Ann Inst Stat Math (2011) 63:1019–1046
DOI 10.1007/s10463-010-0277-6

Local linear hazard rate estimation and bandwidth selection

Dimitrios Bagkavos

Received: 23 February 2009 / Revised: 27 May 2009 / Published online: 23 February 2010

Abstract A new kernel based local linear estimate of the hazard rate, under the random right censorship model is proposed in this article. We study its finite sample and asymptotic properties and prove its asymptotic normality. Then we bring in three popular methods for bandwidth selection to the hazard setting as potential bandwidth choice rules for the estimate. We discuss their practical implementation and through Monte Carlo simulations we use four distributions with different hazard rate shapes to compare their performance over various sample sizes and levels of censoring.

Keywords Hazard rate · Bandwidth selection · Kernel · Censored data · Asymptotic normality

1 Introduction

The hazard rate function is formally defined as

$$\lambda(x) = \lim_{dx \to 0+} \frac{\mathrm{P}(x \leq X < x + dx | x \leq X)}{dx} \tag{1}$$

and expresses the probability that an item with lifetime $X > 0$ will experience an event which is the primary interest of the study, at most once, in the interval $(x, x + dx)$ given that no such event occurred up to time x. Naturally, estimation of the hazard rate function is a concept of major importance in survival analysis as it has many

D. Bagkavos (✉)
Information Resources International, Rostoviou 17,
11526 Athens, Greece
e-mail: dimitrios.bagkavos@gmail.com

applications in fields as diverse as reliability, medicine, demography and insurance, to name but a few.

Frequently in survival analysis it is not possible to observe a complete random sample from the X population. An effective model for such situations is the right censorship model under which a proportion of the lifetimes available are censored on the right by another random variable, possibly from a different distribution than that of X and which is associated with some other risk factor than that associated with the X variable. Exact mathematical formulation is given in Sect. 2.

The shape of the hazard rate helps in apprehending the mechanism that affects survival and so, when no shape constrains are directly imposed by the data, nonparametric methods turn out to be particularly useful in estimating the true curve. In this context, kernel based estimation of the hazard rate function has received particular attention in the literature. A summary of early studies on this topic can be found in Gefeller and Michels (1992). Since Jones (1993) introduced the local linear fitting method to density estimation (see also Fan and Gijbels 1996 for a more general treatment primarily in the regression context) that alleviates the underestimation problem of kernel estimates at the boundary, a substantial number of research has been focused on applications of the technique in the hazard setting and a small overview is given later in this section.

The purpose of this article is to develop an estimate of the hazard rate under the right censorship model and the local linear framework, using as basis an empirical estimate of the hazard rate. This can be viewed as an extension of the work of Bagkavos and Patil (2008) to censored data and the same motivation principle for taking this approach apply in the present work as well. Additionally, we bring in and explore the performance of three bandwidth choice methods when applied on hazard rate estimation. The bandwidth parameter is crucial because it controls the amount of local averaging to be performed and therefore affects the quality of estimation. This can be seen by the fact that it is present in the bias and variance expressions of any kernel based estimate. The methods considered in this article, under the same motivation principle with the density setting, are the empirical bias bandwidth selection (EBBS) method of Ruppert (1997), which models the Mean Squared Error (MSE) empirically and selects the bandwidth that minimizes the MSE locally in a neighborhood of the point of estimation, the plug-in method of Cheng (1997) which uses local linear and cubic fitting to develop a bandwidth selector that minimizes the Mean Integrated Squared Error (MISE) and the method of Hurvich et al. (1997) which selects bandwidth on the basis of the Aikaike's Information Criterion (AIC) minimization.

The local linear principle was first put to use in the hazard setting in the works of Müller et al. (1997), Wang et al. (1998) and Nielsen (1998a). The first two papers developed local linear estimates of the hazard rate from transformed aggregated lifetable data, while Nielsen (1998a) extended the Nadaraya-Watson type local constant estimator of Linton and Nielsen (1995) to the local linear case much as it is known from regression, notably in the multivariate setting first. Nielsen and Tanggaard (2001) developed a general class of hazard rate estimates based on local fitting and under a counting processes framework which facilitates broad patterns of data, not necessarily i.i.d. and quite general censoring mechanisms. Nielsen (2003) under the same framework compared local linear and conventional hazard rate estimators via an extensive simulation study. In applied work the method has also known wide usability. Apart from the works of Müller et al. (1997) and Wang et al. (1998) which were initiated having in mind the analysis of oldest-old mortality data from lifetables, Fledelius et al. (2004) in analyzing old mortality data first argued that the use of local linear fitting as alternative to parametric models is valuable as the choice of a parametric model in this and in similar situations tends to be rather difficult. Moreover, in the same paper it was first suggested that the theoretical properties of the method can be expanded if combined with bias reduction techniques such as the multiplicative bias correction method of Nielsen (1998b). See also Bagkavos and Patil (2008), Nielsen (1998a) and the references therein for applications in reliability and medicine related problems. The wide acceptance of the method initiated further theoretical developments. Mammen and Nielsen (2007) extended the method in the multi-dimensional case while estimating old age mortality with a calendar effect. There, it is demonstrated that parallel to the univariate case the technique's nice properties are preserved in the multivariate setting. Moreover it was identified that especially when combined with bias correction techniques, the excellent properties of the method prove to be very advantageous in the frequent sparse data situations of old age mortality estimation. On another direction, Nielsen et al. (2009) developed the density version of the work of Nielsen and Tanggaard (2001) and identified the need for development of suitable for practical use bandwidth selectors.

It has to be noted that the estimate studied in this article can be viewed as a special case of the local linear estimate of Nielsen and Tanggaard (2001). Specifically, under the i.i.d. data case and the right censorship model, estimator $\hat{\alpha}_{2,W}$ of Nielsen and Tanggaard (2001) is equivalent to the estimate proposed here, given though with a different mathematical formulation. From this perspective, this work extends the results on the asymptotic properties of estimator $\hat{\alpha}_{2,W}$ to the present setup. This work is also related to the works of Müller et al. (1997) and Wang et al. (1998) in the sense that an estimate of similar structure is studied there. Our approach is different in that we estimate the hazard rate function directly from continuously generalizes the asymptotic results of Wang et al. (1998).

The contribution of this article is that it develops a kernel local linear hazard rate estimate under the random right censorship model so that no boundary adjustments are required, provides its finite sample and asymptotic properties and establishes its asymptotic normality. Further, it extends three bandwidth choice methods in the hazard setting and derives neat expressions for their implementation. Finally, simulation evidence is given on the performance of the estimate when using the proposed bandwidth selectors.

We note here that the present research on bandwidth selection also has the prospective to be extended to the multivariate setting as it has been exhibited by Mammen and Nielsen (2007). This would be of particular importance to mortality estimation which is often of multivariate nature and since this would enable incorporation of calendar-time effects and allow observation of trends over time.

The rest of the article is organized as follows. First, in Sect. 2 we define an estimate of the hazard rate based on the local linear method, we study its finite sample and asymptotic properties and prove its asymptotic normality. In Sect. 3 we consider three bandwidth selections the methods and give details for their implementation and in

Sect. 4 we study their performance via simulations and discuss the results. Proofs of all theorems as well as auxiliary lemmas are given in Sect. 5.

2 Local linear hazard rate estimation

Let $T_1, T_2, \ldots, T_N$ be a sample of i.i.d. survival times censored on the right by i.i.d. random variables $U_1, U_2, \ldots, U_N$, which are independent from the T_i's. Let f_T be the common probability density and F_T the distribution function of the T_i's. Denote by H the distribution function of the U_i's. Typically the randomly right censored observed data are denoted by the pairs (X_i, δ_i), $i = 1, 2, \ldots, N$ with $X_i = \min\{T_i,\ U_i\}$ and $\delta_i = 1_{\{T_i \leq U_i\}}$ where $1_{\{\cdot\}}$ is the indicator random variable of the event $\{\cdot\}$. The distribution function of X_i's is $1 - F = (1 - F_T)(1 - H)$. By the definition of conditional probability the hazard rate function (1) can be written as

$$\lambda(x) = \frac{f_T(x)}{1 - F_T(x)}, \quad F_T(x) < 1,$$

on an interval $[0, T]$ of the real line, with $T = \sup\{x : F_T(x) < 1 - \varepsilon\}$ for a small $\varepsilon > 0$. Partition the interval into n disjoint subintervals $\{I_i, i = 1 \ldots n\}$ of equal length $\Delta = T/n$ and denote with $x_i = (i - \frac{1}{2})\Delta$, $i = 1 \ldots n$, the center of the interval I_i. A natural estimate of the hazard rate is

$$c_j = \frac{d_j}{\Delta n_j}, \quad j = 1, 2, \ldots, n$$

with

$$d_j = \sum_{i=1}^{N} 1_{\{X_i \in I_j, \delta_i = 1\}}, \quad n_j = \sum_{i=1}^{N} 1_{\{X_i > \Delta(j-1)\}}$$

because c_j is an empirical estimate of

$$q(x_j) = \mathrm{P}(X \in (x_j, x_j + \Delta) | X > x_j), \quad j = 1, 2, \ldots, n$$

which if Δ is small should be close to $\lambda(x_j)$. Selection of Δ in practice is discussed in Sect. 4. From theorem A.1, (i) and (ii) in Wang et al. (1998) we have that $\mathbb{E}c_j \approx \lambda(x_j)$ and $\mathbb{V}\mathrm{ar}(c_j) \approx \lambda(x_j)(1 - F(x_j))^{-1}$. By taking the bin centers x_j as the design points, the data (c_j, x_j), $j = 1, 2, \ldots, n$ are approximately independent and therefore the hazard rate estimation problem could be regarded as a nonparametric regression problem of the form $c_j = m(x_j) + \sigma(x_j)\varepsilon_j$, $j = 1, \ldots, n$ where $m(x_j) = \lambda(x_j)$ and $\sigma(x_j) = \lambda(x_j)(1 - F(x_j))^{-1}$. Here we handle this problem by following the local linear fitting approach. According to standard local polynomial regression theory, consider

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \{c_i - \beta_0 - \beta_1(x_i - x)\}^2 K_h(x_i - x), \tag{2}$$

where

$$K_h(u) = h^{-1} K\left(u/h\right)$$

is a kernel function which assigns weight to each point and h, usually called *bandwidth*, controls the size of the local neighborhood. Let $\hat{b}_\mu(x)$, $\mu = 0, 1$, be the solution of (2). Then, a natural estimate of $\lambda(x)$ is $\hat{b}_0(x)$. Define

$$T_{n,l}(x) = \sum_{i=1}^{n} c_i K\left(\frac{x_i - x}{h}\right)(x_i - x)^l, \quad l = 0, 1,$$

$$S_{n,l}(x) = \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)(x_i - x)^l, \quad l = 0, 1, 2,$$

$$R_{n,l}(x) = \sum_{i=1}^{n} K^2\left(\frac{x_i - x}{h}\right)(x_i - x)^l, \quad l = 0, 1, 2,$$

$$L_{n,l}(x) = \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)(x_i - x)^l \lambda(x_i), \quad l = 0, 1, 2.$$

Then, by standard calculations, see for example Bagkavos and Patil (2008), the estimate of the hazard function is

$$\hat{b}_0(x) \equiv \hat{\lambda}_L(x) = \frac{T_{n,1}(x)S_{n,1}(x) - T_{n,0}(x)S_{n,2}(x)}{S_{n,1}(x)S_{n,1}(x) - S_{n,0}(x)S_{n,2}(x)}.$$

In order to assess the practical performance of $\hat{\lambda}_L(x)$ we need the asymptotic properties of the estimator. For that, we need the following definitions and assumptions. Let

$$s_{j,\cdot} = \int_{-\cdot}^{+\infty} u^j K(u)\, du, \quad j = 0, 1, 2 \tag{3}$$

and

$$K_{(\cdot)}(u) = \frac{s_{2,\cdot} - s_{1,\cdot} u}{s_{2,\cdot} s_{0,\cdot} - s_{1,\cdot}^2} K(u) I_{[-\cdot,+\infty]}(u).$$

Throughout this paper we assume that the hazard function $\lambda(x)$ and its first two derivatives are bounded and that the kernel function K is supported on $[-1, 1]$ with

$$\int K^2 < +\infty, \quad \int |u^2 K| < +\infty, \quad \int K = 1 \quad \text{and} \quad \int uK = 0.$$

The finite sample properties of $\hat{\lambda}_L(x)$ will be needed in Sect. 3 and are given in the following theorem.

Theorem 1 *The bias and variance expressions of $\hat{\lambda}_L(x)$ are given by*

$$\mathbb{E}\hat{\lambda}_L(x) = \frac{S_{n,1}(x)L_{n,1}(x) - S_{n,2}(x)L_{n,0}(x)}{S_{n,1}^2(x) - S_{n,2}(x)S_{n,0}(x)},$$

$$\mathbb{V}\text{ar}\left\{\hat{\lambda}_L(x)\right\} = \frac{\lambda(x)}{1-F(x)} \times \frac{S_{n,2}^2(x)R_{n,0}(x) - 2S_{n,2}(x)S_{n,1}(x)R_{n,1}(x) + S_{n,1}^2(x)R_{n,2}(x)}{(S_{n,1}^2(x) - S_{n,2}(x)S_{n,0}(x))^2}.$$

It is difficult to appraise the finite sample properties of $\hat{\lambda}_L(x)$. For this reason the asymptotic properties of $\hat{\lambda}_L(x)$ are summarized in the following theorem.

Theorem 2 *Suppose that for $l = 0, 1, 2$, $K^{(l)}$ is bounded and absolutely integrable with finite second moments and that $h \to 0$, $Nh \to +\infty$, $\Delta/h \to 0$. Then, if x is away from the boundary,*

$$\mathbb{E}\left\{\hat{\lambda}_L(x)\right\} - \lambda(x) = \frac{h^2}{2}\lambda''(x)\int u^2K(u)\,du + o\left(h^2\right),$$

$$\mathbb{V}\text{ar}\left\{\hat{\lambda}_L(x)\right\} = \frac{1}{Nh}\frac{\lambda(x)}{1-F(x)}\int K^2(u)\,du + o\left((Nh)^{-1}\right).$$

If x is a boundary point, that is, $x = ph$, $p \geq 0$, then,

$$\mathbb{E}\left\{\hat{\lambda}_L(x)\right\} - \lambda(x) = \frac{h^2}{2}\lambda''(x)\int_{-p}^{+\infty} u^2K_{(p)}(u)\,du + o\left(h^2\right),$$

$$\mathbb{V}\text{ar}\left\{\hat{\lambda}_L(x)\right\} = \frac{1}{Nh}\frac{\lambda(x)}{1-F(x)}\int_{-p}^{+\infty} K_{(p)}^2(u)\,du + o\left((Nh)^{-1}\right).$$

Remark 1 The asymptotic properties of estimator $\hat{\lambda}_L(x)$ in theorem 2 show that the estimate achieves automatic corrections in the boundary while in the interior the estimate behaves like a conventional kernel estimate (e.g., as estimator of $\hat{\lambda}_T(x)$ of Tanner and Wong 1983). This can be seen also by the fact that when x is an interior point, $x = ph$, $p \geq 1$ and thus $s_{0,p} = 1$, $s_{1,p} = 0$. Hence, $K_{(p)}(u) = K(u)$.

Remark 2 In the special case where the censoring random variable has all its mass at ∞ then $H(x) = 0$, $F_T(x) = F(x)$ for all x in $[0, T]$ and the estimate becomes estimator $\hat{\lambda}_L(x)$ of Bagkavos and Patil (2008). In this sense, theorem 2 generalizes theorem 1 of Bagkavos and Patil (2008) in the right censored data setting.

The asymptotic distribution of $\hat{\lambda}_L(x)$ is a useful asset as for example it can be used in construction of confidence intervals and hypothesis tests. It is given in the form of the following theorem.

Theorem 3 *Assuming that K is compatible with both F_T and H, $\hat{\lambda}_L(x)$ has an asymptotic normal distribution as $N \to +\infty$, $h \to 0$, $Nh \to +\infty$.*

For implementation of a kernel estimate in real life applications, one has to specify the bandwidth parameter h and the kernel function K. Of the two, as it has been widely argued in the literature, more important is choice of h. In the next section we investigate three methods for choosing h.

3 Bandwidth selection

In this section we discuss in detail practical implementation of $\hat{\lambda}_L(x)$ by extending three bandwidth selection rules to the hazard setting. We then highlight the necessary changes on these rules so that they can be applied on the local linear multiplicative bias corrected estimate of Nielsen and Tanggaard (2001). The methods we extend are the 'solve the equation' plug-in method of Cheng (1997), the bandwidth selection by the Akaike information criterion minimization of Hurvich et al. (1997) and the empirical bias bandwidth selection method of Ruppert (1997).

We start with the plug-in rule. From theorem 2 it is easy to see that the Mean Integrated Square Error (MISE) of $\hat{\lambda}_L(x)$ is

$$\begin{aligned}\mathrm{MISE}(\hat{\lambda}_L(x)) = \frac{1}{4}h^4\mu_2^2(K)\theta_{2,2} + \frac{1}{Nh}R(K)\int \frac{\lambda(x)}{1-F(x)}\,dx \\ + o\left(\frac{1}{Nh}\right) + O\left(h^4\right),\end{aligned}$$

where $\mu_2(K)$ is the second moment of the kernel K and

$$R(g(x)) = \int g^2(x)\,dx, \quad \theta_{\mu,\nu} = \int \lambda^{(\mu)}(x)\lambda^{(\nu)}(x)\,dx$$

for positive integers μ, ν. Asymptotically, as $N \to +\infty$, $h \to 0$ and $Nh \to +\infty$ the MISE can be well approximated by the AMISE (asymptotic MISE) which is defined as

$$\mathrm{AMISE}(\hat{\lambda}_L(x)) = \frac{1}{4}h^4\mu_2^2(K)\theta_{2,2} + \frac{1}{Nh}R(K)\int \frac{\lambda(x)}{1-F(x)}\,dx.$$

The minimum, with respect to h, value of AMISE is attained at

$$h_{\mathrm{AMISE}} = \left\{\frac{R(K)M}{N\mu_2^2(K)\theta_{2,2}}\right\}^{\frac{1}{5}} \tag{4}$$

with M being

$$M = \int_0^T \frac{\lambda(x)}{1-F(x)}\,dx = \int_0^T \frac{f_T(x)}{(1-F(x))(1-F_T(x))}\,dx.$$

As an estimate of M we use

$$\hat{M} = \int_0^T \frac{\hat{\lambda}_L(x)}{1 - \hat{F}(x)}\, dx$$

where

$$\hat{F}(x) = \frac{1}{N} \sum_{i=1}^{N} \frac{\delta_i}{H^*(X_i)} W\left(\frac{x - X_i}{h}\right), \quad W(x) = \int_{-\infty}^{x} K(t)\, dt$$

and $\hat{H}^*$ is an estimate of $1 - H$, typically taken to be the Kaplan-Meier estimator, slightly modified in order to avoid division by zero, i.e.,

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

with (Z_i, Λ_i) being the ordered (X_i, δ_i), $i = 1, \ldots, n$. Moreover, the quantity $\theta_{2,2}$ in (4) is unknown and therefore has to be estimated. This is a typical problem of plug-in methods and a usual way to proceed, is to reapply local fitting to obtain an estimate of the functional. Bagkavos and Patil (2008) developed such an estimate based on local cubic fitting and binning approximations and studied its asymptotic properties. The estimate can be readily used here and is given by

$$\tilde{\theta}_{2,2} = \int (\tilde{\theta}_2(x))^2 dx \simeq \frac{1}{n} \sum_{i=1}^{n} (\tilde{\theta}_2(x_i))^2,$$

where, with a denoting the bandwidth used,

$$\begin{aligned} \tilde{\theta}_2(x) &= \frac{2}{a^3} \sum_{i=1}^{n} K_2\left(\frac{x_i - x}{a}\right) c_i, \\ K_2(u) &= e_3^T S^{-1} (1, u, u^2, u^3)^T K(u), \end{aligned} \tag{5}$$

$S = (S_{i+j-2}(x))_{1 \le i, j \le 4}$, $e_3^T = (0, 0, 1, 0)$ and

$$S_i = \int_{-\infty}^{+\infty} x^i K(x)\, dx, \quad i = 0, \ldots, 6.$$

We note here that kernel estimators of this type were discussed in the density setting and without boundary adjustment by Hall and Marron (1987) and

Jones and Sheather (1991) among others. Hall and Marron (1987) argued that estimates such as $\tilde{\theta}_{2,2}$ should be used without the diagonal terms incorporated (the terms resulting for the same summation indices in applying the definition of $\tilde{\theta}_2(x)$ on $\tilde{\theta}_{2,2}$) as these are deterministic and can be thought of as adding needless bias. Jones and Sheather (1991) showed how these terms can be used in a beneficial manner and we now illustrate that their arguments apply also for the bandwidth selector we use here. From theorem 2 of Bagkavos and Patil (2008), the squared bias of $\tilde{\theta}_{2,2}$ is of order $N^{-2}a^{-10}$ while variance is of order $N^{-2}a^{-9}$. Therefore bias dominates variance and this means that bandwidth selection can be based only on bias reduction. Now, from (5) and the kernel assumptions of Sect. 2 we have that $R(K_2) > 0$. Combining this with the bias equation of $\tilde{\theta}_{2,2}$ in Theorem 2 of Bagkavos and Patil (2008) we conclude that bias due to using the diagonal elements is positive, as opposed to the smoothing bias term $o(a^2)$ which is negative. Therefore, one can choose a so that bias cancels all together asymptotically by taking

$$a_{\text{AMSE}} = \left\{ \frac{24\chi M R(K_2)}{N\theta_{2,4}\mu_4(K_2)} \right\}^{\frac{1}{7}}, \tag{6}$$

with

$$\chi = \begin{cases} -1 & \text{if } \theta_{2,4} < 0 \\ 5/2 & \text{if } \theta_{2,4} > 0. \end{cases}$$

which gives an optimal bandwidth of $O(N^{-1/7})$. Given the order of the variance, this leads to a minimum MSE of order $N^{-5/7}$. Now, following the proof of theorem 2 of Bagkavos and Patil (2008) it is straightforward to prove that a version of $\tilde{\theta}_{2,2}$, say $\tilde{\theta}^*_{2,2}$ with the diagonal entries removed, has bias expression

$$\mathbb{E}\left\{\tilde{\theta}^*_{2,2}\right\} - \theta_{2,2} = \frac{a^2}{6}\theta_{2,4}\mu_4(K_2) + O\left(N^{-1}a^{-4}\right) + o\left(a^2\right).$$

The variance expression of $\tilde{\theta}^*_{2,2}$ is the same with that of $\tilde{\theta}_{2,2}$. The leading bias term above and the variance order lead to an MSE optimal bandwidth of order $N^{-2/11}$ which, in turn, leads to a minimum MSE of order $N^{-4/11}$. Hence the MSE of $\tilde{\theta}_{2,2}$ compares better to the MSE of $\tilde{\theta}^*_{2,2}$ and this advocates for its use when implemented with a such that its MSE is minimized.

The 'solve the equation' principle starts by a relationship between h and a which from (4) and (6) is $a_{\text{AMSE}} = C(K)D(\theta)h^{5/7}_{\text{AMISE}}$ with

$$C(K) = \left\{ \frac{24R(K_2)\mu_2^2(K)}{R(K)\mu_4(K_2)} \right\}^{1/7}, \quad D(\theta) = \left\{ \frac{\chi\theta_{2,2}}{\theta_{2,4}} \right\}.$$

Let

$$a(h) = C(K)D(\theta)h^{5/7}.$$

Then, the AMISE optimal bandwidth is the solution with respect to h of the equation

$$h_= \left\{ \frac{R(K)\hat{M}}{N\mu_2^2(K)\tilde{\theta}_{2,2}(a(h))} \right\}^{\frac{1}{5}}. \tag{7}$$

When no analytic solution to (7) is feasible, one may use a numerical procedure such a the Newton-Raphson method. Still (7) depends on $D(\theta)$ which in practice is unknown because it contains $\theta_{2,2}$ and $\theta_{2,4}$. According to the conventional 'solve the equation' approach one would go a stage further and apply local polynomial fitting for estimation of $\theta_{2,4}$ before using a reference to a parametric model. However, such an approach is subject to inherit large amount of variability from the data which results in the bandwidth selector to become unstable. Moreover it requires computations of inverses of 6×6 matrices and thus in a considerable decrease in speed. For these two reasons we use a parametric reference at this stage; this has been also advocated by Cheng (1997). Since using an exponential distribution, which admits root-n consistent parameter estimates is not suitable here, in absence of (even partial) information for the underlying density, as a default rule we suggest a two parameter Weibull distribution with the parameters estimated by maximum likelihood. The choice of this particular model is justified by its wide use in survival analysis and by its flexibility as it can mimic the behavior of other distributions such as the Rayleigh and the normal. Therefore, as an estimate of $D(\theta)$ we use $\chi\hat{\theta}_{2,2}\hat{\theta}_{2,4}^{-1}$ with

$$\hat{\theta}_{2,4} = \int (k(k-1)(k-2))^2(k-3)(k-4)\rho^8(\rho x)^{k-8}\,dx$$
$$\hat{\theta}_{2,2} = \int k(k-1)(k-2)\rho^3(\rho x)^{k-3}\,dx$$

where k, ρ are the scale and location parameters of the Weibull model estimated by solving, with respect to k and ρ, the system of (3.18) and (3.19), page 41 in Cox and Oakes (1984). If there is information for the underlying density then this rule should be adjusted accordingly. Thus the suggested bandwidth results from solving (7) for h after substituting the estimate of $D(\theta)$ in the definition of $a(h)$.

We now turn our attention to the Akaike's Information Criterion (AIC) bandwidth selection method which was first introduced for this purpose in the literature in Hurvich et al. (1997). In the case of local linear hazard rate estimation the objective is to minimize, with respect to h, the function

$$\mathrm{AIC}(h) = \log\{\mathrm{RSS}\} + \frac{n + \mathrm{tr}(S)}{n - [\mathrm{tr}(S) + 2]} \tag{8}$$

with

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(c_i - \hat{\lambda}_L(x_i)\right)^2$$

and

$$\mathrm{tr}(S) = K(0) \sum_{i=1}^{n} \frac{S_{n,2}(x_i)}{S_{n,2}(x_i)S_{n,0}(x_i) - S_{n,1}^2(x_i)}.$$

Minimization of (8) over h, is done in the interval $(0, X_{(N)})$, where $X_{(N)}$ denotes the largest sample observation. As the optimization technique, one may use a nonlinear minimization function subject to Box constrains. In the next section (Sect. 4) this is implemented by using the function `nlminb` of S-plus. As pointed out in Hurvich et al. (1997) such an approach may not always return the true minimum, and for this reason one may try to run the minimization function several times changing the starting values. However, this did not seem to be a problem in producing the simulation results and illustrations of the next section. As an alternative though, we have also experimented with minimizing (8) by a grid search again in the interval $(0, X_{(N)})$. The results showed that this approach tends to oversmooth the resulting estimate and moreover it is considerably slower. For this reason in Sect. 4 we focus and report only results generated with the nonlinear minimization approach.

Empirical bias bandwidth selection (EBBS) was initiated in Ruppert (1997). The idea behind the method is to minimize the MSE over h by empirical estimation of the squared bias, without involving asymptotic expressions and the variance as a conditional finite sample expression. To estimate the bias of $\hat{\lambda}_L$ at x, first we have to define neighboring values of h which is a bandwidth at which the bias of $\hat{\lambda}_L(x)$ will be estimated. We confine each neighborhood to contain $m > 1$ bandwidth values. Then, formally the neighborhood of h can be written as the equally spaced points $(h_1, \dots, h_m)$. m as well as the lower and upper bandwidth values are specified by the user. However, in Sect. 4 we use $m = 40$, $h_1 = 0$ and $h_m = X_{(N)}$. Calculating $\hat{\lambda}_L(x)$ at each h_i, yields $\hat{\lambda}_L(x; h_i)$, $i = 1, 2, \dots, m$. The next step is to use the data pairs $(\hat{\lambda}_L(x; h_i), h_i)$, $i = 1, 2, \dots, m$ to fit a polynomial of order $p + t$ by ordinary least squares regression. p is the order of the polynomial used to model $\hat{\lambda}_L(x)$ and t is the number of error terms in a Taylor expansion for the bias. In our case $p = 1$ and $t = 1$ which leads to

$$\hat{\lambda}_L(x; h) \simeq \gamma_0(x) + \gamma_2(x)h^2. \tag{9}$$

Then, the bias of $\hat{\lambda}_L(x; h)$ is estimated by $\hat{\gamma}_2(x)h^2$ with $\hat{\gamma}_2(x)$ being the estimate of $\gamma_2(x)$ found by the regression model. As an estimate of the variance at each h_i, $i = 1, \dots, m$ we use its finite sample form as this is given in theorem 1, with the unknown quantities replaced by their estimates calculated at each h_i, i.e.,

$$\mathbb{V}\mathrm{ar}\left\{\hat{\lambda}_L(x; h_i)\right\} = \frac{\hat{\lambda}_L(x; h_i)}{1 - \hat{F}(x; h_i)}$$
$$\times \frac{S_{n,2}^2(x; h_i)R_{n,0}(x; h_i) - 2S_{n,2}(x; h_i)S_{n,1}(x; h_i)R_{n,1}(x; h_i) + S_{n,1}^2(x; h_i)R_{n,2}(x; h_i)}{(S_{n,1}^2(x; h_i) - S_{n,2}(x; h_i)S_{n,0}(x; h_i))^2}.$$

In the formula above, similarly to $\hat{\lambda}_L(x; h_i)$ the quantities $S_{n,\cdot}(x; h_i)$, $R_{n,\cdot}(x; h_i)$ and $F(x; h_i)$ denote the quantities $S_{n,\cdot}(x)$, $R_{n,\cdot}(x)$ and $F(x)$ calculated with bandwidth h_i. For each h_i in $(h_1, \ldots, h_m)$ the procedure is applied to all bin centers $x_1, \ldots, x_n$ and the resulting MSE's are averaged across each h_i. The selected bandwidth is the one that corresponds to the first local minimum of the averaged MSEs. This approach corresponds to the global bandwidth option, as this was suggested in Ruppert (1997). The reason for taking this route is that in the fixed design set up the EBBS procedure is applied on the partitioned data pairs (c_j, x_j), $j = 1, 2, \ldots, n$ which typically is of 'small' size.

It is feasible to extend these three bandwidth rules to more sophisticated estimators such as the multiplicative bias corrected local linear hazard estimate introduced in Nielsen and Tanggaard (2001). This can be very useful for practitioners as this estimate is frequently used in mortality estimation, see for example Fledelius et al. (2004). The estimate is motivated by the formula $\lambda(x) = \hat{\lambda}(x) g_m(x)$ with $g_m(x) = \lambda(x)\hat{\lambda}(x)^{-1}$ being the multiplicative error of the pilot estimate $\hat{\lambda}(x)$. Here, as pilot estimate we use $\hat{\lambda}_L(x)$. Employing the local linear technique as in Nielsen and Tanggaard (2001) yields an estimate of the multiplicative error g_m given by

$$\hat{g}_m(x) = \frac{T^*_{n,1}(x) S^*_{n,1}(x) - T^*_{n,0}(x) S^*_{n,2}(x)}{S^*_{n,1}(x) S^*_{n,1}(x) - S^*_{n,0}(x) S^*_{n,2}(x)}$$

with

$$T^*_{n,l}(x) = \sum_{i=1}^{n} c_i K\left(\frac{x_i - x}{a}\right)(x_i - x)^l \hat{\lambda}_L(x_i), \quad l = 0, 1,$$
$$S^*_{n,l}(x) = \sum_{i=1}^{n} K\left(\frac{x_i - x}{a}\right)(x_i - x)^l \hat{\lambda}^2_L(x_i), \quad l = 0, 1, 2.$$

Then the multiplicative bias corrected local linear estimate is defined as $\hat{\lambda}_m(x) = \hat{\lambda}_L(x)\hat{g}_m(x)$. Implementation of $\hat{\lambda}_L(x)$ in the definition of $\hat{\lambda}_m(x)$ and $\hat{\lambda}_L(x_i)$ in the definitions of $T^*_{n,l}(x)$ and $S^*_{n,l}(x)$ is done with the same bandwidth h. Hence it remains to extend the three rules to selection of a. Since many specific parts of the three rules will be common in both determining h and a, we confine ourselves here in highlighting the major differences.

Starting with the plug-in rule, from theorem 3 of Nielsen and Tanggaard (2001) the AMISE of $\hat{\lambda}_m(x)$ is

$$\text{AMISE}\left(\hat{\lambda}_m(x)\right) = \frac{1}{16} a^8 \mu_2^4 R\left(\left(\lambda''(x)\lambda^{-1}(x)\right)''\right) + (nh)^{-1} R(T_k) \int \frac{\lambda(x)}{1 - F(x)}\, dx$$

with $T_k(u) = 2K(u) - (K * K)(u)$, where $*$ denotes convolution, is a fourth order kernel obtained by twicing, see also Nielsen and Tanggaard (2001). Minimizing with respect to a and substituting the unknown quantities by their estimates as in the $\hat{\lambda}_L(x)$

case the AMISE optimal bandwidth becomes

$$a_{\text{AMISE}} = \left(8R(T_k)\hat{M} \left\{ N\mu_2^4 R\left(\left(\tilde{\theta}_2(x)\hat{\lambda}_L^{-1}(x) \right)'' \right) \right\}^{-1} \right)^{-\frac{1}{9}}$$

and bandwidth selection for $\tilde{\theta}_2$ is identical as in with the $\tilde{\theta}_{2,2}$ case before. The AIC bandwidth selection criterion extends in a straightforward way for estimator $\hat{g}_m(x)$. The selected bandwidth is the one that minimizes (8), but now

$$\text{RSS} = \sum_{i=1}^{n} \left(c_i \hat{\lambda}_L^{-1}(x_i) - \hat{g}_m(x_i) \right)^2 \tag{10}$$

and

$$\text{tr}(S) = K(0) \sum_{i=1}^{n} \frac{S^*_{n,2}(x_i)}{S^*_{n,2}(x_i)S^*_{n,0}(x_i) - S^*_{n,1}(x_i)S^*_{n,1}(x_i)}. \tag{11}$$

Finally, for the EBBS rule, it suffices to give the bias and variance formulas for MSE estimation as the other steps remain unchanged. The asymptotic bias of $\hat{g}_m$ is

$$\mathbb{E}\hat{g}_m(x) = 1 - \frac{h^2}{2}\mu_2 \frac{\lambda''(x)}{\lambda(x)} + O(h^4)$$

and so the bias of $\hat{g}_m(x)$ can be modeled as in (9) with $\hat{\lambda}_L(x; h)$ replaced by $\hat{g}_m(x; a)$, which denotes $\hat{g}_m(x)$ calculated with bandwidth a. Finally the finite sample variance can be estimated at each bandwidth to be tested by

$$\mathbb{V}\text{ar}\left\{ \hat{g}_m(x) \right\} \approx \frac{\text{RSS}}{n - \sum_{i=1}^{n} (2S(x_i) - Q(x_i))}$$

where RSS is given by (10), $S(x_i)$ is the summand of (11) and Q is defined by

$$Q(x) = \frac{S^{*\,2}_{n,2}(x)R_{n,0}(x) - 2S^*_{n,2}(x)S^*_{n,1}(x)R_{n,1}(x) + S^{*\,2}_{n,1}(x)R_{n,2}(x)}{(S^*_{n,2}(x)S^*_{n,0}(x) - S^*_{n,1}(x)S^*_{n,1}(x))^2}.$$

Ruppert et al. (1995) used an estimate of similar nature in the regression setting; see Turlach and Wand (1996) for a general way of constructing variance estimates. The reason for using such an estimate here is to indicate a possible alternative in extensions of the EBBS method on estimators that their finite sample variance is not tractable.

Theoretical comparison of the three bandwidth selectors does not seem possible. For this reason, in the next section we investigate via simulations the practical performance the three bandwidth rules proposed here when applied on $\hat{\lambda}_L(x)$.

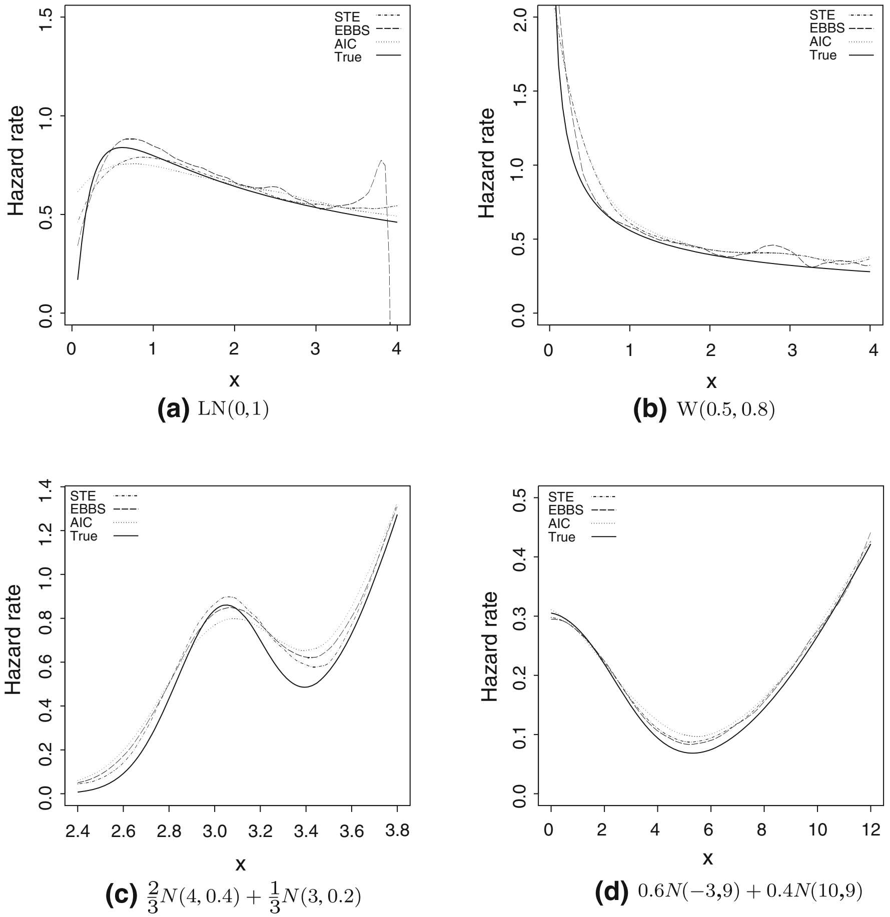
4 Numerical study and graphical illustration of the estimates

In this section we use distributional data to illustrate visually and through simulations the practical performance of $\hat{\lambda}_L(x)$ using the three bandwidth rules discussed in Sect. 3. As a vehicle we use four hazard rate functions which are frequently met in reliability studies, see for example Lai and Xie (2006) and Navarro and Hernandez (2004). These are the hazard rate function of the standard Lognormal distribution (LN(0, 1)), the Weibull distribution with shape parameter 0.5 and scale parameter 0.8 (W(0.5, 0.8)) and the truncated to $(0, +\infty)$ normal mixtures $\frac{2}{3}N(4, 0.4) + \frac{1}{3}N(3, 0.2)$ and $0.6N(-3, 9) + 0.4N(10, 9)$.

To implement random censoring we independently generate random censoring times from the uniform $U[0, k]$ distribution where k is selected so that the desired percentage of censoring is achieved on average across all iterations. The Epanechnikov kernel is used throughout in all implementations. Moreover, in all realizations of the estimate, the length of each partition is calculated by $\Delta = (X_{(N)} - X_{(1)})/n$, where $X_{(N)}$ and $X_{(1)}$ denote the last and first observation of the ordered sample respectively and n is the number of grid points at which we calculate $\hat{\lambda}_L(x)$. In all examples in this section $n = 80$. It has to be noted that in general selection of Δ may be viewed as selection of the number of bins of an histogram. Although not reported, all results in this section have been also calculated with $\Delta = 3.49\hat{\sigma} N^{-1/3}$, $\hat{\sigma}$ being the sample standard deviation. That is the normal reference rule, see for example Wand (1997). However, the simulation results and the associated inference drawn are very similar to the results and inference reported in Tables 1 and 2 later in this section and thus, we suggest using the first approach as it is simpler, faster and yields equivalent results.

In Fig. 1a–d we plot for each hazard function the average of 20 realizations of $\hat{\lambda}_L(x)$ using the three bandwidth rules, all calculated with sample size $N = 200$ and 10% censoring. Apart from the AIC case in the LN(0, 1) example of Fig. 1a, $\hat{\lambda}_L(x)$ demonstrates good behavior in the left boundary. The plots indicate that all bandwidth selectors produce through $\hat{\lambda}_L(x)$ acceptable estimates of the hazard rate function, with the EBBS method to have a slight advantage in the W(0.5, 0.8) and $0.6N(-3, 9) + 0.4N(10, 9)$ examples (Fig. 1b, d). Moreover it appears that the EBBS method is more efficient in capturing changes in the shape of the curvature as it can be seen in Fig. 1a–c. However, this comes at a cost of a higher bias near the tail of the region of estimation as seen in Fig. 1a and b.

Next we compare the approximate MISE's of estimator $\hat{\lambda}_L(x)$ when using the three bandwidth rules. We consider for this comparison four different sample sizes, $N = 100, 200, 400$ and 1,000 and four levels of censoring, i.e., censoring at 0, 10, 20 and 30% of the sample. For each distribution, sample size and level of censoring we compute the approximate MISE of each estimate as follows. The differences $(\hat{\lambda}_L(x_i) - \lambda(x_i))^2$, with $\lambda(x_i)$ being the true hazard rate at x_i, are calculated for all grid points x_i, $i = 1, \ldots, n$ and then integrated using Simpson's extended rule. This is done three times, one for each bandwidth selection method. The averaged integrated differences across 100 iterations are reported on the tables for each estimate. Note that in every iteration the sample in use is common for all bandwidth rules.



(a) LN(0,1)

(b) W(0.5, 0.8)

(c) $\frac{2}{3}N(4, 0.4) + \frac{1}{3}N(3, 0.2)$

(d) $0.6N(-3,9) + 0.4N(10,9)$

Fig. 1 Averages of 20 estimates of $\hat{\lambda}_L$ using the STE, EBBS and AIC bandwidth rules. Sample size is 200 and amount of censoring is 10%. The *solid black line* is the true hazard rate

The results are given in Tables 1 and 2 where the STE column corresponds to the plug-in method while the EBBS and AIC columns are rather self-explanatory. The overall conclusion is that the EBBS procedure seems to be doing slightly better than the plug-in and the AIC methods. However, this does not translate in uniform superiority across all sample sizes and levels of censoring. In Table 1, the W(0.5, 0.8) MISE figures indicate that in this case the plug-in and the EBBS method are rather equivalent, with EBBS to be doing marginally better, while the AIC method demonstrates inferior performance. On the contrary, in the LN(0, 1) simulation at 30% censoring and all sample sizes the AIC procedure is better than the other two methods. For the rest three levels of censoring in general all methods seem rather equivalent. For the $\frac{2}{3}N(4, 0.4) + \frac{1}{3}N(3, 0.2)$ illustration in Table 2, the figures suggest an advantage of EBBS over the other two methods in large samples with the plug-in method to be better in small samples. Exception is the 30% censoring case where the plugin method

Table 1 Approximate MISE's of $\hat{\lambda}_L(x)$ estimating the hazard rate from the normal mixture W(0.5, 0.8) and LN(0, 1) for various sample sizes (N) and amounts of censoring

Censoring	N	STE	EBBS	AIC	STE	EBBS	AIC
		Weibull(0.5, 0.8)			Lognormal(0, 1)		
0%	100	0.1733	0.1639	0.1925	0.3511	0.3509	0.3502
	200	0.0933	0.092	0.0996	0.2481	0.2487	0.2543
	400	0.0522	0.0511	0.0591	0.1956	0.1948	0.1968
	1,000	0.0311	0.0284	0.0301	0.1672	0.1666	0.1554
10%	100	0.1819	0.1824	0.2141	0.3951	0.3940	0.3922
	200	0.0995	0.1024	0.1107	0.2801	0.2793	0.2850
	400	0.057	0.0569	0.0657	0.2198	0.2187	0.2206
	1,000	0.0346	0.0316	0.0334	0.1891	0.1871	0.1742
20%	100	0.302	0.3051	0.3638	0.8705	1.0269	0.8306
	200	0.2157	0.2134	0.2545	0.6913	0.6923	0.6703
	400	0.193	0.1928	0.2214	0.5946	0.5943	0.5982
	1,000	0.1484	0.1473	0.1569	0.5171	0.5086	0.5106
30%	100	1.111	0.963	1.3417	0.5553	1.3172	0.4582
	200	0.9991	0.9471	1.1532	0.5001	0.9584	0.4982
	400	1.0345	1.0779	1.104	0.3761	0.5154	0.3207
	1,000	0.9243	0.9388	0.9519	0.1898	0.2114	0.1602

Table 2 Approximate MISE's of $\hat{\lambda}_L(x)$ estimating the hazard rate from $\frac{2}{3}N(4, 0.4) + \frac{1}{3}N(3, 0.2)$ and $0.6N(-3, 9) + 0.4N(10, 9)$ for various sample sizes (N) and amounts of censoring

Censoring	N	STE	EBBS	AIC	STE	EBBS	AIC
		$\frac{2}{3}N(4, 0.4) + \frac{1}{3}N(3, 0.2)$			$0.6N(-3, 9) + 0.4N(10, 9)$		
0%	100	0.0232	0.0278	0.0311	0.0212	0.0207	0.0156
	200	0.0187	0.0233	0.0274	0.0128	0.0192	0.0093
	400	0.0123	0.0121	0.0181	0.0084	0.0099	0.0071
	1,000	0.009	0.0992	0.0110	0.0057	0.00702	0.00535
10%	100	0.0493	0.0552	0.0537	0.0555	0.0673	0.0812
	200	0.0288	0.0296	0.0356	0.0529	0.057	0.0676
	400	0.0165	0.0149	0.0201	0.0352	0.0377	0.046
	1,000	0.0106	0.0092	0.015	0.0286	0.0302	0.0335
20%	100	0.0902	0.0981	0.1096	0.1395	0.1949	0.2082
	200	0.0674	0.0679	0.0847	0.1351	0.1511	0.1682
	400	0.0499	0.0485	0.0586	0.0979	0.1088	0.1199
	1,000	0.0366	0.0342	0.0403	0.083	0.0879	0.0953
30%	100	0.1883	0.1981	0.1884	0.824	0.78066	0.82403
	200	0.154	0.1542	0.1592	0.4162	0.4329	0.4987
	400	0.1101	0.1134	0.1468	0.3043	0.3198	0.3779
	1,000	0.0942	0.0946	0.0959	0.2818	0.2963	0.3272

is generally better in all sample. On the contrary, in the $0.6N(-3, 9) + 0.4N(10, 9)$ case, the plug-in method is doing generally better than the EBBS method even though marginally and not across all sample sizes and levels of censoring.

Finally, it has to be noted that the marginal superiority of the EBBS method comes with the cost of a much longer computation time especially compared to the plug-in rule. Thus, although in general it is expected that the EBBS will give more precise results, in practical work where time constrains are very frequent, the plug-in method may be more appropriate as the extra computational cost of EBBS or the AIC methods may not be worthy of the added precision in estimation.

5 Proofs

5.1 Proof of theorem 1

In matrix notation $\hat{\lambda}_L(x)$ can be written as

$$\hat{\lambda}_L(x) = e_1^T(X^TWX)^{-1}X^TWC$$

with $C = (c_1, \ldots, c_n)^T$, $e_1 = (1, 0)^T$ and

$$X = \begin{pmatrix} 1 & x_1 & -x \\ 1 & x_2 & -x \\ \cdots & & \\ 1 & x_n & -x \end{pmatrix}, \quad W = \text{diag}_{1\le i\le n} K\left(\frac{x_i - x}{h}\right).$$

Now, note that

$$\mathbb{E}C = (m(x_1), \ldots, m(x_n))^T \equiv (\lambda(x_1), \ldots, \lambda(x_n))^T = M. \tag{12}$$

It is straightforward to show that

$$e_1^T(X^TWX)^{-1} = \frac{1}{\det S}\left(S_{n,2}(x)\ -S_{n,1}(x)\right). \tag{13}$$

where $\det S = S_{n,1}^2(x) - S_{n,0}(x)S_{n,2}(x)$. Also

$$X^TWM = \begin{pmatrix} \sum_{i=1}^n K\left(\frac{x_i-x}{h}\right)\lambda(x_i) \\ \sum_{i=1}^n K\left(\frac{x_i-x}{h}\right)(x_i - x)\lambda(x_i) \end{pmatrix}. \tag{14}$$

Then, by (12)–(14) and simple algebra we get

$$\begin{aligned}\mathbb{E}\hat{\lambda}_L(x) &= e_1^T(X^TWX)^{-1}X^TWM \\ &= \frac{S_{n,1}(x)L_{n,1}(x) - S_{n,2}(x)L_{n,0}(x)}{S_{n,1}^2(x) - S_{n,2}(x)S_{n,0}(x)}.\end{aligned}$$

As for the variance we have

$$\mathbb{V}\text{ar}\left\{\hat{\lambda}_L(x)\right\} = \sigma^2(x)e_1^T(X^TWX)^{-1}(X^TW^2X)(X^TWX)^{-1}e_1.$$

Simple calculations yield

$$X^TW^2X = \begin{pmatrix} R_{n,0}(x) & R_{n,1}(x) \\ R_{n,1}(x) & R_{n,2}(x) \end{pmatrix} \equiv R$$

and thus, after a little algebra from the first to the second step below

$$\begin{aligned}\mathbb{V}\text{ar}\left\{\hat{\lambda}_L(x)\right\} &= \frac{\sigma^2(x)}{(\det S)^2}\left(S_{n,2}(x) \;\; -S_{n,1}(x)\right) R \begin{pmatrix} S_{n,2}(x) & -S_{n,1}(x) \\ -S_{n,1}(x) & S_{n,0}(x) \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \frac{\lambda(x)}{1-F(x)} \\ &\quad \times \frac{S_{n,2}^2(x)R_{n,0}(x) - 2S_{n,2}(x)S_{n,1}(x)R_{n,1}(x) + S_{n,1}^2(x)R_{n,2}(x)}{(S_{n,1}^2(x) - S_{n,2}(x)S_{n,0}(x))^2}.\end{aligned}$$

□

5.2 Proof of theorem 2

The proof is identical with the proof of theorem 1 in Bagkavos and Patil (2008) with the only difference that equations 15, 16 and 17 of lemma 4 there should be proved for the random right censorship model. However, this has already been established by Wang et al. (1998) in the proof of theorem A.1 there. Specifically, first note that in their notation the bin center of the I_j, $j = 1, 2, \ldots, n$ interval is defined as $t_j = \Delta(j-1)+\Delta/2$ so that $\Delta(j-1)$ denotes the start of the interval. Now, since $F < 1$ we have that $F^n(\Delta(j-1)) = o(n^{-1})$ as $n \to +\infty$. Using this in (A.13) and in the subsequent equations for the proof of $\mathbb{E}[\tilde{q}^2(t_j)]$ and $\mathbb{E}[\tilde{q}(t_i)\tilde{q}(t_j)]$, pages 155 and 156 respectively in the proof of theorem A.1 in Wang et al. (1998) we see that equations 15, 16 and 17 of lemma 4 in Bagkavos and Patil (2008) are readily verified for the case of random right censorship. □

5.3 Proof of theorem 3

The proof is based on the Hajek projection lemma (Hajek 1968, lemma 4.1) according to which, for a statistic W based on the i.i.d sample $Y_1, Y_2, \ldots, Y_N$ the standardized versions of W and its projection on the subspace of all such independent terms, say $\hat{W}$, have the same asymptotic distribution. Our proof is along the same lines with the corresponding proof of Tanner and Wong (1983) (Sect. 3 there) who first employed the method. For this reason, here we will only highlight the major differences. In our case the i.i.d sample consists of the pairs (X_i, δ_i), $i = 1, \ldots, N$ and $W = \hat{\lambda}_L(x)$.

First note that $\hat{\lambda}_L(x)$ can be written as

$$\hat{\lambda}_L(x) = \sum_{j=1}^{n} \sum_{i=1}^{N} K_N \left(\frac{x_j - x}{h} \right) \frac{1_{\{X_i \in I_j, \delta_i = 1\}}}{n_j}$$

with

$$K_N(u) = \frac{\left[S_{n,2}(x) - S_{n,1}(x) hu \right] K(u)}{\Delta \left\{ S_{n,2}(x) S_{n,0}(x) - S_{n,1}^2(x) \right\}} I_{[-p,\, +\infty)}(u).$$

We have that

$$W = \sum_{j=1}^{n} \sum_{i=1}^{N} K_N \left(\frac{x_j - x}{h} \right) \frac{1_{\{X_i \in I_j, \delta_i = 1\}}}{n_j} = \sum_{i=1}^{N} W_i$$

where

$$W_i = \sum_{j=1}^{n} K_N \left(\frac{x_j - x}{h} \right) \frac{1_{\{X_i \in I_j, \delta_i = 1\}}}{n_j}.$$

Let $Y_i = (X_i, \delta_i)$, $i = 1, \ldots, N$ which are i.i.d. The approximation of W is given by

$$\hat{W} = \sum_{i=1}^{N} \mathbb{E}(W | Y_i) - (N - 1) \mathbb{E} W.$$

In order to show that W and $\hat{W}$ have the same asymptotic distribution the following conditions which are implied by Hajek's lemma and are easy to verify must hold

$$\mathbb{E}(W - \hat{W})^2 = \mathbb{V}\text{ar}(W) - \mathbb{V}\text{ar}(\hat{W}), \quad \mathbb{E}\hat{W} = \mathbb{E}W. \tag{15}$$

We should also show that

$$\text{Var}(\hat{W}) / \text{Var}(W) \to 1 \text{ as } N \to +\infty. \tag{16}$$

From lemma 1 in Sect. 5.4 we have that

$$\hat{W} - \mathbb{E}\hat{W} = \sum_{i=1}^{N} \left\{ \mathbb{E}\left(W_i | Y_i \right) + (N - 1) \mathbb{E}\left(W_r | Y_i \right) - \mathbb{E}W \right\}. \tag{17}$$

Now from (23) in lemma 2 and lemma 3 in Sect. 5.4, (17) becomes

$$\begin{aligned}\hat{W} - \mathbb{E}\hat{W} &= \sum_{i=1}^{N} \left\{ \frac{m(Y_i)}{Nh} \sum_{j=1}^{n} K_N\left(\frac{x_j - x}{h}\right) \frac{1 - F^N(x_j)}{1 - F(x_j)} \right. \\ &\quad \left. + \int K_{(p)}\left(\frac{s-x}{h}\right) \lambda(s)\, ds - \mathbb{E}W \right\} + o(1) \\ &= \sum_{i=1}^{N} \left\{ \frac{m(Y_i)}{Nh} \sum_{j=1}^{n} K_N\left(\frac{x_j - x}{h}\right) \frac{1 - F^N(x_j)}{1 - F(x_j)} \right. \\ &\quad \left. - \int K_{(p)}\left(\frac{s-x}{h}\right) \frac{\lambda(s)}{1 - F(s)}\, ds \right\} + o(1) + O(N^{-1}).\end{aligned}$$

Therefore the variance of $\hat{W}$ becomes,

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\hat{W}) &= \mathbb{V}\mathrm{ar}\left\{ \sum_{i=1}^{N} \left\{ \frac{m(Y_i)}{Nh} \sum_{j=1}^{n} K_N\left(\frac{x_j - x}{h}\right) \frac{1 - F^N(x_j)}{1 - F(x_j)} \right.\right. \\ &\quad \left.\left. - \int K_{(p)}\left(\frac{s-x}{h}\right) \frac{\lambda(s)}{1 - F(s)}\, ds \right\} \right\} \\ &\simeq \mathbb{V}\mathrm{ar}\left\{ \sum_{i=1}^{N} \frac{m(Y_i)}{Nh} \sum_{j=1}^{n} K_N\left(\frac{x_j - x}{h}\right) \frac{1}{1 - F(x_j)} \right\}\end{aligned}$$

after using the fact that $F^N(s) = O(N^{-1})$. From Bagkavos and Patil (2008), p. 49 we have that

$$K_N(u) = h^{-1} K_{(p)}(u)(1 + o(1)). \tag{18}$$

Using lemma 2 in Bagkavos and Patil (2008) with $t_i = x_j$, $B = h^{-1}$ and

$$G(x) = K_{(p)}\left(\frac{x_j - x}{h}\right) \frac{1}{1 - F(x_j)}$$

together with (18) gives

$$\mathbb{V}\mathrm{ar}(\hat{W}) \simeq \frac{1}{Nh^2} \mathbb{V}\mathrm{ar}\left\{ m(Y_1) \int K_{(p)}\left(\frac{s-x}{h}\right) \frac{1}{1 - F(s)}\, ds \right\} + o(1).$$

Using again the properties for the variance and a Taylor expansion in the integrand it is easily established that

$$\mathbb{V}\mathrm{ar}(\hat{W}) \simeq \frac{1}{Nh} \frac{\lambda(x)}{1 - F(x)} \int K_{(p)}^2(u)\, du + O(N^{-1}).$$

It easily follows that $\mathrm{Var}(\hat{W}) \simeq \mathrm{Var}(W)$ and hence (16) is proved. By (15) and (16)

$$\mathbb{E}\left(\frac{\hat{W}-\mathbb{E}\hat{W}}{\sqrt{\mathrm{Var}(\hat{W})}}-\frac{W-\mathbb{E}W}{\sqrt{\mathrm{Var}(\hat{W})}}\right)^2=\frac{\mathbb{E}\left(\hat{W}-W\right)^2}{\mathrm{Var}(\hat{W})}=\frac{\mathrm{Var}(\hat{W})-\mathrm{Var}(W)}{\mathrm{Var}(\hat{W})}\to 0.$$

which proves that W and $\hat{W}$ have the same asymptotic distribution. Finally, asymptotic normality of the standardized $\hat{W}$ statistic follows easily by application of the Lyapunov theorem. □

5.4 Auxiliary lemmas

The following lemma appeared without proof first in Tanner and Wong (1983). However, we give the proof here as we believe that it is not immediately obvious.

Lemma 1

$$\hat{W}-\mathbb{E}\hat{W}=\sum_{i=1}^{N}\left\{\mathbb{E}\left(W_i|Y_i\right)+(N-1)\mathbb{E}\left(W_r|Y_i\right)-\mathbb{E}W\right\}. \tag{19}$$

Proof In order to prove (19) first write it as

$$\begin{aligned}\hat{W}-\mathbb{E}\hat{W}&=\sum_{i=1}^{N}\mathbb{E}\left(W_i|Y_i\right)+(N-1)\sum_{i=1}^{N}\mathbb{E}\left(W_r|Y_i\right)-\sum_{i=1}^{N}\mathbb{E}W\\&=\sum_{i=1}^{N}\mathbb{E}\left(W_i|Y_i\right)+(N-1)\sum_{i=1}^{N}\mathbb{E}\left(W_r|Y_i\right)-N\mathbb{E}W.\end{aligned} \tag{20}$$

The Hajek projection formula can be written as

$$\begin{aligned}\hat{W}-\mathbb{E}\hat{W}&=\sum_{i=1}^{N}\mathbb{E}\left(W|Y_i\right)-(N-1)\mathbb{E}\hat{W}-\mathbb{E}\hat{W}\\&=\sum_{i=1}^{N}\mathbb{E}\left(W|Y_i\right)-N\mathbb{E}W+\mathbb{E}W-\mathbb{E}\hat{W}\\&=\sum_{i=1}^{N}\mathbb{E}\left(W|Y_i\right)-N\mathbb{E}W\end{aligned} \tag{21}$$

where we used the fact that $\mathbb{E}W = \mathbb{E}\hat{W}$ in the second step above. So, from (20) and (21) the lemma will be proved if we show that

$$\sum_{i=1}^{N} \mathbb{E}\left(W_i|Y_i\right) + (N-1)\sum_{i=1}^{N} \mathbb{E}\left(W_r|Y_i\right) - N\mathbb{E}W = \sum_{i=1}^{N} \mathbb{E}\left(W|Y_i\right) - N\mathbb{E}W$$

which, since the term $N\mathbb{E}W$ is common on both sides of the equation, reduces to

$$\sum_{i=1}^{N} \left\{\mathbb{E}\left(W_i|Y_i\right) + (N-1)\mathbb{E}\left(W_r|Y_i\right)\right\} = \sum_{i=1}^{N} \mathbb{E}\left(W|Y_i\right). \tag{22}$$

Now,

$$\begin{aligned}\sum_{i=1}^{N} \mathbb{E}\left(W|Y_i\right) &= \mathbb{E}\left(W|Y_1\right) + \mathbb{E}\left(W|Y_2\right) + \cdots + \mathbb{E}\left(W|Y_N\right)\\ &= \mathbb{E}\left(\sum_{i=1}^{N} W_i|Y_1\right) + \mathbb{E}\left(\sum_{i=1}^{N} W_i|Y_2\right) + \cdots + \mathbb{E}\left(\sum_{i=1}^{N} W_i|Y_N\right).\end{aligned}$$

That is,

$$\begin{aligned}\sum_{i=1}^{N} \mathbb{E}\left(W|Y_i\right) &= \sum_{i=1}^{N} \mathbb{E}(W_i|Y_i) + \sum_{i=1}^{N}\sum_{i\neq r} \mathbb{E}(W_r|Y_i)\\ &= \sum_{i=1}^{N} \left\{\mathbb{E}(W_i|Y_i) + \sum_{i\neq r} \mathbb{E}(W_r|Y_i)\right\}\\ &= \sum_{i=1}^{N} \left\{\mathbb{E}(W_i|Y_i) + (N-1)\mathbb{E}(W_r|Y_i)\right\}\end{aligned}$$

with $i \neq r$. This proves (22) and thus the proof of the lemma is completed. □

Lemma 2

$$\mathbb{E}\left\{\frac{1_{\{X_i\in I_j,\delta_i=1\}}}{n_j}|Y_i\right\} = m(Y_i)\left\{\frac{1}{N}\frac{1-F^N(x_j)}{1-F(x_j)} - F^{N-1}(x_j)\right\}, \tag{23}$$

$$\mathbb{E}\left\{\frac{1_{\{X_r\in I_j,\delta_r=1\}}}{n_j}|Y_i\right\} = \frac{1}{N-1}\lambda(x_j) + \frac{1}{N-1}\frac{\lambda(x_j)}{1-F(x_j)} + O\left(N^{-1}\right). \tag{24}$$

Proof

$$\mathbb{E}\left\{\frac{1_{\{X_i\in I_j,\delta_i=1\}}}{n_j}|Y_i\right\}=\mathbb{E}\left\{1_{\{X_i\in I_j,\delta_i=1\}}|Y_i\right\}$$
$$\times\,\mathbb{E}\left\{\frac{1}{1_{\{X_i>x_j\}}+\sum_{k=1,k\neq i}^{N}1_{\{X_k>x_j\}}}|Y_i\right\}. \tag{25}$$

From Tanner and Wong (1983) we have that

$$\mathbb{E}\left\{1_{\{X_i\in I_j,\delta_i=1\}}|Y_i\right\}=f_T(Y_i)(1-H(Y_i))/f(Y_i)\equiv m(Y_i). \tag{26}$$

Also, given Y_i,

$$r\equiv\sum_{k=1,k\neq i}^{N}1_{\{X_k>x_j\}}\sim\text{Binomial}(N-1,1-F(x_j))$$

and thus

$$\mathbb{E}\left\{\frac{1}{1_{\{X_i>x_j\}}+\sum\limits_{\substack{k=1\\k\neq i}}^{N}1_{\{X_k>x_j\}}}|Y_i\right\}=\mathbb{E}\left\{\frac{1}{1+r}\right\}$$
$$=\sum_{r=1}^{N-1}\frac{1}{1+r}\binom{N-1}{r}(1-F(x_j))^rF^{N-r-1}(x_j)$$

with $r=1,\ldots,N-1$. Set $1+r=m$, that is $r=m-1$. Then $m=2,\ldots,N$ and

$$\mathbb{E}\left\{\frac{1}{1+r}\right\}=\sum_{m=2}^{N}\frac{1}{m}\binom{N-1}{m-1}(1-F(x_j))^{m-1}F^{N-m}(x_j)$$
$$=\frac{1}{N}\frac{1}{1-F(x_j)}\sum_{m=0}^{N}\binom{N}{m}(1-F(x_j))^mF^{N-m}(x_j)$$
$$-\frac{1}{N}\frac{1}{1-F(x_j)}\sum_{m=0}^{1}\binom{N}{m}(1-F(x_j))^mF^{N-m}(x_j)$$
$$=\frac{1}{N}\frac{1-F^N(x_j)}{1-F(x_j)}-F^{N-1}(x_j). \tag{27}$$

From (26) and (27), (25) becomes

$$\mathbb{E}\left\{\frac{1_{\{X_i\in I_j,\delta_i=1\}}}{n_j}|Y_i\right\} = m(x_j)\left\{\frac{1}{N}\frac{1-F^N(x_j)}{1-F(x_j)} - F^{N-1}(x_j)\right\}$$

and thus (23) is proved. Now,

$$\begin{aligned}\mathbb{E}\left\{\frac{1_{\{X_r\in I_j,\delta_r=1\}}}{n_j}|Y_i\right\} &= \mathbb{E}\left\{\mathbb{E}\left\{\frac{1_{\{X_r\in I_j,\delta_r=1\}}}{n_j}|Y_i\right\}|Y_r, i<r\right\}\\ &\quad+\mathbb{E}\left\{\mathbb{E}\left\{\frac{1_{\{X_r\in I_j,\delta_r=1\}}}{n_j}|Y_i\right\}|Y_r, i>r\right\}\\ &=\mathbb{E}\Bigg\{\mathbb{E}\left\{1_{\{X_r\in I_j,\delta_r=1\}}|Y_i\right\}\\ &\quad\times\mathbb{E}\left\{\frac{1}{1_{\{X_r>x_j\}}+\sum_{k=1,k\neq r}^{N}1_{\{X_k>x_j\}}}|Y_i\right\}\Bigg|Y_r, i<r\Bigg\}\\ &\quad+\mathbb{E}\Bigg\{\mathbb{E}\left\{1_{\{X_r\in I_j,\delta_r=1\}}|Y_i\right\}\\ &\quad\times\mathbb{E}\left\{\frac{1}{1_{\{X_r>x_j\}}+1_{\{X_i>x_j\}}+\sum_{k=1,k\neq i,r}^{N}1_{\{X_k>x_j\}}}|Y_i\right\}\Bigg|Y_r, i>r\Bigg\}.\end{aligned} \tag{28}$$

Using (26), (28) becomes

$$\mathbb{E}\left\{\frac{1_{\{X_r\in I_j,\delta_r=1\}}}{n_j}|Y_i\right\} = \mathbb{E}\left\{m(x_j)\mathbb{E}\left\{\frac{1}{1+l}\right\}\right\} + \mathbb{E}\left\{m(x_j)\mathbb{E}\left\{\frac{1}{2+l}\right\}\right\} \tag{29}$$

with $l \sim \text{Binomial}(N-2, 1-F(x_j))$. Thus,

$$\begin{aligned}\mathbb{E}\left\{\frac{1_{\{X_r\in I_j,\delta_r=1\}}}{n_j}|Y_i\right\} &= \int_{x_{j-1}}^{x_j} m(y)\mathbb{E}\left\{\frac{1}{1+l}\right\} f(y)\,dy\\ &\quad+\int_{x_{j-1}}^{x_j} m(y)\mathbb{E}\left\{\frac{1}{2+l}\right\} f(y)\,dy.\end{aligned} \tag{30}$$

Now,

$$\mathbb{E}\left\{\frac{1}{1+l}\right\} = \sum_{l=1}^{N-2}\frac{1}{1+l}\binom{N-2}{l}(1-F(x_j))^{l-2}F^{N-l-2}(x_j).$$

Let $1 + l = m$, so that $l = m - 1$. Then m ranges from 2 to $N - 1$ and calculations similar to (27) give

$$\mathbb{E}\left\{\frac{1}{1+l}\right\} = \sum_{m=2}^{N-1} \frac{1}{m}\binom{N-2}{m-1}(1-F(x_j))^{m-1}F^{N-m-1}(x_j)$$
$$= \frac{1}{N-1}\frac{1-F^{N-1}(x_j)}{1-F(x_j)} - F^{N-2}(x_j). \tag{31}$$

Also,

$$\mathbb{E}\left\{\frac{1}{2+l}\right\} = \sum_{l=1}^{N-2} \frac{1}{2+l}\binom{N-2}{l}(1-F(x_j))^{l-2}F^{N-l-2}(x_j).$$

Let $2 + l = m$, so that $l = m - 2$. Then m ranges from 3 to N and therefore

$$\mathbb{E}\left\{\frac{1}{2+l}\right\} = \sum_{m=3}^{N} \frac{1}{m}\binom{N-2}{m-2}(1-F(x_j))^{m-2}F^{N-m}(x_j)$$
$$\times (1-F(x_j))^m F^{N-m}(x_j)$$
$$= \frac{1}{N(N-1)}\frac{1}{(1-F(x_j))^2}\sum_{m=3}^{N}(m-1)\binom{N}{m}(1-F(x_j))^m F^{N-m}(x_j). \tag{32}$$

Note that

$$\sum_{m=3}^{N}\binom{N}{m}(1-F(x_j))^m F^{N-m}(x_j)$$
$$= \sum_{m=0}^{N}\binom{N}{m}(1-F(x_j))^m F^{N-m}(x_j) - \sum_{m=0}^{2}\binom{N}{m}(1-F(x_j))^m F^{N-m}(x_j)$$
$$= 1 - F^N(x_j) - N(1-F(x_j))F^{N-1}(x_j)$$
$$-\frac{N(N-1)}{2}(1-F(x_j))^2F^{N-2}(x_j). \tag{33}$$

Also

$$\sum_{m=3}^{N} m\binom{N}{m}(1-F(x_j))^m F^{N-m}(x_j)$$
$$= N\sum_{m=3}^{N}\binom{N-1}{m-1}(1-F(x_j))^m F^{N-m}(x_j)$$
$$= N - NF^{N-1}(x_j) - NF^{N-2}(x_j)(1-F(x_j)). \tag{34}$$

From (33) and (34), (32) becomes

$$\mathbb{E}\left\{\frac{1}{2+l}\right\} = \frac{1-F^{N-1}(x_j)-F^{N-2}(x_j)(1-F(x_j))}{(N-1)(1-F(x_j))^2}$$
$$-\frac{1-F^{N}(x_j)-N(1-F(x_j))F^{N-1}(x_j)-\frac{N(N-1)}{2}(1-F(x_j))^2F^{N-2}(x_j)}{N(N-1)(1-F(x_j))^2}$$
$$= \frac{1-F^{N-1}(x_j)-F^{N-2}(x_j)(1-F(x_j))}{(N-1)(1-F(x_j))^2} + O\left((N(N-1))^{-1}\right). \tag{35}$$

Finally, substitute (31) and (35) back to (30) to get

$$\mathbb{E}\left\{\frac{1_{\{X_r\in I_j,\delta_r=1\}}}{n_j}|Y_i\right\} = \left(\frac{1}{N-1}\frac{1-F^{N-1}(x_j)}{1-F(x_j)} - F^{N-2}(x_j)\right)\int_{x_{j-1}}^{x_j} m(y)f(y)\,dy$$
$$+\frac{1-F^{N-1}(x_j)-F^{N-2}(x_j)(1-F(x_j))}{(N-1)(1-F(x_j))^2}$$
$$\times\int_{x_{j-1}}^{x_j} m(y)f(y)\,dy + O\left((N(N-1))^{-1}\right). \tag{36}$$

From Wang et al. (1998), page 154 we have that

$$\int_{x_{j-1}}^{x_j} m(y)f(y)\,dy \simeq \lambda(x_j)(1-F(x_j))$$

and therefore (36) becomes

$$\mathbb{E}\left\{\frac{1_{\{X_r\in I_j,\delta_r=1\}}}{n_j}|Y_i\right\} = \frac{1}{N-1}\lambda(x_j) + \frac{1}{N-1}\frac{\lambda(x_j)}{1-F(x_j)} + O\left(N^{-1}\right)$$

which completes the proof of (24). □

Lemma 3

$$\mathbb{E}\{W_r|Y_i\} = \frac{1}{N-1}\int K_{(p)}\left(\frac{s-x}{h}\right)$$
$$\times\left\{\lambda(s)+\frac{\lambda(s)}{1-F(s)}\right\}ds\,dx(1+o(1)+o(N^{-1})).$$

Proof Since $F < 1$, asymptotically, as $N \to +\infty$, we have that $F^N(x_j) = o(N^{-1})$. Using (24) and (18) in the second step below

$$\begin{aligned}\mathbb{E}\{W_r|Y_i\} &= \mathbb{E}\left\{\sum_{j=1}^{n} K_N\left(\frac{x_j - x}{h}\right)\frac{1_{\{X_r \in I_j, \delta_r = 1\}}}{n_j}|Y_i\right\}\\ &= h^{-1}\sum_{j=1}^{n} K_{(p)}\left(\frac{x_j - x}{h}\right)\mathbb{E}\left\{\frac{1_{\{X_r \in I_j, \delta_r = 1\}}}{n_j}|Y_i\right\}(1 + o(1))\\ &= \frac{1}{N-1}\int K_{(p)}\left(\frac{s - x}{h}\right)\left\{\lambda(s) + \frac{\lambda(s)}{1 - F(s)}\right\} ds(1 + o(1) + o(N^{-1}))\end{aligned}$$

after applying lemma 2 in Bagkavos and Patil (2008) with $t_i = x_j$, $B = 1$ and

$$G(x) = K_{(p)}\left(\frac{x_j - x}{h}\right)\left(\lambda(x_j) + \frac{\lambda(x_j)}{1 - F(x_j)}\right).$$

□

Acknowledgments The author thanks two anonymous referees for their useful comments and suggestions that helped in improving this research.

References

Bagkavos, D., Patil, P. N. (2008). Local polynomial fitting in failure rate estimation. *IEEE Transactions on Reliability, 56*, 126–163.

Cheng, M. Y. (1997). Boundary aware estimators of integrated density derivative products. *Journal of the Royal Statistical Society, Series B, 59*, 191–203.

Cox, D. R., Oakes, D. (1984). *Analysis of survival data.* London: Chapman & Hall.

Fan, J., Gijbels, I. (1996). *Local polynomial modelling and its applications.* London: Chapman & Hall.

Fledelius, P., Guillen, M., Nielsen, J. P., Petersen, K. S. (2004). A comparative study of parametric and nonparametric estimators of old-age mortality in Sweden. *Journal of Actuarial Practice, 11*, 103–127.

Gefeller, O., Michels, M. (1992). A review on smoothing methods for the estimation of the hazard rate based on kernel functions. In Y. Dodge, J. Whittaker (Eds.), *Computational statistics* (pp. 459–464). Switzerland: Physica.

Hajek, J. (1968). Asymptotic normality of simple linear statistics under alternatives. *The Annals of Mathematical Statistics, 39*, 325–346.

Hall, P., Marron, S. (1987). Estimation of integrated squared density derivatives. *Statistics and Probability Letters, 6*, 109–115.

Hurvich, C., Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297–307.

Hurvich, C., Simonoff, J., Tsai, C. (1997). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B, 60*, 271–293.

Jones, M. C. (1993). Simple boundary correction for density estimation. *Statistics and Computing, 3*, 135–146.

Jones, M. C., Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistics and Probability Letters, 11*, 511–514.

Lai, C. D., Xie, M. (2006). *Stochastic ageing and dependence for reliability.* New York: Springer.

Linton, O., Nielsen, J. P. (1995). Kernel estimation in a nonparametric marker dependent hazard model. *Annals of Statitics, 23*, 1735–1748.

Mammen, E., Nielsen, J. P. (2007). A general approach to the predictability issue in survival analyses. *Biometrika, 94*, 873–892.

Müller, H.-G., Wang, J. L., Capra, W. B. (1997). From lifetables to hazard rates: The transformation approach. *Biometrika, 84*, 881–892.

Navarro, J., Hernandez, P. (2004). How to obtain bathtub-shaped failure rate models from normal mixtures. *Probability in the Engineering and Informational Sciences, 18*, 511–531.

Nielsen, J. P. (1998a). Marker dependent hazard estimation from local linear estimation. *Scandinavian Actuarial Journal, 2*, 113–124.

Nielsen, J. P. (1998b). Multiplicative bias correction in kernel hazard estimation. *Scandinavian Journal of Statistics, 25*, 541–553.

Nielsen, J. P. (2003). Variable bandwidth kernel hazard estimators. *Journal of Nonparametric Statistics, 15*, 355–376.

Nielsen, J. P., Tanggaard, C. (2001). Boundary and bias correction in kernel hazard estimation. *Scandinavian Journal of Statistics, 28*, 675–698.

Nielsen, J. P., Tanggaard, C., Jones, C. (2009). Local linear density estimation for filtered survival data. *Statistics, 42*, 167–186.

Ruppert, D. (1997). Empirical bias bandwidths for local polynomial regression and density estimation. *Journal of the American Statistical Association, 92*, 1049–1062.

Ruppert, D., Sheather, S. J., Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association, 90*, 1257–1270.

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

Turlach, B. A., Wand, M. P. (1996). Fast computation of auxiliary quantities in local polynomial regression. *Journal of Computational and Graphical Statistics, 5*, 337–350.

Wand, M. P. (1997). Data-based choice of histogram bin width. *American Statistician, 51*, 59–64.

Wand, M., Jones, M. C. (1995). *Kernel smoothing*. London: Chapman & Hall

Wang, J. L., Müller, H.-G., Capra, W. B. (1998). Analysis of oldest-old mortality: Lifetables revisited. *Annals of Statitics, 28*, 126–163.