Smoothing Hazard Rates

Contribution to the Encyclopedia of Biostatistics

Jane-Ling Wang
Department of Statistics, University of California, Davis, CA 95616, U.S.A.
Email: wang@wald.ucdavis.edu

Version: June 22, 2003

Abstract. The nonparametric approach to estimate hazard rates for lifetime data is flexible, model-free and data-driven. No shape assumption is imposed other than that the hazard function is a smooth function. Such an approach typically involves smoothing of an initial hazard estimate, with arbitrary choice of smoother. We describe methods for grouped lifetime data observed at certain time intervals and for continuously observed lifetime data. There are some intrinsic differences between the smoothing approaches for these two types of data. More specifically, smoothing an initial hazard estimate based on the life table is adopted for grouped lifetime data; while for continuous data, smoothing is employed to increments of the Nelson-Aalan cumulative hazard estimate aiming at the derivative of the cumulative hazard function. A few nonparametric hazard regression methods are also discussed.

Keywords: Life table; Nelson-Aalen estimator; Nonparametric smoothing methods; Bandwidth choices; Boundary effects; Hazard regression.

1 Introduction

In the analysis of lifetime data or time-to-event data, a primary interest is to assess the risk of an individual at certain times (or ages) (*see* **Survival Analysis**, **Overview**). Let T denote a lifetime variable with distribution function $F(t) = \Pr(T \leq t)$ and probability density function $f(t) = dF(t)/dt$. The risk of an individual at age t can be measured by the so called "**hazard rate**" or "hazard function", which is defined as:

$$\lambda(t) = f(t)/[1 - F(t)], \text{ for } F(t) < 1. \tag{1}$$

That is, $\lambda(t)dt$ represents the instantaneous chance that an individual will die in the interval $(t, t + dt)$ given that this individual is alive at age t. The hazard rate provides the trajectory of risk and is widely used also in other fields. Engineers refer to it as "failure rate function" and demographers refer to it as "force of mortality function". The term "lifetime" simply denotes the time until the occurrence of an event of interest.

While parametric models provide convenient ways to analyze lifetime data, the necessary model assumptions, when violated, can lead to erroneous analyses and thus need to be checked carefully (*see* **Parametric Models in Survival Analysis**). We give a brief survey on hazard rate estimation in this article. No shape restriction on the hazard rate is assumed except for smoothness. Such a model-free approach is data driven and can be used for parametric model checking. The nonparametric approach of hazard rate estimation typically involves the smoothing of an initial hazard estimate. The brief survey of various smoothing hazard rate estimators provided here covers grouped lifetime data on the one hand and continuously observed lifetime data on the other.

For grouped data, the observations occur in the form of scatter-plots (t_i, q_i), where q_i is an initial hazard estimate at the midpoint t_i of the ith time interval. Smoothing for such data corresponds to a scatter-plot smoothing or **nonparametric regression** step. As for continuously observed data, hazard

rate estimation resembles **density estimation** (smoothing the increments of a cumulative function estimate). Almost any density estimation method can be adapted for hazard rate smoothing. The simplest such method is the kernel method which should however be employed with care in the boundary region. More details are given later in Section 5.

2 Smoothing Hazard Rates for Grouped Data: Nonparametric Graduation of Lifetables

The earliest nonparametric hazard rate estimate was the **life table** estimate based on grouped lifetimes (*see* **Grouped Survival Times**), which has been known for centuries. Assume for simplicity that lifetimes are grouped into intervals of unit length with midpoints $t_1, ..., t_p$. Let n_i denote the number of individuals alive (or at risk) at the beginning of interval i, and d_i denote the number of observed deaths during this interval. An *ad hoc* estimate of the hazard rate for the ith interval is the so called death rate, $q_i = d_i/n_i$ (for intervals of length Δ the death rate is replaced by $d_i/(\Delta n_i)$). A plot of the raw death rates at various times t_i typically yields a curve that is ragged, indicating high variability; see Figure 1 for an example concerning the death rates of 1,000 female Mediterranean fruit flies. Dead flies were counted daily, and q_i is the death rate at day i.

Since the actual hazard rate λ is typically assumed to be a smooth function, smoothing the death rates provides an aesthetically improved estimate (see Fig. 1 for two versions of smoothed death rates). A smoothing procedure, when applied properly, also improves the statistical performance of the resulting hazard rate estimator.

For example, the smoothed death rates typically have a faster convergence rate than the unsmoothed death rates. The smoothing of death rates was pioneered by actuaries who referred to these smoothing methods as "linear graduation" or "nonparametric graduation", in contrast to "analytic gradu-

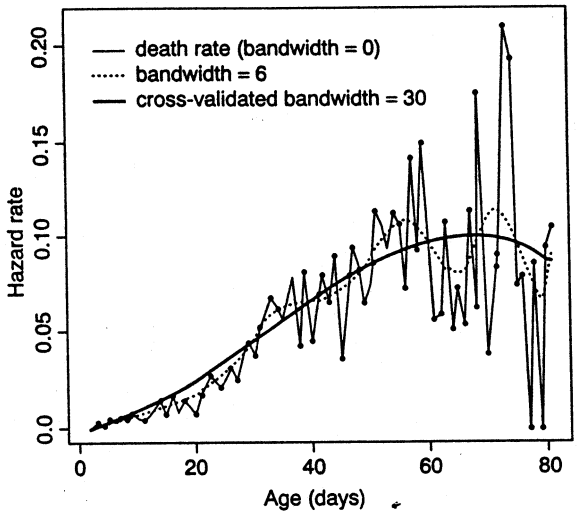

Figure 1: Three hazard rate estimates for the survival of 1,000 female Mediterranean fruit flies. (a) death rates (thin line) (b) smoothed hazard rate with fixed bandwidth $b = 6$ (solid line) (c) smoothed hazard rate with least squares cross-validated bandwidth choice $b = 30$ (bold line).

ation" based on parametric models (*see* **Actuarial Methods**). The term "linear" refers to the fact that these nonparametric graduation methods yield hazard estimates of the form

$$\hat{\lambda}(t) = \sum_{i=1}^{p} c_i(t) q_i \quad , \text{where } \sum_{i=1}^{p} c_i(t) = 1, \text{ at each time } t. \tag{2}$$

That is, the resulting hazard estimate at age t is a weighted average of the death rates with weights $c_i(t)$ specified by the method of graduation and adjusted locally at each age t.

The graduation (or smoothing) process typically reduces the variance of the resulting hazard estimates at the expense of introducing biases. The graduated or smoothed hazard estimate converges to the true hazard rate at a slower rate than the $\sqrt{n}$ rate which holds for a parametric (or analytic) graduated hazard estimate.

Moving averages, local weighted **least squares** methods and the so-called Whittaker-Henderson estimates have been the earliest proposals among a variety of different possible graduation methods, and are commonly adopted by actuaries (see Borgan [6] and Hoem [33]). Any nonparametric regression

method can be used to graduate **life tables** in order to obtain a smooth hazard rate estimate. One just applies the chosen smoother, which could be a spline or kernel method, to the scatter plot $\{(t_i, q_i),\, i = 1, ..., p\}$. The Whittaker-Henderson estimate resembles a spline estimate. The kernel method for graduation (see Copas and Haberman [8], Bloomfield and Haberman [4]) is conceptually simple but needs to be applied with caution in the boundary region of the data, owing to its large bias there.

For the graduation of grouped data, we recommend the local polynomial method which is also called the locally weighted least squares method. This graduation method has been credited to the famous mathematician J.P. Gram, perhaps best known for his contributions to **Gram-Schmidt** orthogonalization. See Hoem [32] and Seal [52] for historical reviews. Specifically, in his doctoral dissertation, Gram [21] suggested a weighted least squares method to fit a smooth curve locally by polynomials. The explicit form of Gram's estimate using a local linear fit is given in equation (3) below.

The local polynomial method is well suited for graduating initial hazard estimates based on life tables. As a least squares based procedure, it is simple to interpret, and automatically includes boundary corrections. For the kernel method, boundary corrections require the implementation of special boundary kernels. Both kernel and local polynomial methods are theoretically more tractable than the spline method, especially for lifetime data which are often incomplete. Some asymptotic results for the local polynomial estimator are reviewed in the next section.

We note that the death rate q_i can be replaced by any initial estimate of the hazard rate. For example, the central death rate, $q_{c_i} = 2d_i/(n_i + n_{i+1})$, is a good alternative. If death rates are used in (2), it is recommended (see (11) of next section and [61]) to include a transformation of the smoothed death rates $\hat{\lambda}(t)$, and to use $-\log(1 - \hat{\lambda}(t))$ as the final hazard estimate. This transformation reduces the bias resulting from grouping the data. This bias can be substantial at extreme ages (i.e., for large t) and may result in

inconsistent estimates of the hazard rate. If the central death rates are used in (2), another transformation (*see* (13) of next section and [43]) of the smoothed central death rates is recommended instead.

As for the choice of the smoother in (2), it is a judgement call, and typically the choice of an adequate smoothing parameter is more important. The sampling or asymptotic properties of the resulting hazard rate estimator are much more complicated than in the standard regression setting, as the q_i or other initial hazard estimates are not independent of each other. The incompleteness of lifetime data further complicates theoretical analysis. Therefore, much is yet to be explored in hazard rate estimation based on smoothing life tables.

For an overview and details of the kernel smoothing method, see Wand and Jones [60]; for the spline method, Greene and Silverman [26]; and for the local polynomial method, Fan and Gijbels [14].

3 More on Local Polynomial Hazard Smoothing for Grouped Data

In addition to the grouping, we shall assume that the lifetimes $T_1, T_2, \cdots, T_n$, based on a cohort of n individuals, are subject to random censoring by $C_1, C_2, \cdots, C_n$. Let $I_1, I_2, ..., I_p$ denote a partition of p ordered intervals over a time interval of length L. For the jth individual, the value of $\delta_j = 1_{\{X_j=T_j\}}$ is known but not the actual value of $X_j = \min(T_j, C_j)$. It is only known that $X_j \in I_i$ for some i. Observed are (d_i, n_i), where $d_i = \sum_{j=1}^{n} 1_{\{X_j \in I_i, \delta_j=1\}}$ is the number of observed deaths in the interval I_i, and $n_i = \sum_{j=1}^{n} 1_{\{X_j \in I_k, \text{ for some } k \geq i\}}$ is the number of individuals at risk at the beginning of the interval I_i.

For simplicity of presentation we shall assume that the intervals I_i are of equal length Δ and that the first interval starts at zero. The non-equal length case can be handled similarly as in nonparametric regression with

non-equidistant design points and will not be discussed here. The grouped data can thus be summarized in lifetable form which consists of data pairs $(t_i, q_i),\ i = 1, ..., p$. Here, $t_i = \Delta(i - \frac{1}{2})$ is the midpoint of the ith interval I_i and $q_i = \tilde{q}(t_i) = d_i/(\Delta n_i)$ is the death rate (out of those alive) for interval I_i. A closer look at $\hat{q}$ reveals that it is an empirical estimate of the population death rate defined by

$$q(t) = \Delta^{-1} \Pr(T \in (t - \frac{\Delta}{2}, t + \frac{\Delta}{2}) | T > t - \frac{\Delta}{2}),$$

and one expects $q(t)$ to be close to the true hazard function $\lambda(t)$, provided that Δ is small.

The local polynomial smoother due to Gram [21,22], is based on smoothing the lifetable data $\{(t_i, q_i),\ i = 1, ..., p\}$ by locally fitting a polynomial of fixed degree r. Thus, given a bandwidth or window of size $b = b_n$, for estimation at age t, a polynomial $g(x - t)$ of degree r is fitted to all lifetable data points (t_i, q_i) for which $|t - t_i| \leq b$. The coefficients of the polynomial $g(\cdot)$ are obtained via the weighted least squares criterion and the value of the fitted polynomial at t (i.e., the intercept) is the hazard estimate. A common choice is to fit local linear polynomials (i.e., $r = 1$).

For $r = 1$, this estimate, denoted by $\hat{q}(t)$, is equal to the minimizer for a_0 of

$$\sum_{i=1}^{p} w_i K((t - t_i)/b) \{q_i - [a_0 + a_1(t_i - t)]\}^2 . \tag{3}$$

Here w_i are case weights, typically chosen as $w_i = n_i$, and K is a nonnegative kernel function satisfying

$$V = \int K^2(x)dx < \infty. \tag{4}$$

We recommend to use either the Epanechnikov kernel

$$\begin{array}{l} K(x) = .75(1 - x^2), \quad -1 \leq x \leq 1, \\ \text{or the Gaussian kernel} K(x) = (2\pi)^{-1/2} e^{-x^2/2}. \end{array} \tag{5}$$

The bandwidths should satisfy

$$b_n \to 0 \text{ and } nb_n \to \infty; \tag{6}$$

The weighted least squares method is used for two reasons. First, in the spirit of smoothing methods, it gives remote observations less influence in a way that can be controlled by choice of bandwidth and kernel in (3). Second, it allows to address the high degree of heteroscedasticity (*see* **Scedasticity**) of the lifetable estimate q_i, through the choice of the case weights w_i in (3). Bias and variance expressions are derived in Wang et al. [61] and summarized below.

First we define a constant that appears in the leading bias term:

$$B = \frac{1}{2}\int x^2 K(x)dx \tag{7}$$

Under the kernel and bandwidth conditions (4) and (6), and if in addition

$$\Delta \to 0, \text{and } \Delta \log n/b \to 0 \text{ as } n \to \infty, \tag{8}$$

we have for t with $F(t) < 1$ and $G(t) < 1$, and B and V as in (4),(7),

$$bias(\hat{q}(t)) = -\frac{\Delta}{2}\lambda^2(t) + \frac{\Delta^2}{24}[\lambda^{(2)}(t) + 4\lambda^3(t)] + b^2\lambda^{(2)}(t)B + o(b^2) + o(\Delta^2) \tag{9}$$

$$var(\hat{q}(t)) = \frac{1}{nb}\left\{\frac{\lambda(t)}{[1-F(t)]\,[1-G(t)]}V + o(1)\right\}. \tag{10}$$

Bias Reduction Transformation

Note that the leading term of the variance in (10) is the same as for the kernel estimate for continuously observed data in (21). The terms in (9) involving b correspond to the bias due to smoothing and are also the same as for continuously observed data with $k = 2$ in (20). The terms

involving Δ in (9) correspond to an additional bias due to the grouping of the data. This additional bias can be improved by the transformation $\phi(x) = -\log(1-\Delta x)/\Delta$, which is motivated by the relation

$$\Delta q(t) = 1 - \frac{1 - F(t+\frac{\Delta}{2})}{1 - F(t-\frac{\Delta}{2})} = 1 - \exp\left[-\int_{t-\frac{\Delta}{2}}^{t+\frac{\Delta}{2}} \lambda(x)dx\right] \approx 1 - e^{-\Delta\lambda(t)}.$$

Thus, we propose the transformed estimate

$$\phi(\hat{q}(t)) = -\log(1 - \Delta\hat{q}(t))/\Delta, \tag{11}$$

which has the same variance expression (10) as $\hat{q}$ has, but a bias of smaller order:

$$bias(\phi(\hat{q}(t))) = \frac{\Delta^2}{24}\lambda^{(2)}(t) + b^2\lambda^{(2)}(t)B + o(b^2) + o(\Delta^2). \tag{12}$$

Comparing (9) and (12), we see that $\hat{q}(t)$ has an additional bias, $-\frac{\Delta}{2}\lambda^2(t) + \frac{\Delta^2}{6}\lambda^3(t)$, as compared to $\phi(\hat{q}(t))$. In addition to this bias reduction there are other advantages in using $\phi(\hat{q}(t))$ rather than $\hat{q}(t)$, especially when hazards at extreme ages are of primary interest (*see* Wang et al. [61] for details). If the central death rate, q_{c_i}, is used in (3) instead of the death rate, q_i, a different transformation is proposed in Müller et al. [43], given by:

$$\psi(\hat{q}_c(t)) = \frac{1}{\Delta}\log\frac{2 + \Delta\hat{q}_c(t)}{2 - \Delta\hat{q}_c(t)} \tag{13}$$

We close this section by pointing out that the rate of convergence of $\hat{q}(t)$, $\phi(\hat{q}(t))$, $\hat{q}_c(t)$ or $\psi(\hat{q}_c(t))$, and the choice of the bandwidth b can be derived analogous to that of the kernel estimate $\hat{\lambda}$ in Section 5, with Δ playing a role in the asymptotic bias term. The program to compute $\hat{q}(t)$ in (3) or $\hat{q}_c(t)$ and their corresponding transformed estimates, $\phi(\hat{q}(t))$ in (11) or $\psi(\hat{q}_c(t))$ is very simple, and so is the computation of the cross-validated bandwidths as employed in [43] and [61].

The hazard rate estimate, based on the least squares cross-validated bandwidth, calculated from the lifetimes for 1,000 female Mediterranean fruit flies

is plotted in Figure 1. The lifetimes are grouped into days. Here the cross-validated bandwidth is fairly large ($b = 30$), owing to the large variation of the death rates after day 60. The hazard plot was truncated at day 81 when there were only 10 flies left.

4 Smoothing Hazard Rates for Continuously Observed Data

The grouped data situation discussed in the previous section is common for demographic data that were observed at fixed time points or grouped for convenience. The estimation of hazard rates for continuously observed data is conceptually close to **density estimation**. To see this, consider, instead of (1), the hazard rate function as the derivative of the cumulative hazard function $\Lambda(t) = \int_0^t \lambda(x)dx$. A hazard rate estimate can thus be obtained, analogous to a density estimate, by smoothing the increments of an estimate of $\Lambda(t)$.

Watson and Leadbetter [63,64] were the first to propose and study such a smoothed hazard estimator using the empirical cumulative hazard estimate $\Lambda_n(t)$ based on an independent and identically distributed (i.i.d.) sample of lifetimes (that is, the $\Lambda_n(t)$ in (15) with all $\delta_{[j]} = 1$). They propose the following convolution type hazard estimator.

$$\hat{\lambda}_n(t) = \int W_n(t-x)d\Lambda_n(t), \tag{14}$$

where W_n is a sequence of smooth functions approaching the Dirac delta-function for large n. This delta-sequence method is quite general and covers several types of smoothing methods, including the kernel method (with $W_n(x) = b_n^{-1}K(x/b_n)$). Another type of hazard estimator proposed in Watson and Leadbetter [64] is of a ratio type,

$$\tilde{\lambda}_n(t) = \hat{f}_n(t)/[1 - \hat{F}_n(t)], \tag{15}$$

where $\hat{f}_n$ can be any density estimate of the lifetime density f and $\hat{F}_n$ is an empirical estimate of the lifetime distribution function F. Both types of hazard estimators have the same asymptotic variance but different asymptotic biases (Rice and Rosenblatt [50]). The convolution type estimator $\hat{\lambda}_n$ has prevailed owing to its theoretical tractability (exact **mean square errors** available) and aesthetic superiority over the ratio type estimator $\tilde{\lambda}_n$.

A complete **random sample** of lifetimes as assumed above is often unavailable. In reality, lifetime data are often incomplete owing to **staggered entry**, loss to follow-up, or early termination of a study. For simplicity of presentation we focus on the random **censoring** case for the rest of the entry. Basic references for hazard estimation for other incomplete data such as left truncated and right censored data can be found in Uzunogullari and Wang [59] and Gu [27]. The related problem of estimating transition intensities for a two-state Markov Process was explored in Keiding and Andersen [35].

Under the random censorship model, the actual lifetime T_i of an individual may be censored by another random variable C_i. One observes instead (X_i, δ_i), where $X_i = \min(T_i, C_i)$, the minimum of the lifetime and censoring time of the ith individual, and $\delta_i = 1_{\{X_i = T_i\}}$, which is one if the actual lifetime is observed and zero otherwise. We shall assume that the censoring times $C_1, C_2, ..., C_n$ have a common distribution function G and that they are independent of the lifetimes $T_1, ..., T_n$. Let $(X_{(i)}, \delta_{[i]})$, $i = 1, 2, ..., n$, be the ordered sample with respect to $X_i{}'s$ (that is, $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$, and $\delta_{[i]}$ is the corresponding censoring indicator of $X_{(i)}$).

Hazard estimators in this situation are ordinarily obtained by smoothing the increments of the **Nelson-Aalen estimator** $\Lambda_n(\cdot)$ for the cumulative hazard function $\Lambda(t)$. Let $N_n(t) = \sum_{i=1}^{n} 1_{\{X_i \leq t, \delta_i = 1\}}$, and $Y_n(t) = \sum_{i=1}^{n} 1_{\{X_i \geq t\}}$. The Nelson-Aalen estimator $\Lambda_n(\cdot)$, which is instrumental in survival analysis for censored data, is defined as

$$\Lambda_n(t) \quad = \int_0^t \frac{1_{\{Y_n(s)>0\}}}{Y_n(s)} dN_n(s) = \sum_{i=1}^{n} \frac{\delta_{[i]} 1_{\{X_{(i)} \leq t\}}}{n-i+1} \tag{16}$$

if there are no tied observations. Properties of the random step function $\Lambda_n(t)$ have been studied extensively, see for example, Andersen et al. [1], Section IV.1 for details.

Kernel Estimators

Substituting the Λ_n in (16) into (14) and choosing $W_n(x) = b^{-1}\,K((t-x)/b)$, for a particular choice of kernel K and bandwidth $b = b_n$, we arrive at the kernel hazard estimator:

$$\begin{aligned}\hat{\lambda}(t) &= \int \tfrac{1}{b} K(\tfrac{t-x}{b}) \Lambda_n(x), \\ &= \sum_{i=1}^{n} \tfrac{1}{b} K(\tfrac{t-X_{(i)}}{b}) \tfrac{\delta_{[i]}}{n-i+1},\end{aligned} \tag{17}$$

if there are no tied observations.
Asymptotic properties on consistency are typically obtained under the following assumptions: (i) the true hazard rate is k-times differentiable for a $k \geq 0$; (ii) the bandwidths satisfy (6); and (iii) the kernel is of order k, defined as:

$$\begin{aligned}&\int K(x)dx = 1, \qquad \int K^2(x)dx < \infty, \quad \int x^j K(x)dx = 0 \text{ for } 1 < j < k, \\ &\int x^k K(x)dx \qquad \text{is finite but nonzero.}\end{aligned} \tag{18}$$

The choice of the bandwidth is of crucial importance and regulates the trade off between the bias and variance of the estimator in (17). A small bandwidth yields a less smooth curve, with smaller bias but larger variance, as compared to a larger bandwidth (see (20) and (21)). Bandwidth choice is particularly crucial for hazard estimation near the right boundary of the data as the variance increases to infinity there. More discussions on bandwidth choice is provided in the next section.

As for the choice of the kernel, smoothness of the kernel determines the smoothness of the corresponding kernel estimate, and the order of the kernel determines the order of the bias (see (20)) and thus the rate of convergence.

Often, nonnegative kernels are used in practice, and the Epanechnikov kernel in (5) has certain optimality properties (see Müller [40].

The kernel hazard estimate is the simplest and thus widely adopted smooth hazard estimator. It has been studied extensively in the literature, for example, by Ramlau-Hansen [48,49], Yandell [66], Tanner and Wong [58], Burke and Horvath [7], Diehl and Stute [11] and Müller and Wang [41].

Spline Estimators

Another commonly adopted smoothing method is the spline method. There are several types of spline methods. The most widely investigated spline method for hazard smoothing is the penalized likelihood approach. Let $\eta(t) = \log \lambda(t)$ be the log hazard function. The log likelihood function for censored data is:

$$\ell(\eta) = \sum_{i=1}^{n} \left\{ \delta_i \eta(X_i) - \int_0^{X_i} e^{\eta} \right\},$$

which is unbounded if no shape restriction on η is imposed. A penalty $J(\eta)$, measuring the roughness of η, is therefore incorporated and the penalized likelihood estimate $\hat{\eta}$ of η is the maximizer of the penalized log likelihood

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i \eta(X_i) - \int_0^{X_i} e^{\eta} \right\} - \frac{\alpha}{2} J(\eta), \tag{19}$$

among all η in a Hilbert space. Here α is a smoothing parameter. Smaller α yields a better fit but a more variable (rough) curve. A typical choice of $J(\eta)$ is $\int [\eta^{(2)}(x)]^2 dx$, which leads to a cubic spline with knots at all $X's$. More specifically, $\hat{\eta}$ is two-times continuously differentiable and is a piecewise cubic polynomial between any two consecutive $X's$. The smoothing parameter α plays a similar role as the bandwidth b in a kernel estimate. **Cross-validation** is a common way to determine the value of α. See O'Sullivan [44, 45] for computational details and Gu [27] for asymptotic results.

In (19), the roughness of $\log \lambda(t)$ is penalized so as to avoid nonnegative constraints on the hazard function. Other forms of penalty functions were proposed in Anderson and Senthilselvan [2], Senthilselvan [46], and Antoniadis and Grégoire [3]. The penalty function J determines the kind of spline resulted from (19). For example, the penalty $J(\eta) = \int [\lambda'(X)]^2 dx$ is employed in Anderson and Sethilselvan [2], and the resulting hazard estimate is a piecewise quadratic spline. Note that this hazard estimate may yield negative values under heavy censoring.

The above spline estimates have knots at each of the observed X values and are called smoothing splines in the literature (see Green and Silverman [20, Chapter 2]. Another type of spline method is regression splines or B-Splines which adopt a fixed number of knots and basis functions. See Rosenberg [51] and Kooperberg et al. [36] for details and ways to select the number and location of knots. A hazard function estimate with flexible tails, called HEFT, is proposed in [36] by estimating the log-hazard function using cubic splines.

Other Hazard Rate Estimators

The ratio type hazard estimator in (14), also due to Watson-Leadbetter, has been extended to censored data as well and was studied by Blum and Susarla [5], Földes, Rejtö and Winter [16] and Lo, Mack and Wang [38].

Hjort [31] advocated the use of semiparametric approaches to estimate hazard rates. The approach is to start with a possibly crude parametric estimate and to improve it via some nonparametric procedures. The motivation is to reduce the bias of a parametric estimate via nonparametric correction locally, and yet to arrive at an estimate that is less variable than a fully nonparametric one.

For reviews of earlier results on hazard rate estimation see Padgett [46] and Gefeller and Michels [18], and Singpurwalla and Wong [55] for uncensored data.

5 More on Kernel Hazard Estimators for Continuously Observed Data

The rate of convergence of the kernel hazard estimate (17) depends on the order of the kernel, the bandwidth and the differentiability of the hazard function. Typically, the order k of the kernel is chosen to be an even number with $k = 2$ being the standard choice. The resulting bias and variance are respectively:

$$bias(\hat{\lambda}(t)) = b^k[\lambda^{(k)}(t)B_k + o(1)], \tag{20}$$

$$var(\hat{\lambda}(t)) = \frac{1}{nb}\left\{\frac{\lambda(t)}{[1-F(t)][1-G(t)]}V + o(1)\right\}, \tag{21}$$

where $B_k = (-1)^k/k! \int x^k K(x)dx$ and V is as in (4).

The influence of the bandwidth b and the trade off between the bias and variance is seen from (20) and (21). The optimal rate for the mean squared error (MSE) of $\hat{\lambda}(t)$ is attained when the $(bias)^2$ and variance are of the same order. This results in an optimal MSE rate of convergence of $n^{2k/(2k+1)}$, which is $n^{4/5}$ for the standard choice of $k = 2$. This rate is slower than the usual parametric rate of n regardless of the order of k. For the asymptotic distribution, we further assume that $d = \lim_{n\to\infty} nb^{2k+1}$ exists for some $0 \le d < \infty$. Then

$$(nb)^{1/2}(\hat{\lambda}(t) - \lambda(t)) \xrightarrow{\mathcal{D}} N\left(d^{1/2}\lambda^{(k)}(x)B_k, \frac{\lambda(t)}{[1-F(t)][1-G(t)]}V\right). \tag{22}$$

Extensions to the estimation of derivatives of hazard functions have been considered as well (Müller and Wang [41]). These essentially involve a change in the kernel. Derivatives are of interest to detect rapid changes in hazard rates or for data based bandwidth choices, as the optimal bandwidths in (23) or (24) depend on the derivatives of the hazard rates. Again, the order k of the kernel affects the convergence rate and also asymptotic constants.

Bandwidth Choice

The bandwidth for a kernel hazard estimate can be fixed at all points (global bandwidth b) or can vary for different points (local bandwidth $b(t)$). Usually a global bandwidth is employed for a smooth density or regression estimate owing to its simplicity. However, for the hazard estimation situation discussed here there are compelling reasons to adopt local rather than global bandwidth choices. According to (21) the variance of the kernel estimate $\hat{\lambda}(t)$ eplodes to infinity as t approaches the right boundary of the data. Thus the variance tends to dominate the bias in the right tail and this needs to be compensated for by a larger bandwidth.

The optimal local bandwidth of $\hat{\lambda}(t)$ which minimizes the leading term of $MSE(\hat{\lambda}(t))$ is:

$$b^*(t) = n^{-1/(2k+1)} \left\{ \frac{1}{2k} \frac{\lambda(t)}{[1-F(t)][1-G(t)]} \frac{V}{[\lambda^{(k)}(t) B_k]^2} \right\}^{1/(2k+1)} \tag{23}$$

To find the optimal global bandwidth, we have to restrict the range of t to a compact interval $[0, \tau]$ with $F(\tau) < 1$ and $G(\tau) < 1$. The global optimal bandwidth which minimizes the leading term of $MISE(\hat{\lambda}) = E \int_0^\tau \left[\hat{\lambda}(x) - \lambda(x)\right]^2 dx$ is

$$b_{opt} = n^{-1/(2k+1)} \left\{ \frac{1}{2k} \int_0^\tau \frac{\lambda(x)}{[1-F(x)][1-G(x)]} dx \frac{V}{B_k^2 \int_0^\tau \left[\lambda^{(k)}(y)\right]^2 dy} \right\}^{1/(2k+1)}. \tag{24}$$

Note that both the local and global optimal bandwidths in (23) and (24) involve unknown quantities. In practice one has to find alternatives. There is an extensive literature on bandwidth selection and "cross-validation" and "plug-in" techniques are popular. See Patil [47] and Müller and Wang [41][42] for details. A bootstrap method to select the global bandwidth has been advocated in González-Manteiga, Cao and Marron [20] as an alternative. In addition to the local bandwidth choice in (23), which adopts different bandwidths at different time point t, choosing bandwidths as the distance of t to its $k-th$ nearest-neighbor among the remaining uncensored observations is a convenient way to adapt to the data by allowing for varying degrees of

smoothing; see Tanner [57s], Tanner and Wong [58] and Gefeller and Dette [17] for detailed descriptions. Other data-adaptive local or global bandwidth choices for hazard estimates can be derived analogously to the density estimation case as discussed in Silverman [54, Section 3.4] and Wand and Jones [60, Chapter 3].

Boundary Effects

We close this section with a cautionary remark that the kernel smoothing method needs to be employed very carefully near the boundary as there is a bias problem in such regions, usually referred to in the literature as boundary effects. Boundary effects may be attributed to the fact that the support of the kernel exceeds the available range of data and are not unique to hazard estimates.

An unmodified kernel estimate is unreliable in the boundary region, which is the region within one bandwidth of the largest or smallest observations. To remedy the boundary effects, different kernels, referred to as "boundary kernels" can be used within the boundary region. As a consequence, varying kernels are employed at each location t and the bandwidths are affected accordingly. The resulting kernel estimate with varying kernels and varying local bandwidths takes the form

$$\hat{\lambda}(t) = \int \frac{1}{b(t)} K_t \left(\frac{t-x}{b(t)} \right) d\Lambda_n(x), \tag{25}$$

where both the bandwidth $b = b(t)$ as well as the kernel $K = K_t$ depend on the point t. Details for the choices of the kernel K_t and bandwidths $b(t)$ can be found in Müller and Wang [42].

Simulation Comparison of Hazard estimators and Software

A very informative and extensive simulation study was carried out in Hess [30] to compare the aforementioned kernel-based hazard estimators with various local and global bandwidth choices and boundary corrections, the kernel-based hazard estimators in Gefeller and Dette [17] with varying bandwidth

methods based on kth nearest-neighbor, and the spline-based estimators in Kooperberg et al. [36]. The results indicated advantages of using **HADES**, the aforementioned local optimal bandwidth choice and boundary correction in Mueller and Wang [42]. There is significant improvement (over 50% on the average) in mean square error over the global bandwidth choice if a local optimal bandwidth is employed. Boundary corrections will lend additional efficiency. The locally optimal bandwidth estimators in [42] with only left boundary correction also outperformed two publicly available procedures, the spline estimator in [36] and the nearest-neighbor estimator in [17]. The latter is based on the procedures in Tanner[57] and Tanner and Wong [58].

A library of Fortran and S-Plus programs for the HADES estimator in [42] and for the nearest-neighbor estimator in [17] is available under a package called "muhaz" at the website of the authors of [30] :http://odin.mdacc.tmc.edu/anonftp/ To get the S-code follow the link: ftp://odin.mdacc.tmc.edu/pub/S/muhaz.tar.gz The corresponding R program for **muhaz** is also publicly available at: cran.r-project.org/doc/packages/muhaz

The S-plus code of the spline estimator in Kooperberg et al.[36] called, **HEFT** is publicly available from the StatLib software library.

6 Hazard Regression

Estimating a Baseline Hazard Function

So far we discussed hazard smoothing for a homogeneous population. Often the risk of an individual varies according to the values of some **covariates**. Thus the hazard function of an individual with covariate $Z \in \Re^d$ is $\lambda(t, Z)$ and regression techniques are required. A semi-parametric approach with a regression parameter β and a nonparametric baseline hazard function $\lambda_0(t)$ is often adopted. Examples include **Cox's proportional hazards regression model** where $\lambda(t, Z) = \lambda_0(t) \exp(\beta^T Z)$, and the **accelerated failure-time model** where $\lambda(t, Z) = \lambda_0(\exp(\beta^T Z)t) \cdot \exp(\beta^T Z)$.

A smooth estimate of the baseline hazard is preferable and often necessary to obtain consistent estimates of $\lambda(t, Z)$. Anderson and Senthilselvan [2] applied the penalized maximum likelihood approach, and Gray [23] and Wells [65] applied the kernel method to estimate the baseline hazard function in Cox's proportional hazard model. Andersen et. al [1, Section VII.2.5] give several examples of estimated baseline hazard functions.

The Cox proportional model has been extended in Dabrowska [10] to allow covariate dependent baseline hazard function. The model is: $\lambda(t, Z) = \lambda_0(t, X_t)exp[\beta^T Z_t]$, where X_t and Z_t are predictable covariate processes or covariate vectors. Another type of extension is to employ, as in Wang [62], an unknown link function in the proportional model, where $\lambda(t, Z) = \lambda_0(t)g(\beta^T Z)$ with g completely unknown and estimated via local **partial likelihood** method. Etezadi-Amoli and Ciampi [13] also investigated another extension of Cox's proportional hazards and accelerated failure time models of the form: $\lambda(t, Z) = \lambda_0(g_1(\alpha^T Z)t)g_2(\beta^T Z)\cdot$, where $\lambda_0(t)$ denotes the baseline hazard function which is estimated by the regression spline method.

Generalized Additive Proportional Hazards Model

Another type of **proportional hazards** model allows an arbitrary covariate effect of the form:

$$\lambda(t, Z) = \lambda_0(t) \exp[g(Z)], \tag{26}$$

where g is an unspecified smooth function of Z. LeBlanc and Crowley [37] use the CART (Classification and Regression Trees) algorithm to estimate the relative risk g (*see* **Tree-structured statistical Methods**), Gentleman and Crowley [19] and Fan, Gijbels and King [15] use local full or partial likelihood methods to estimate g. Although this is the most general proportional hazards model, it is difficult to estimate $g(Z)$ when the covariate Z is of high dimension, say $d \geq 3$. An extremely large sample size would be needed. This is called the "curse of dimensionality". Dimension reduction models

and methods are thus called for. Among these, the additive regression model is a promising alternative to (26).

Under the additional assumption that g is additive in (26), i.e. $g(z) = \sum_{i=1}^{d} g_i(z_i)$, Hastie and Tibshirani [29] and O'Sullivan [44, 45] use smoothing splines to estimate g (*see* **Generalized Additive Model**). Sleeper and Harrington [56] use B-splines, and Gray [24] uses penalized splines with fixed knots to estimate g and incorporate time-varying coefficients. Apart from the minor differences in the various spline methods, all the aforementioned methods adopt the partial likelihood approach with a penalty for each g_i to be estimated.

Let (X_i, Z_i, δ_i), $i = 1, ..., n$ denote the observed data and $Y_1 < ... < Y_k$ denote the k distinct failure times with d_i failures at time Y_i. The penalized log partial likelihood with smoothing parameters $\alpha_1, ..., \alpha_d$ is:

$$\ell(g_1, ..., g_d) = \sum_{i=1}^{k} \delta_i \left\{ \sum_{j \in D_i} g(Z_j) - d_i \log \left(\sum_{j \in R_i} e^{g(Z_j)} \right) \right\} - \frac{1}{2} \sum_{i=1}^{d} \alpha_i \int \left[g_i^{(2)}(t) \right]^2 dt,$$

where D_i is the set of indices of the failures at observed failure time X_i, and R_i is the set of indices of individuals at risk at time X_i. Minimizing $\ell(g_1, ..., g_d)$ then yields the smoothing spline estimates $(\hat{g}_1, ..., \hat{g}_d)$. Calculations of the estimates can be very time-consuming. See Hastie and Tibshirani [28, Section 8.3] for computational issues.

Nonparametric Hazard Regression

A completely nonparametric approach to estimate $\lambda(t, Z)$ is desirable sometimes. Kooperberg, Stone and Truong [36] used **loglinear regression** splines and their tensor products to estimate $\log \lambda(t, Z)$. Gu [27] considered the penalized likelihood approach. Doss and Li [12] used linear polynomials in Z to fit $\lambda(t, Z)$ locally in a neighborhood of Z. Martingale convergence theory for **counting processes** was used to derive the weak convergence of their hazard estimate.

For continuously observed lifetimes, one can obtain a hazard regression estimate for $\lambda(t, Z)$ by smoothing the increments of any cumulative hazard estimate $\Lambda(t, Z)$. Such a cumulative hazard estimate can be found in Dabrowska [9] and is further studied by McKeague and Utikal [39]. Again, any of the smoothing methods discussed so far can be extended to a nonparametric hazard regression estimate.

Note that by grouping the data along the time axis and the covariate axis, one can also apply any nonparametric regression smoother to grouped data. Gray [24] illustrates this grouping method through a local linear polynomial smoother and kernel regression.

Lexis Diagram

An interesting application of nonparametric hazard regression is the **Lexis diagram** in which individual life-lines are represented as line segments between (time at birth, 0) and (time, age) of death. Here time at birth can be used in a broad sense, i.e., as the onset time of a disease. If mortality of individuals varies according to time of birth, a covariate Z based on an individual's calendar time of birth can be incorporated to model individual risks at age t represented by $\lambda(t, Z)$. Keiding [34] suggests to use bivariate versions of nonparametric smoothing methods, as discussed above, to estimate $\lambda(t, Z)$, provided that the influence of Z on the hazard function is continuous in Z.

References

[**1**] Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

[**2**] Anderson, J. and Senthilselvan, A. (1980). Smooth estimates for the hazard function. *J. R. Statist. Soc. B* **42**, 322-327.

[**3**] Antoniadis, A. and Grégoire, G. (1990). Penalized likelihood estimation for rates with censored survival data. *Scand. J. Statist.* **17**, 43-63.

[**4**] Bloomfield, D. and Haberman, S. (1987). Graduation: Some experiments with kernel methods. *J. Inst. Actuaries* **114**, 339-369.

[**5**] Blum, J.R. and Susarla, V., (1980). Maximal deviation theory of density and failure rate function estimates based on censored data. In: Krishnaiah, P.R. (ed.) *Multivariate Analysis* vol. V, pp. 213-222. New York: North Holland.

[**6**] Borgan, Ø. (1979). On the theory of moving average graduation. *Scand. Actuarial J.* **1979**, 83-105.

[**7**] Burke, M.D. and Horváth, L. (1984). Density and failure rate estimation in a competing risks model. *Sankhyā Ser. A* **46**, 135-154.

[**8**] Copas, J. and Haberman, S. (1983). Nonparametric graduation using kernel methods. *J. Inst. Actuaries* **110**, 135-156.

[**9**] Dabrowska, D.M. (1987). Non-parametric regression with censored survival time data. *Scand. J. Statist.* **14**, 181-197.

[**10**] Dabrowska, D. (1997). Smoothed Cox regression. *Ann. Statist.* **25**, 1510-1540.

[**11**] Diehl, S. and Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. *J. Mult. Analy.* **25**, 299-310.

[**12**] Doss, H. and Li, G. (1995). An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.* **23**, 787-823.

[**13**] Etezadi-Amoli, J. and Ciampi, A. (1987). Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function. *Biometrics* **43**, 181-192.

[**14**] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* London: Chapman and Hall.

[**15**] Fan, J, Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *Ann. Statist.* **25**, 1661-2690.

[**16**] Földes, A., Rejtö, L. and Winter, B.B. (1981). Strong consistency properties of nonparametric estimators for randomly censored data. II: Estimation of density and failure rate. *Period. Math. Hung.* **12**, 15-29.

[**17**] Gefeller, O. and Dette, H.(1992). Nearest neighbor kernel estimation of the hazard function from censored data. *J. Statistical and Computational Simulations*, **43**, 93-101.

[**18**] Gefeller, O. and Michels, P. (1992). A review on smoothing methods for the estimation of the hazard rate based on kernel functions, in Dodge, Y. and Whittaker, J. (eds). *Computational Statistics*, Physica-Verlaz, Switzerland, 459-464.

[**19**] Gentleman, R. and Crowley, J. (1991). Local full likelihood estimation for the proportional hazard model. *Biometrics*, **47**,1283-1296.

[**20**] González-Manteiga, W., Cao R. and Marron, J.S. (1996). Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *J. Am. Statist. Assoc.* **91**, 1130-1140.

[**21**] Gram, J.P. (1879). *Om Rækkeudviklinger, bestemte ved Hjælp af de mindste Kvadraters Methode.* Copenhagen: A.F. H ϕst & Sϕn.

[**22**] Gram, J.P. (1883). Ueber Entwickelung reeller Functionen in Reihen mittelst der Methode der Kleinsten Quadrate. *J. Math.* **94**, 41-73.

[**23**] Gray, R. (1990). Some diagnostic methods for Cox regression models through hazard smoothing. *Biometrics* **46**, 93-102.

[**24**] Gray, R. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Am. Statist. Assoc.* **87**, 942-951.

[**25**] Gray, R. (1996). Hazard regression using ordinary nonparametric regression smoothers. *J. Comput. and Graphical Statist.* **5**, 190-207.

[**26**] Green, P.J. and Silverman, B.W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman and Hall.

[**27**] Gu, C. (1996). Penalized likelihood hazard estimation: A general procedure. *Statistica Sinica* **6**, 861-876.

[**28**] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* London: Chapman and Hall.

[**29**] Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **46**, 1005-1016.

[**30**] Hess, K. R., Serachitopol, D.M. and Brown, B.W. (1999). Hazard function estimators: a simulation study. *Statist. in Medicine.* **18**, 3075-3088.

[**31**] Hjort, N. (1991). Semiparametric estimation of parametric hazard rates. J.P. Klein and P.K. Goel, eds. *Survival Analysis: State of the Art*, 211-236 Kluwer: Dordrecht.

[**32**] Hoem, J. (1983). The reticent trio: Some little-known discoveries in life insurance mathematics by L.H.F. Oppermann, T.N. Thiele, and J.P. Gram. *International Statist. Rev.* **51**, 213-221.

[**33**] Hoem, J. (1984). A contribution to the statistical theory of linear graduation. *Insurance: Math. and Econ.* **3**, 1-17.

[**34**] Keiding, N. (1990). Statistical inference in the Lexis diagram. *Phil Trans. Roy. Soc. London A* **332**, 487-509.

[**35**] Keiding, N. and Andersen, P.K. (1989). Nonparametric estimation of transition intensities and transition probabilities: a case study of a two-state Markov process. *Appl. Statist.* **38**, 319-329.

[**36**] Kooperberg, C., Stone, C.J. and Truong, Y.K. (1995). Hazard regression. *J. Am. Statist. Assoc.* **90**, 78-94.

[**37**] LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics* **48**, 411-425.

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

[**39**] McKeague, I. W. and Utikal, K.J. (1990). Inference for a nonlinear counting process regression model. *Ann. Statist.* **18**, 1172-1187.

[**40**] Müller, H.G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Springer: New York.

[**41**] Müller, H.G. and Wang, J.L. (1990). Locally adaptive hazard smoothing. *Probab. Th. Rel. Fields* **85**, 523-538.

[**42**] Müller, H.G. and Wang, J.L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* **50**, 61-76.

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

[**44**] O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Statist. Comput.* **9**, 531-542.

[**45**] O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput.* **9**, 363-379.

[**46**] Padgett, W.J. (1988). Nonparametric estimation of density and hazard rate functions when samples are censored. P.R. Krishnaiah and C.R. Rao, eds., *Handbook of Statistics,* vol. **7**, pp. 313-331. Elsevier Science Publishers B.V.

[**47**] Patil, P.N. (1993). Bandwidth choice for nonparametric hazard rate estimation. *J. Statist. Plann. Inf.* **35**, 15-30.

[**48**] Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11** , 453-466.

[**49**] Ramlau-Hansen, H. (1983). The choice of a kernel function in the graduation of counting process intensities. *Scand. Actuar. J.*, 165-182.

[**50**] Rice, J. and Rosenblatt, M. (1976). Estimation of the log survivor function and hazard function. *Sankhyā Ser. A* **38**, 60-78.

[**51**] Rosenberg, P.S. (1995). Hazard function estimation using B-splines. *Biometrics* **51**, 874-887.

[**52**] Seal, H.L. (1981). Graduation by piecewise cubic polynomials: a historical review. Blätter, *Deutsche Gesellshaft für Versicherungsmathematik* **15**, 89-114.

[**53**] Senthilselvan, A. (1987). Penalized likelihood estimation of hazard and intensity functions. *J. R. Statist. Soc. B* **49**, 170-174.

[**54**] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall: London.

[**55**] Singpurwalla, N.D. and Wong, M.-Y. (1983). Estimation of the failure rate – a survey of nonparametric methods. Part I: Non-Bayesian methods. *Commun. Statist.-Theor. Meth.***12**, 559-588.

[**56**] Sleeper, L.A. and Harrington, D.P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *J. Am. Statist. Assoc.* **85**, 941-949.

[**57**] Tanner, M. A. (1983). A note on the variable kernel estimator of the hazard function from randomly censored data. *Ann. Statist.* **11**, 994-998.

[**58**] Tanner, M. A. and Wong, W.H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *Ann. Statist.***11**, 989-993.

[**59**] Uzunogullari, Ü. and Wang, J.-L. (1992). A comparison of hazard rate estimators for left truncated and right censored data. *Biometrika***79**, 297-310.

[**60**] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall: London.

[**61**] Wang, J.L., Müller, H.G. and Capra, W.B. (1998). Analysis of oldest-old mortality: Lifetables revisited. *Annals of Statistics***26**, 126-163.

[**62**] Wang, W. (2001). Proportional hazard regression model with unknown link function and applications to longitudinal time-to-event data. Ph.D. Thesis, University of California, Davis.

[**63**] Watson, G.S. and Leadbetter, M.R. (1964). Hazard analysis. I. *Biometrika* **51**, 175-184.

[**64**] Watson, G.S. and Leadbetter, M.R. (1964). Hazard analysis. II. *Sankhyā Ser. A* **26**, 101-116.

[**65**] Wells, M.T. (1994). Nonparametric kernel estimation in counting processes with explanatory variables. *Biometrika* **81**, 759-801.

[**66**] Yandell, B.S. (1983). Nonparametric inference for rates with censored survival data. *Ann. Statist.* **11**, 1119-1135.