in[b^{(1)}, b^{(2)}].\end{aligned}$$

It follows that

$$\hat{\beta}(x, b) = (E\hat{h}(x, b) - h(x))(1 + o_p(1)), \tag{5.8}$$

and (5.7), (5.8) imply that

$$\hat{\mathrm{MSE}}(x, b) = \mathrm{MSE}(x, b)(1 + o_p(1)), \tag{5.9}$$

where $o_p(1)$ is uniform in $b \in [b^{(1)}, b^{(2)}]$. Defining

$$g(x, s) = s^{2k}(h^{(k)}(x)\, B_{0,k})^2 + \frac{1}{s}\frac{h(x)}{L(x)} V_{0,k}$$

and

$$b(s_n) = s_n n^{-1/(2k+1)},$$

we observe that according to (5.1) and (5.3),

$$n^{2k/(2k+1)} \mathrm{MSE}(x, b(s_n)) = g(x, s_n)(1 + o(1)),$$

where $o(1)$ is uniform over sequences $b(s_n)$ such that $b^{(1)} \leqq b(s_n) \leqq b^{(2)}$.

Since $g(x, s)$ achieves a unique minimum at s^* and is strictly convex, it follows that given $\varepsilon > 0$, there exist $s_\ell < s^* < s_u$ such that for sufficiently large n, the minimizer $\hat{s}$ satisfies $P(\hat{s} \in [s_\ell, s_u]) > 1 - \varepsilon$. Owing to the uniformity of the $o(1)$- and $o_p(1)$-term,

$$\sup_{s \in [s_l, s_u]} |n^{2k/(2k+1)} \mathrm{MSE}(x, b(s)) - g(x, s)| = o_p(1),$$

and it follows by standard arguments that $\hat{s}(x) \xrightarrow{P} s^*(x)$. □

References

Abramson, I.: Arbitrariness of the pilot estimator in adaptive kernel methods. J. Multivariate Anal. **12**, 562–567 (1982)

Billingsley, P.: Weak convergence of probability measures. New York: Wiley 1968

Breslow, N.E., Crowley, J.: A large sample study of the life table and product limit estimates under random censorship. Ann. Stat. **2**, 437–453 (1974)

Diehl, S., Stute, W.: Kernel density estimation in the presence of censoring. J. Multivariate Anal. **25**, 299–310 (1988)

Krieger, A.M., Pickands, J. III.: Weak convergence and efficient density estimation at a point. Ann. Stat. **9**, 1066–1078 (1981)

Lehmann, E.L.: Theory of point estimation. New York: Wiley 1983

Lo, S.H., Mack, Y.P., Wang, J.L.: Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. Probab. Th. Rel. Fields **80**, 461–473 (1989)

Lo, S.H., Singh, K.: The product-limit estimator and the bootstrap: some asymptotic representations. Probab. Th. Rel. Fields **71**, 455–465 (1986)

Mack, Y.P., Müller, H.G.: Adaptive nonparametric estimation of a multivariate regression function. J. Multivariate Anal. **23**, 169–182 (1987)

Marron, J.S., Padgett, W.J.: Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. Ann. Stat. **15**, 1520–1535 (1987)

Müller, H.G.: Empirical bandwidth choice for nonparametric kernel regression by means of pilot estimators. Stat. Decis. [Suppl.] **2**, 193–206 (1985)

Müller, H.G., Stadtmüller, U., Schmitt, T.: Bandwidth choice and confidence intervals for derivatives of noisy data. Biometrika **74**, 743–750 (1987)

Nelson, W.: Theory and applications of hazard plotting for censored failure data. Technometrics **14**, 945–966 (1972)

Padgett, W.J., McNichols, D.T.: Nonparametric density estimation from censored data. Commun. Stat. Theory Methods **13**, 1581–1611 (1984)

Parzen, E.: On estimation of a probability density and mode. Ann. Math. Stat. **35**, 1065–1076 (1962)

Ramlau-Hansen, H.: Smoothing counting process intensities by means of kernel functions. Ann. Stat. **11**, 453–466 (1983)

Rice, J.: Boundary modification for nonparametric kernel regression. Commun. Stat. Theory Methods **13**, 893–900 (1984)

Rosenblatt, M.: Remarks on some nonparametric estimators of a density function. Ann. Math. Stat. **27**, 832–837 (1956)

Rosenblatt, M.: Curve estimates. Ann. Math. Stat. **42**, 1815–1842 (1971)

Stute, W.: Optimal bandwidth selection in pointwise density and hazard rate estimation when censoring is present. Technical Report, University of Giessen 1985

Tanner, M., Wong, W.H.: The estimation of the hazard function from randomly censored data by the kernel method. Ann. Statist. **11**, 989–993 (1983)

Yandell, B.S.: Nonparametric inference for rates with censored survival data. Ann. Stat. **11**, 1119–1135 (1983)