Dynamic likelihood hazard rate estimation

Nils Lid Hjort

University of Oslo and University of Oxford

ABSTRACT. The best known methods for estimating hazard rate functions in survival analysis models are either purely parametric or purely nonparametric. The parametric ones are sometimes too biased while the nonparametric ones are sometimes too variable. In the present paper a certain semiparametric approach to hazard rate estimation, proposed in Hjort (1991), is developed further, aiming to combine parametric and nonparametric features. It uses a dynamic local likelihood approach to fit the locally most suitable member in a given parametric class of hazard rates, and amounts to a version of nonparametric parameter smoothing within the parametric class. Thus the parametric hazard rate estimate at time s inserts a parameter estimate that also depends on s. We study bias and variance properties of the resulting estimator and methods for choosing the local smoothing parameter. It is shown that dynamic likelihood estimation often leads to better performance than the purely nonparametric methods, while also having capacity for not losing much to the parametric methods in cases where the model being smoothed is adequate.

KEY WORDS: *dynamic likelihood, hazard rate, kernel smoothing, local goodness of fit, local modelling, semiparametric estimation*

1. Introduction and summary. This paper concerns a class of semiparametric type methods of estimating hazard rate functions in models for life history data. The best known methods for estimating such hazard rates are those that are either purely parametric or purely nonparametric. The parametric methods are usually biased since parametric models are usually imperfect, and the nonparametric methods often have high estimation variance. There should accordingly be room for methods that somehow lie between the parametric and the nonparametric ones. One might hope that such methods are better than the nonparametric ones if the true hazard is in the vicinity of the parametric model, while not being much worse than the parametric ones if the parametric model is true.

Although results can be obtained in a more general framework of counting process models we shall mainly be content to illustrate and investigate ideas for the 'random censorship' model, which is the simplest and perhaps most important special case of such models for censored life-time data. It postulates that life-times $X_1^0,\ldots,X_n^0$ from a population are i.i.d. with density $f(.)$, cumulative distribution $F(.)$, and hazard rate function $\alpha(.)$ given by $\alpha(s)=f(s)/F[s,\infty)$; $\alpha(s)\,\mathrm{d}s$ is the probability of failing in $[s,s+\mathrm{d}s)$ given that an individual is still at risk at time s. The life-time X_i^0 may not be directly observed, however, because of a possibly interfering censoring variable C_i; only $X_i=\min(X_i^0,C_i)$ and the indicator variable $\delta_i=I\{X_i^0\le C_i\}$ are observed. For simplicity and concreteness we stipulate that the C_i's are independent of the life-times and i.i.d. according to a distribution with cumulative function G. In particular the n pairs (X_i,δ_i) are i.i.d. Finally we shall assume that data are obtained on a finite time horizon basis, say on $[0,T]$ for a known and finite T. This is convenient for some of the martingale convergence theory and is not a practical limitation.

The parametric approach is to postulate that $\alpha(s)=\alpha(s,\theta)$ for a suitable family, indexed by some one- or multi-dimensional θ. Typical examples include the exponential, the Weibull, the simple frailty model with $\alpha(s)=\theta_1/(1+\theta_2 s)$, the piecewise constant hazard rate model, the Gompertz–Makeham distribution, the gamma, and the log-normal. Properties of the maximum likelihood method for estimating θ with censored data have been studied by Borgan (1984) and others under the condition that the model is correct, i.e. that there really is some θ_0 with $\alpha(s)=\alpha(s,\theta_0)$ on

$[0,T]$. In practice the model is never perfect, however, and it is useful to study estimation methods outside model conditions, where the best parameter is to be thought of as being 'least false' or 'most suitable', as opposed to 'true'. The large-sample behaviour of several estimation methods in this wider setting has been explored in Hjort (1992). Some results about this are reviewed in Section 2 and are used in later sections.

In Section 3 a dynamic likelihood approach to parametric hazard rate estimation is presented. It takes as its basis any given parametric hazard function and consists of inserting a local parameter estimate $\widehat{\theta}(s)$ in $\alpha(s,\theta)$ at time s, producing

$$\widehat{\alpha}(s) = \alpha(s, \widehat{\theta}(s)),$$

where the parameter estimate is obtained using only information on those individuals that have survived up to $s - \frac{1}{2}h$ and what happens to them on $[s - \frac{1}{2}h, s + \frac{1}{2}h]$. This amounts to a kind of nonparametric parameter smoothing within a given parametric class. A more general estimator involving smoothing with a kernel function is also discussed. Bias and variance properties are studied in Section 3 for one-dimensional and in Section 4 for multi-dimensional families. It turns out that

$$\mathrm{E}\widehat{\alpha}(s) \doteq \alpha(s) + \tfrac{1}{2}\beta_K h^2 b(s) \quad \text{and} \quad \mathrm{Var}\,\widehat{\alpha}(s) \doteq \frac{\gamma_K}{nh}\frac{\alpha(s)}{y(s)},$$

where β_K and γ_K are characteristics of the kernel function used and $y(s)$ is the limiting proportion of individuals still at risk at time s. The $b(s)$ is a certain bias factor, the size of which depends on both $\alpha''(s)$ and characteristics of the underlying parametric model used. These results match closely those of the most usual nonparametric method, that of smoothing the empirical cumulative hazard function, for which

$$\mathrm{E}\widetilde{\alpha}(s) \doteq \alpha(s) + \tfrac{1}{2}\beta_K h^2 \alpha''(s) \quad \text{and} \quad \mathrm{Var}\,\widetilde{\alpha}(s) \doteq \frac{\gamma_K}{nh}\frac{\alpha(s)}{y(s)}.$$

In Section 5 situations are characterised where the new method performs better than the traditional nonparametric method. Methods for choosing the local smoothing parameter h are discussed in Section 6, including the arduous one that for each s expands the $s \pm \frac{1}{2}h$ interval until a goodness of fit criterion rejects the model. Overall it transpires that a suitable dynamic likelihood estimator often can perform better than the purely nonparametric ones, while at the same time not losing much to parametric ones when the true hazard is close to the parametric hazard. Finally some supplementing results and remarks are offered in Section 7.

This paper expands in several ways on the basic results that were already presented in Hjort (1991). That paper also proposed two further semiparametric estimation schemes, one using orthogonal expansions to correct on an initial parametric guess, and one Bayesian procedure that employs a nonparametric prior around a given parametric hazard model.

2. Purely nonparametric and purely parametric estimation. This section introduces some basic notation and reviews properties of the Nelson–Aalen estimator for the cumulative hazard function in the nonparametric case and of the maximum likelihood and maximum weighted likelihood estimators in the parametric case. These will be used in later sections. Since our ambition is to go beyond ordinary parametric methods the behaviour of these must be considered also outside the conditions of the postulated parametric model.

2.1. NONPARAMETRIC ESTIMATION. Let $N(t) = \sum_{i=1}^n I\{X_i \le t, \delta_i = 1\}$ be the counting process and $Y(s) = \sum_{i=1}^n I\{X_i \ge s\}$ the at risk process, and form from these the Nelson–Aalen estimator

$$\widehat{A}(t)=\int_0^t \frac{\mathrm{d}N(s)}{Y(s)}=\sum_{i=1}^n \frac{\delta_i}{Y(X_i)} I\{X_i\le t\} \tag{2.1}$$

for the cumulative hazard rate $A(t)=\int_0^t \alpha(s)\,\mathrm{d}s$. Its properties are best explained using the martingale $B(t)=N(t)-\int_0^t Y(s)\alpha(s)\,\mathrm{d}s$. Let $y(s)$ be the limit in probability of $\widehat{y}(s)=Y(s)/n$, i.e. the limiting proportion of individuals under risk at time s, and equal to $F[s,\infty)G[s,\infty)$ under present circumstances, where $G(.)$ is the censoring distribution. A basic large-sample property of B is that $B(t)/\sqrt{n}$ goes to a Gaußian martingale $V(t)$ with independent increments and noise level $\mathrm{Var}\,\mathrm{d}V(s)=y(s)\alpha(s)\,\mathrm{d}s$, and, more generally, that $\int_0^t H_n(s)\,\mathrm{d}B(s)/\sqrt{n}$ tends to $\int_0^t h(s)\,\mathrm{d}V(s)$ in distribution, in cases where $H_n(.)$ is previsible (its value at s is known at $s-$) and converges to the deterministic $h(.)$. It follows from these facts that

$$\begin{aligned}\sqrt{n}\{\mathrm{d}\widehat{A}(s)-\mathrm{d}A(s)\}&=\frac{I\{Y(s)\ge 1\}}{\widehat{y}(s)}\frac{\mathrm{d}B(s)}{\sqrt{n}}-I\{Y(s)=0\}\,\mathrm{d}A(s)\\&\doteq_d \widehat{y}(s)^{-1}\mathrm{d}B(s)/\sqrt{n}\to_d y(s)^{-1}\,\mathrm{d}V(s)\end{aligned} \tag{2.2}$$

in the large-sample limit. In particular $\mathrm{d}\widehat{A}(s)$ is very nearly unbiased for $\mathrm{d}A(s)$ and $\sqrt{n}\{\widehat{A}(t)-A(t)\}$ tends to the Gaußian martingale $\int_0^t y(s)^{-1}\,\mathrm{d}V(s)$ with variance $\int_0^t y(s)^{-1}\alpha(s)\,\mathrm{d}s$. See for example the recent book Andersen, Borgan, Gill & Keiding (1993, Chapter II) for more details. The usual nonparametric way of estimating the hazard rate itself is to smooth the Nelson–Aalen and take the derivative, see (5.1).

2.2. Maximum likelihood estimation. A parametric model is of the form $\alpha(t)=\alpha(t,\theta)$, where $\theta=(\theta_1,\ldots,\theta_p)'$ is some p-dimensional parameter. The log-likelihood for the observed data can be written $L_n(\theta)=\int_0^T\{\log\alpha(t,\theta)\,\mathrm{d}N(t)-Y(t)\alpha(t,\theta)\,\mathrm{d}t\}$, see for example Andersen et al. (1993, Chapter VI). This defines the maximum likelihood estimator $\widehat{\theta}$.

To explain the large-sample behaviour of this estimator, let $U_n(\theta)=n^{-1}\int_0^T \psi(t,\theta)\{\mathrm{d}N(t)-Y(t)\alpha(t,\theta)\}\,\mathrm{d}t$ be the p-vector of first partial derivatives of $n^{-1}L_n(\theta)$, where we write $\psi(t,\theta)=\frac{\partial}{\partial\theta}\log\alpha(t,\theta)$. Under natural regularity conditions $U_n(\theta)$ tends in probability to $u(\theta)=\int_0^T y(t)\,\psi(t,\theta)\{\alpha(t)-\alpha(t,\theta)\}\,\mathrm{d}t$, with $y(t)$ as above. The maximum likelihood estimator, which solves $U_n(\widehat{\theta})=0$, converges in probability to the particular parameter value θ_0 that solves $u(\theta_0)=0$. We think of this as the 'least false' or 'agnostic' parameter value, and it minimises the distance measure

$$d[\alpha,\alpha(.,\theta)]=\int_0^T y\big[\alpha\{\log\alpha-\log\alpha(.,\theta)\}-\{\alpha-\alpha(.,\theta)\}\big]\,\mathrm{d}t \tag{2.3}$$

between true model and approximating model. This is proved in Hjort (1992). In later sections we shall also need the large-sample distribution, and quote the following result from Hjort (1992). Consider the $p\times p$-matrix $\psi^*(t,\theta)=\partial^2\log\alpha(t,\theta)/\partial\theta\partial\theta'$ and the function $E(t)=\int_0^t y(s)\psi(s,\theta_0)\{\alpha(s)-\alpha(s,\theta_0)\}\,\mathrm{d}s$ (in particular $E(0)=E(T)=0$). Define $p\times p$-matrices

$$\begin{aligned}J&=\int_0^T\Big[y(t)\psi(t,\theta_0)\psi(t,\theta_0)'\alpha(t,\theta_0)-y(t)\psi^*(t,\theta_0)\{\alpha(t)-\alpha(t,\theta_0)\}\Big]\,\mathrm{d}t,\\M&=\int_0^T\Big[y(t)\psi(t,\theta_0)\psi(t,\theta_0)'\alpha(t)+\{\psi(t,\theta_0)E(t)'+E(t)\psi(t,\theta_0)'\}\alpha(t,\theta_0)\Big]\,\mathrm{d}t.\end{aligned}$$

Then $\sqrt{n}(\widehat{\theta}-\theta_0)\to_d \mathcal{N}_p\{0,J^{-1}MJ^{-1}\}$. Note that under model conditions $\alpha(t)$ is indeed equal to $\alpha(t,\theta_0)$, the expressions for J and M simplify and become equal, and we have the more familiar-looking limit distribution $\mathcal{N}_p\{0,J^{-1}\}$, a result proved by Borgan (1984).

2.3. M-ESTIMATORS. We shall also need some general results about weighted likelihood estimators, from Hjort (1992, Section 5). Consider $\int_0^T G_n(t)\{\log\alpha(t,\theta)\,\mathrm{d}N(t) - Y(t)\alpha(t,\theta)\,\mathrm{d}t\}$ instead of the ordinary log-likelihood (which uses $G_n(t)=1$), and let $\widehat\theta_g$ maximise. This estimator also solves $\int_0^T G_n(t)\psi(t,\theta)\{\mathrm{d}N(t) - Y(t)\alpha(t,\theta)\,\mathrm{d}t\} = 0$, and belongs to the class of M-estimators for this counting process model, see Hjort (1985) and Andersen et al. (1993, Chapter VI). Assume that the weight function $G_n(t)$ is previsible and goes in probability to $g(t)$. The first result is that this estimator is consistent for the particular least false parameter value $\theta_{0,g}$ that minimises the distance function

$$d_g[\alpha,\alpha(.,\theta)] = \int_0^T gy\big[\alpha\{\log\alpha - \log\alpha(.,\theta)\} - \{\alpha - \alpha(.,\theta)\}\big]\,\mathrm{d}t, \tag{2.4}$$

a generalisation of (2.3). It also solves $\int_0^T g(t)y(t)\psi(t,\theta)\{\alpha(t) - \alpha(t,\theta)\}\,\mathrm{d}t = 0$. Secondly,

$$\sqrt{n}(\widehat\theta_g - \theta_{0,g}) \to_d \mathcal{N}_p\{0, J_g^{-1}M_gJ_g^{-1}\}, \tag{2.5}$$

where J_g and K_g are appropriate generalisations of those appearing above. In fact

$$\begin{aligned} J_g &= \int_0^T gy\big[\psi_0\psi_0'\alpha_0 - \psi_0^*(\alpha-\alpha_0)\big]\,\mathrm{d}t,\\ M_g &= \int_0^T \big[g^2y\psi_0\psi_0'\alpha + g\{\psi_0E_g' + E_g\psi_0'\}\alpha_0\big]\,\mathrm{d}t, \end{aligned} \tag{2.6}$$

in which $E_g(t) = \int_0^t gy\psi_0(\alpha-\alpha_0)\,\mathrm{d}s$, and where $\alpha_0 = \alpha(s,\theta_{0,g})$, $\psi_0 = \psi(s,\theta_{0,g})$. Note that both $E_g(0)$ and $E_g(T)$ are equal to 0, and that the expressions for J_g and M_g simplify when the model happens to be correct.

3. Dynamic likelihood estimation. Of course the parametric estimation method of 2.2 works best if the postulated model is adequate, i.e. if there really is a single θ_0 that secures $\alpha(s) \doteq \alpha(s,\theta_0)$ throughout $[0,T]$. Otherwise there is modelling bias present and it could for example be advantageous to use different θ_0's in different sub-intervals. We shall pursue a somewhat more extreme version of this idea, namely to fit a local estimate $\widehat\theta(s)$ for each s, and then use $\alpha(s,\widehat\theta(s))$ in the end.

3.1. DYNAMIC LIKELIHOOD. The dynamic or local likelihood estimation proposal is to use the M-estimator apparatus with a 'window function' $G_n(t) = g(t) = I\{t\in W\}$, where $W = [s-\frac12h, s+\frac12h]$ is a local interval around a given fixed s. So let $\widehat\theta(s)$ maximise

$$L_W(\theta) = \int_W \{\log\alpha(t,\theta)\,\mathrm{d}N(t) - Y(t)\alpha(t,\theta)\,\mathrm{d}t\}. \tag{3.1}$$

The resulting *dynamic likelihood hazard rate estimator* is

$$\widehat\alpha(s) = \alpha(s,\widehat\theta(s)). \tag{3.2}$$

Note that $L_W(\theta)$, the local log-likelihood at window W around s, is a bona fide log-likelihood, namely that based on those individuals that have survived up to $s-\frac12h$ and information about what happens to them in $[s-\frac12h, s+\frac12h]$. Showing this is not difficult by first noting that this group of individuals have

probability density	$= \alpha(t,\theta)\exp\{-\int_{s-h/2}^t \alpha(u,\theta)\,\mathrm{d}u\}$	for $t\in[s-\frac12h, s+\frac12h]$,
and chance	$= \exp\{-\int_{s-h/2}^{s+h/2}\alpha(u,\theta)\,\mathrm{d}u\}$	of further surviving $s+\frac12h$.

Consciously disregarded, for example, is information about individuals failing in $[0, s - \frac{1}{2}h)$. Including such a $[1 - \exp\{-A(s - \frac{1}{2}h, \theta)\}]^{n_0}$ term would have strengthened the likelihood and made our θ estimator more precise – but only if the parametric form of the hazard is correct also to the left of $s - \frac{1}{2}h$. The crucial idea here is to only trust the parametric form locally, and this leads to the (3.1) log-likelihood. Of course if h is large, which should correspond to trusting the model over the full range, then we get back the full log-likelihood and ordinary maximum likelihood.

The $\widehat{\theta}(s)$ estimator aims at the locally most suitable parameter value $\theta_0(W) = \theta_0(s)$ that minimises (2.4) with $g = I_W$, or, equivalently, solves $\int_W y(t)\psi(t,\theta)\{\alpha(t) - \alpha(t,\theta)\}\,\mathrm{d}t = 0$. Its large-sample behaviour is described by (2.5), which suggests

$$\mathrm{E}\widehat{\theta}(s) \doteq \theta_0(s), \quad \mathrm{VAR}\,\widehat{\theta}(s) \doteq J_W^{-1} M_W J_W^{-1}/n,$$

where J_W and M_W are as in (2.6) with $g(t) = I\{t \in W\}$. This transforms into corresponding properties for $\widehat{\alpha}(s)$ by Taylor expansions and delta-method arguments:

$$\begin{aligned} \mathrm{E}\alpha(s,\widehat{\theta}(s)) &\doteq \alpha(s,\theta_0(s)), \\ \mathrm{Var}\,\alpha(s,\widehat{\theta}(s)) &\doteq \alpha(s,\theta_0(s))^2\psi(s,\theta_0(s))' J_W^{-1} M_W J_W^{-1}\psi(s,\theta_0(s))/n. \end{aligned} \tag{3.3}$$

These approximations are valid if h is fixed and n is large. But we are also interested in becoming increasingly fine-tuned about the $s \pm \frac{1}{2}h$ interval as n grows. In order to study the bias and variance properties more closely, observe first that if $z(t)$ is a twice differentiable function defined in a neighbourhood of s, then $\int_W z(t)\,\mathrm{d}t \doteq z(s)h + \frac{1}{24}z''(s)h^3$ by a simple Taylor argument. From this and the defining equation for $\theta_0(s)$ we see that

$$y(s)\psi(s,\theta)\{\alpha(s) - \alpha(s,\theta)\} + \tfrac{1}{24}\big(y\psi(.,\theta)(\alpha - \alpha(s,\theta))\big)''(s)h^2 \doteq 0,$$

for the particular value $\theta = \theta_0(s)$, where $(fgh)''(s)$ means the second derivative of the $f(s)g(s)h(s)$ function evaluated at s. This implies generally that $\alpha(s,\theta_0(s)) = \alpha(s) + O(h^2)$. One can also show from this that

$$\mathrm{E}\alpha(s,\widehat{\theta}(s)) = \alpha(s,\theta_0(s)) + O(1/n) = \alpha(s) + O(h^2 + 1/n).$$

In order for the bias of the (3.2) estimator to go to zero it is therefore necessary that $h \to 0$ as $n \to \infty$.

At the moment we shall be content to give a bias formula for the case of a one-parameter family $\alpha(s,\theta)$, for which

$$\alpha(s,\theta_0(s)) \doteq \alpha(s) + \frac{h^2}{24}\Big[\alpha''(s) - \alpha_0''(s) + 2\{\alpha'(s) - \alpha_0'(s)\}\Big\{\frac{y'(s)}{y(s)} + \frac{\psi_0'(s)}{\psi_0(s)}\Big\}\Big]. \tag{3.4}$$

In this formula $\alpha_0'(s)$ means the derivative of $\alpha(s,\theta)$ w.r.t. s, and then inserted $\theta = \theta_0(s)$, and similarly for $\alpha_0''(s)$ and $\psi_0'(s)$. The case of multi-parametric classes of hazard rates is handled in Section 4.

Turning next to the variance matrix, one finds after using the (2.6) expressions and the previously established $O(h^2)$ result for the bias that $J_W = y(s)\psi_0(s)\psi_0(s)'\alpha(s,\theta_0(s))\,h + O(h^3)$ and $M_W = y(s)\psi_0(s)\psi_0(s)'\alpha(s)\,h + O(h^3)$, under smoothness assumptions on $\alpha(.)$ and $y(.)$, and writing for simplicity $\psi_0(s)$ for $\psi(s,\theta_0(s))$. We note here for the one-parameter case that $\mathrm{Var}\,\widehat{\theta}(s) \doteq (nh)^{-1}\{y(s)\alpha(s)\psi_0(s)^2\}^{-1}$, which in its turn implies

$$\mathrm{Var}\,\alpha(s,\widehat{\theta}(s)) \doteq \frac{1}{nh}\frac{\alpha(s)}{y(s)}. \tag{3.5}$$

Thus $nh \to \infty$ is necessary for the variance to go to zero, and this together with $h \to 0$ suffices for consistency of the (3.2) estimator.

3.2 Special case: estimating the local constant. The simplest model to try out is the one having $\alpha(s,\theta) = \theta$, a constant hazard. The local hazard estimate and its limit in probability are

$$\widehat{\alpha}(s) = \widehat{\theta}(s) = \frac{\int_W \mathrm{d}N(t)}{\int_W Y(t)\,\mathrm{d}t} \to_p \frac{\int_W y(t)\alpha(t)\,\mathrm{d}t}{\int_W y(t)\,\mathrm{d}t} = \theta_0(s), \tag{3.6}$$

again with $W = [s - \frac{1}{2}h, s + \frac{1}{2}h]$. The estimate is of the type total occurrence over total exposure, and the underlying local least false parameter is a local y-weighted average of the true hazard rate. By earlier efforts

$$\mathrm{E}\widehat{\alpha}(s) \doteq \alpha(s) + \frac{h^2}{24}\Big\{\alpha''(s) + 2\alpha'(s)\frac{y'(s)}{y(s)}\Big\} \quad \text{and} \quad \mathrm{Var}\,\widehat{\alpha}(s) \doteq \frac{1}{nh}\frac{\alpha(s)}{y(s)}. \tag{3.7}$$

This can also be verified directly. Further attention to these details is given in the next subsection.

A general remark about the dynamic likelihood method is that the particular parametric model used should be allowed to be quite crude, since we only employ it as a local approximation to the true hazard rate. This example illustrates this. (3.7) shows that even when $\alpha(.)$ simplistically is modelled as being locally a constant the result is a reasonable nonparametric estimator.

3.3. Kernel smoothed dynamic likelihood. The dynamic likelihood method of Sections 3.1 and 3.2 can be generalised to kernel smoothed variants. Let $K(u)$ be a symmetric kernel function with support $[-\frac{1}{2}, \frac{1}{2}]$ and integral 1. Define the local kernel smoothed likelihood estimator $\widehat{\theta}(s)$ to maximise

$$\widetilde{L}_W(\theta) = \int_W K\big(h^{-1}(t-s)\big)\big\{\log\alpha(t,\theta)\,\mathrm{d}N(t) - Y(t)\alpha(t,\theta)\,\mathrm{d}t\big\}. \tag{3.8}$$

The hazard rate estimator is as in (3.2) with this more general estimate of θ. The previously defined local likelihood estimate corresponds to the special case $K(u) = 1$ on $[-\frac{1}{2}, \frac{1}{2}]$. This uniform choice has perhaps some special appeal since the dynamic log-likelihood $L_W(\theta)$ then can be interpreted as a genuine log-likelihood for a subgroup of the individuals under study. The current smoothed likelihood is more of a mathematical construction, but turns out to produce estimators with slightly better properties, for good choices of $K(u)$.

We can draw on the general results of 2.3 to find approximate bias and variance for the maximiser of (3.8). Let $\beta_K = \int u^2 K(u)\,\mathrm{d}u$ and $\gamma_K = \int K(u)^2\,\mathrm{d}u$. (2.6) with Taylor expansion quickly gives

$$\begin{aligned} J_W &= y(s)\psi_0(s)\psi_0(s)'\alpha(s,\theta_0(s))\,h + O(h^3), \\ M_W &= \gamma_K y(s)\psi_0(s)\psi_0(s)'\alpha(s)\,h + O(h^3), \end{aligned} \tag{3.9}$$

The multi-parameter case requires more precise expansions, since the inverse of J_W is needed and $\psi_0(s)\psi_0(s)'$ has rank 1. Leaving the multi-parameter case for Section 4, consider an arbitrary one-parameter family $\alpha(s,\theta)$, where $\widehat{\theta}(s)$ solves $\int_W K(h^{-1}(t-s))\psi(t,\theta)\{\mathrm{d}N(t) - Y(t)\alpha(t,\theta)\,\mathrm{d}t\} = 0$. It aims at the locally least false $\theta_0 = \theta_0(W)$ that solves $\int_W K(h^{-1}(t-s))\psi(t,\theta)y(t)\{\alpha(t) - \alpha(t,\theta)\}\,\mathrm{d}t = 0$, or $\int K(u)\psi(s+hu,\theta)y(s+hu)\{\alpha(s+hu) - \alpha(s+hu,\theta)\}\,\mathrm{d}u = 0$. Taylor expansion shows that $\int K(u)z(s+hu)\,\mathrm{d}u = z(s) + \frac{1}{2}\beta_K h^2 z''(s) + O(h^4)$ for smooth $z(.)$ functions, and this, in conjunction with (3.3) and (3.9), leads to

$$\mathrm{E}\alpha(s,\widehat{\theta}(s)) \doteq \alpha(s) + \tfrac{1}{2}h^2\beta_K b(s) \quad \text{and} \quad \mathrm{Var}\,\alpha(s,\widehat{\theta}(s)) \doteq \frac{\gamma_K}{nh}\frac{\alpha(s)}{y(s)}, \tag{3.10}$$

where the bias factor is

$$b(s)=\alpha''(s)-\alpha_0''(s)+2\{\alpha'(s)-\alpha_0'(s)\}\Big\{\frac{y'(s)}{y(s)}+\frac{\psi_0'(s)}{\psi_0(s)}\Big\}. \qquad (3.11)$$

The fact that $\int uK(u)\,\mathrm{d}u=0$ is used here. When $K(u)$ is uniform we get back (3.4) and (3.5). Observe that the approximate variance does not depend on the parametric family employed (to the order of approximation used).

3.4. Special case: local constant with a kernel. Let us illustrate this for the special case where $\alpha(s,\theta)=\theta$. Then

$$\widehat\alpha(s)=\widehat\theta(s)=\frac{\int_W K(h^{-1}(t-s))\,\mathrm{d}N(t)}{\int_W K(h^{-1}(t-s))Y(t)\,\mathrm{d}t}=\frac{\sum_{|x_i-s|\le h/2}K(h^{-1}(x_i-s))\delta_i}{\sum_{i=1}^n\int_{W\cap[0,x_i]}K(h^{-1}(t-s))\,\mathrm{d}t}, \qquad (3.12)$$

a locally weighted occurrence over locally weighted exposure estimate. Here and later on x_i denotes the observed value of $X_i=\min(X_i^0,C_i)$. Previous efforts give

$$E\widehat\alpha(s)=\alpha(s)+\tfrac12\beta_Kh^2\{\alpha''(s)+2\alpha'(s)y'(s)/y(s)\}+O(h^4),$$

and variance $\gamma_K(nh)^{-1}\alpha(s)/y(s)$ as before. This generalises (3.7).

One theoretical advantage that (3.12) has over the (3.6) estimator is that it has smaller mean squared error, for several natural choices of kernel K, see 6.1. A more immediate practical advantage is that K can be chosen to make it smoother than the (3.6) version, which is discontinuous at time points s where $s\pm\frac12h$ is equal to observed failure times. (3.12) is continuous when $K(\pm\frac12)=0$, and has a continuous derivative if K is chosen such that $K'(\pm\frac12)=0$.

4. Dynamic likelihood for multi-parameter families. The dynamic likelihood and kernel smoothed dynamic likelihood ideas of Section 3 can be applied for any smooth parametric family of hazards, but the basic bias and variance properties have so far only been derived for one-parameter families. We saw in (3.9), for example, that the multi-parameter case requires more careful expansions. It is not clear at the outset that we gain in precision by smoothing e.g. a two-parameter hazard family. We should perhaps expect larger windows to be required to be able to estimate both parameters with reasonable precision.

4.1. A running Gompertz estimator. The hazard function model $\alpha(t)=a\exp(\beta t)$ is sometimes called the Gompertz model. Concentrating on a fixed s with fixed window $W=s\pm\frac12h$, we may reparametrise the hazard as

$$\alpha(t,\theta,\beta)=a\exp(\beta s)\exp(\beta(t-s))=\theta\exp(\beta(t-s))\quad\text{for } t\in[s-\tfrac12h,s+\tfrac12h], \qquad (4.1)$$

and interpret θ as the 'local level' and β as the 'local slope'. Define $\widehat\theta(s)$ and $\widehat\beta(s)$ as those maximising the kernel smoothed dynamic likelihood

$$\widetilde L_W(\theta,\beta)=\int_W K(h^{-1}(t-s))\big[\{\log\theta+\beta(t-s)\}\,\mathrm{d}N(t)-Y(t)\theta\exp(\beta(t-s))\,\mathrm{d}t\big]. \qquad (4.2)$$

One has

$$\widehat\theta(s,\beta)=\frac{\int_W K(h^{-1}(t-s))\,\mathrm{d}N(t)}{\int_W K(h^{-1}(t-s))Y(t)\exp(\beta(t-s))\,\mathrm{d}t}, \qquad (4.3)$$

and the resulting profile dynamic likelihood can be shown to be concave in β, and accordingly not very difficult to maximise. The maximiser found is then inserted into (4.3) to give $\widehat{\theta}(s)$. Note that the general dynamic likelihood recipe gives

$$\widehat{\alpha}(s) = \alpha(s, \widehat{\theta}(s), \widehat{\beta}(s)) = \widehat{\theta}(s), \tag{4.4}$$

simply, so the β parameter estimate is only somewhat silently present. From the general theory of Section 2.3 we know that $\widehat{\theta}(s)$ and $\widehat{\beta}(s)$ aim at certain appropriate least false parameter values $\theta_0 = \theta_0(s)$ and $\beta_0 = \beta_0(s)$, depending on the window W, and that $\sqrt{n}(\widehat{\theta}(s) - \theta_0, \widehat{\beta}(s) - \beta_0)$ goes to a zero-mean normal with covariance matrix $J_W^{-1} M_W J_W^{-1}$. Here J_W and M_W are as in (2.6) with $g = I_W$. We now set out to provide informative approximations for these and for the least false local parameters.

The least false parameter values are such that they solve the two equations $\int_W K(h^{-1}(t - s))\psi(t, \theta_0, \beta_0)y(t)\{\alpha(t) - \theta_0 \exp(\beta_0(t - s))\}\,\mathrm{d}s = 0$, where $\psi(t, \theta, \beta) = (1/\theta, t - s)$. The first equation gives

$$\theta_0 = \frac{\int_W K(h^{-1}(t-s))y(t)\alpha(t)\,\mathrm{d}t}{\int_W K(h^{-1}(t-s))y(t)\exp(\beta_0(t-s))\,\mathrm{d}t} = \frac{\int K(u)y(s+hu)\alpha(s+hu)\,\mathrm{d}u}{\int K(u)y(s+hu)\exp(\beta_0 hu)\,\mathrm{d}u},$$

where the latter integrals are over the support $[-\frac{1}{2}, \frac{1}{2}]$ for the kernel function $K(u)$. Upon using $\int K(u)z(s + hu)\,\mathrm{d}u = z(s) + \frac{1}{2}\beta_K h^2 z''(s) + O(h^4)$ again, one finds after some calculations that

$$\theta_0 \doteq \alpha(s) + \tfrac{1}{2}\beta_K h^2[\{y(t)\alpha(t)\}''(s) - \alpha(s)\{y(t)\exp(\beta_0(t-s))\}''(s)]/y(s) = \alpha(s) + \tfrac{1}{2}\beta_K h^2 b(s, \beta_0),$$

say, up to $O(h^4)$ terms, where in fact $b(s, \beta_0) = \alpha''(s) - \alpha(s)\beta_0^2 + 2\{y'(s)/y(s)\}\{\alpha'(s) - \alpha(s)\beta_0\}$. Similarly the second equation gives $\int K(u)u\,y(s + hu)\{\alpha(s + hu) - \theta_0 \exp(\beta_0 hu)\}\,\mathrm{d}u = 0$, which upon using $\int K(u)u\,z(s + hu)\,\mathrm{d}u = \beta_K z'(s)h + O(h^3)$ delivers $\beta_0 = \alpha'(s)/\alpha(s) + O(h^2)$. This can be plugged into $b(s, \beta_0)$ above to give

$$\alpha(s, \theta_0(s), \beta_0(s)) = \theta_0 = \alpha(s) + \tfrac{1}{2}\beta_K h^2\{\alpha''(s) - \alpha'(s)^2/\alpha(s)\} + O(h^4). \tag{4.5}$$

Note that the bias is only $O(h^4)$ at s if the true $\alpha(.)$ is locally like a Gompertz hazard.

Next consider the matrices that determine the approximate variances for $\widehat{\theta}(s)$ and $\widehat{\beta}(s)$. From (2.6),

$$\begin{aligned} J_W = \int_W K(h^{-1}(t-s))y(t)\Big[&\begin{pmatrix} 1/\theta_0^2 & (t-s)/\theta_0 \\ (t-s)/\theta_0 & (t-s)^2 \end{pmatrix} \theta_0 \exp(\beta_0(t-s)) \\ &+ \begin{pmatrix} 1/\theta_0^2 & 0 \\ 0 & 0 \end{pmatrix} \{\alpha(t) - \theta_0 \exp(\beta_0(t-s))\}\Big]\,\mathrm{d}t. \end{aligned}$$

We find

$$J_{11} = h\int K(u)y(s+hu)\theta_0^{-2}\alpha(s+hu)\,\mathrm{d}u = hy(s)\alpha(s)/\theta_0^2 + \tfrac{1}{2}\beta_K h^3(y\alpha)''(s)/\theta_0^2 + O(h^5)$$

for the (1,1) element. Similar calculations give

$$J_W = h\Big[\begin{pmatrix} y(s)\alpha(s)/\theta_0^2 & 0 \\ 0 & 0 \end{pmatrix} + \beta_K h^2 \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}\Big] + O(h^5),$$

where in fact $a_{11} = \frac{1}{2}(y\alpha)''(s)/\theta_0^2$, $a_{12} = y'(s) + y(s)\beta_0$, and $a_{22} = y(s)\theta_0$. Next look at

$$\begin{aligned} M_W = \int_W \Big[&K(h^{-1}(t-s))^2 y(t) \begin{pmatrix} 1/\theta_0^2 & (t-s)/\theta_0 \\ (t-s)/\theta_0 & (t-s)^2 \end{pmatrix} \alpha(t) \\ &+ K(h^{-1}(t-s)) \begin{pmatrix} 2E_1(t)/\theta_0 & E_2(t)/\theta_0 + E_1(t)(t-s) \\ E_2(t)/\theta_0 + E_1(t)(t-s) & 2(t-s)E_2(t) \end{pmatrix}\Big]\,\mathrm{d}t, \end{aligned}$$

where $E_1(t)$ and $E_2(t)$ are the components of the $E(t)$ function defined after (2.6). It turns out that $E_1(t) = O(h^3)$ while $E_2(t) = O(h^4)$, so the second part of the M_W matrix is of a smaller size than the first. We find after some expansion work that

$$M_W = h\Big[\begin{pmatrix} \gamma_K y(s)\alpha(s)/\theta_0^2 & 0 \\ 0 & 0 \end{pmatrix} + \delta_K h^2 \begin{pmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{pmatrix}\Big] + \begin{pmatrix} O(h^4) & O(h^5) \\ O(h^5) & O(h^5) \end{pmatrix},$$

where $\gamma_K = \int K(u)^2\,du$ and $\delta_K = \int u^2 K(u)^2\,du$, and where $b_{11} = \frac{1}{2}(y\alpha)''(s)/\theta_0^2$, $b_{12} = (y\alpha)'(s)/\theta_0$, and $b_{22} = y(s)\alpha(s)$.

To reach expressions for $J_W^{-1} M_W J_W^{-1}$ we need to work with a matrix of the form $(cE_{11} + h^2 A)^{-1}(dE_{11} + h^2 B)(cE_{11} + h^2 A)^{-1}$, where E_{11} is the matrix with 1 as (1,1) element and zeros elsewhere. The result, after lengthy but elementary calculations, is of the form

$$J_W^{-1} M_W J_W^{-1} = \begin{pmatrix} h^{-1}\gamma_K \alpha(s)/y(s) + c_{11}h + O(h^2) & h^{-1}c_{12} + O(h) \\ h^{-1}c_{12} + O(h) & h^{-3}(\delta_K/\beta_K^2)/\{y(s)\alpha(s)\} + h^{-1}c_{22} + O(1) \end{pmatrix},$$

for certain c_{ij}. We are primarily interested in the approximate variance for the local $\widehat{\theta}(s)$, in view of (4.4), and this is $(nh)^{-1}\gamma_K\alpha(s)/y(s) + O(h/n)$, precisely as in the one-dimensional case (3.10). Hence bias and variance properties are of the same form as in the one-dimensional case, but with a different bias factor, inherited from the model one smooths.

4.2. Dynamic likelihood for a general multi-parameter model. Suppose the hazard rate model is of the type $\alpha(t) = a\gamma(t,\beta)$, i.e. a constant parameter a times a function which depends on a possibly multi-dimensional parameter β but not on a. Reparametrise locally to

$$\alpha(t) = a\gamma(s,\beta)\{\gamma(t,\beta)/\gamma(s,\beta)\} = \theta\exp\{C(t,\beta) - C(s,\beta)\} \quad \text{for } t \in [s - \tfrac{1}{2}h, s + \tfrac{1}{2}h]. \tag{4.6}$$

The score type function of the model is $\psi(t,\theta,\beta) = (1/\theta, C^*(t,\beta) - C^*(s,\beta))$, where $C^*(t,\beta) = \frac{\partial}{\partial\beta}C(t,\beta)$. Notice that the local estimate $\widehat{\beta}(s)$ is only 'silently present', in that it is used only in conjunction with finding the local $\widehat{\theta}(s)$, as with (4.3) and (4.4).

Now the calculations of the Gompertz model above can be repeated with the necessary modifications. As in that case one finds $\theta_0 = \alpha(s) + \frac{1}{2}\beta_K h^2 b(s,\beta_0)$ with a similar $b(s,\beta_0)$, and also that $\theta_0 c(s,\beta_0) = \alpha'(s) + O(h^2)$, where $c(t,\beta) = \frac{\partial}{\partial t}C(t,\beta)$. This leads to

$$\begin{aligned} \mathrm{E}\widehat{\alpha}(s) &= \alpha(s) + \tfrac{1}{2}\beta_K h^2\big[\alpha''(s) - \alpha(s)\{c(s,\beta_0)^2 + c'(s,\beta_0)\} \\ &\qquad + 2\{y'(s)/y(s)\}\{\alpha'(s) - \alpha(s)c(s,\beta_0)\}\big] + O(h^4) \\ &= \alpha(s) + \tfrac{1}{2}\beta_K h^2\big[\alpha''(s) - \theta_0\{c(s,\beta_0)^2 + c'(s,\beta_0)\} + 2\{y'(s)/y(s)\}O(h^2)\big] + O(h^4) \\ &= \alpha(s) + \tfrac{1}{2}\beta_K h^2\{\alpha''(s) - \alpha_0''(s)\} + O(h^4), \end{aligned}$$

where $\alpha_0''(s)$ is the second derivative of the model's hazard rate $\theta\exp\{C(t,\beta) - C(s,\beta)\}$ w.r.t. t, evaluated at s, and with the local least false parameters $\theta_0 = \theta_0(W)$ and $\beta_0 = \beta_0(W)$ inserted. We also have $\alpha_0''(s) = \theta_0\{c(s,\beta_0)^2 + c'(s,\beta_0)\} = \alpha'(s)^2/\alpha(s) + \alpha(s)c'(s,\beta_0) + O(h^2)$. Note that (4.5) is a special case.

One next finds that the (1,1) element of the appropriate $J_W^{-1} M_W J_W^{-1}$ matrix is yet again equal to $h^{-1}\gamma_K\alpha(s)/y(s) + O(h)$, albeit with a more involved expression for the constant in the secondary $O(h)$ term. The basic properties for the dynamic likelihood estimator are accordingly once more of the familiar type

$$\mathrm{E}\widehat{\alpha}(s) = \alpha(s) + \tfrac{1}{2}\beta_K h^2\{\alpha''(s) - \alpha_0''(s)\} + O(h^4) \quad \text{and} \quad \mathrm{Var}\,\widehat{\alpha}(s) \doteq \frac{\gamma_K}{nh}\frac{\alpha(s)}{y(s)}, \tag{4.7}$$

with a bias term $\frac{1}{2}\beta_K h^2 b(s)$ appropriate to the parametric model employed. As noted above there are also alternative useful expressions for the $b(s)$ term, since we can move out $O(h^2)$ terms.

4.3. A RUNNING WEIBULL ESTIMATOR. As an example of the previous general machinery, consider the Weibull model, which uses $\alpha(t) = abt^{b-1}$ for certain parameters a and b. We reparametrise to $\alpha(t) = \theta(t/s)^\beta$, where $\theta = abs^{b-1}$ and $\beta = b-1$. Let $\widehat{\theta}(s)$ and $\widehat{\beta}(s)$ maximise the kernel smoothed dynamic log-likelihood

$$\int_W K(h^{-1}(t-s))\big[\{\log\theta + \beta(\log t - \log s)\}\,\mathrm{d}N(t) - Y(t)\theta(t/s)^\beta\,\mathrm{d}t\big].$$

Then use $\widehat{\alpha}(s) = \widehat{\theta}(s)$ in the end. The results above imply $\theta_0\beta_0 = s\alpha'(s) + O(h^2)$, and the bias is

$$\tfrac{1}{2}\beta_K h^2\{\alpha''(s) - \theta_0(\beta_0^2 - \beta_0)/s^2\} + O(h^4) = \tfrac{1}{2}\beta_K h^2\{\alpha''(s) - \alpha'(s)^2/\alpha(s) + \alpha'(s)/s\} + O(h^4). \quad (4.8)$$

The approximate variance is yet again $\gamma_K(nh)^{-1}\alpha(s)/y(s)$. Note that the bias is only $O(h^4)$ if the true hazard is locally a Weibull hazard.

4.4. A DYNAMIC NONINCREASING ESTIMATOR. As a final example, consider the simple frailty model with hazard rate $a/(1+\beta t)$. This is the hazard rate in a population where each individual has a constant hazard rate but where these vary in the population according to a gamma distribution with mean a and variance β. The local parametrisation is $\theta(1+\beta s)/(1+\beta t)$ for $t \in [s-\frac{1}{2}h, s+\frac{1}{2}h]$. Even though the model can tolerate a small negative value for β we shall in this example take it a priori as a nonnegative quantity. So let $\widehat{\theta}(s)$ and $\widehat{\beta}(s)$ maximise

$$\int_W K(h^{-1}(t-s))\big[\{\log\theta + \log(1+\beta s) - \log(1+\beta t)\}\,\mathrm{d}N(t) - Y(t)(1+\beta s)\,\mathrm{d}t/(1+\beta t)\big],$$

and use $\widehat{\alpha}(s) = \widehat{\theta}(s)$ in the end. Then

$$\mathrm{E}\widehat{\alpha}(s) = \alpha(s) + \tfrac{1}{2}\beta_K h^2\{\alpha''(s) - 2\alpha'(s)^2/\alpha(s)\} + O(h^4) \quad\text{and}\quad \mathrm{Var}\,\widehat{\alpha}(s) \doteq \frac{\gamma_K}{nh}\frac{\alpha(s)}{y(s)}. \quad (4.9)$$

5. Comparison with the traditional kernel estimator. Estimators developed in Sections 3 and 4 can now be compared with the classical nonparametric estimator.

5.1. THE SMOOTHED NELSON–AALEN ESTIMATOR. The traditional nonparametric estimator is a kernel smooth of the (2.1) estimator of the cumulative,

$$\widetilde{\alpha}(s) = \int_W h^{-1}K(h^{-1}(t-s))\,\mathrm{d}\widehat{A}(t) = \sum_{|x_i - s|\le h/2} h^{-1}K(h^{-1}(x_i - s))\delta_i/Y(s_i). \quad (5.1)$$

When $K(u)$ is uniform this becomes $\{\widehat{A}(s+\frac{1}{2}h) - \widehat{A}(s-\frac{1}{2}h)\}/h$, for example. From the properties of $\widehat{A}$ reviewed in Section 2 it is not difficult to derive

$$\mathrm{E}\widetilde{\alpha}(s) \doteq \alpha(s) + \tfrac{1}{2}\beta_K h^2\alpha''(s) \quad\text{and}\quad \mathrm{Var}\,\widetilde{\alpha}(s) \doteq \frac{\gamma_K}{nh}\frac{\alpha(s)}{y(s)}. \quad (5.2)$$

See also Ramlau-Hansen (1983), Yandell (1983), and Tanner and Wong (1983), who all studied estimators of this type, and Andersen et al. (1993, Chapter IV). It is remarkable that the new estimators $\alpha(s, \widehat{\theta}(s))$ and the traditional one have exactly the same approximate variance and the same type of approximate bias, when they use the same kernel and the same bandwidth; see (3.10), (4.4) and (4.7).

5.2. WHEN IS THE DYNAMIC METHOD ALWAYS BETTER? The dynamic kernel smoothed likelihood estimator has approximate bias $\frac{1}{2}\beta_K h^2 b(s)$, with a $b(s)$ function depending on the underlying parametric family used. In view of the comparison already made above it follows that the new method is always as good as or better than the Ramlau-Hansen–Yandell estimator, with the same kernel and the same window size, provided only $|b(s)| \le |\alpha''(s)|$.

For the one-parameter situation the question is whether

$$|\alpha''(s) - \alpha_0''(s) + 2\{\alpha'(s) - \alpha_0'(s)\}\{y'(s)/y(s) + \psi_0'(s)/\psi_0(s)\}| \le |\alpha''(s)|. \tag{5.3}$$

This can easily happen if the parametric family is only moderately acceptable. For the special case (3.12), for which $\alpha_0'(s)$, $\alpha_0''(s)$ and $\psi_0'(s)$ are absent, the inequality might take place in regions where α is convex and increasing, or concave and decreasing. When there is no censoring $y = \exp(-A)$ and $y'/y = -\alpha$, and then the criterion for when (3.12) is better than then the traditional (5.1) becomes $0 \le \alpha(s)\alpha'(s)/\alpha''(s) \le 1$.

For the multi-parametric families of Section 4 we have established $b(s) = \alpha''(s) - \alpha_0''(s)$, with $\alpha_0''(s)$ stemming from the model used, and with certain useful alternative expressions. The dynamic likelihood estimator is better than (5.1), when the same h is used, whenever $|\alpha''(s) - \alpha_0''(s)| \le |\alpha''(s)|$, which can be rewritten

$$0 \le \frac{\alpha_0''(s)}{\alpha''(s)} \le 2 \tag{5.4}$$

when the second derivative of the true hazard is not zero. If the parametric model used is locally correct, then the ratio is 1 and the bias is $O(h^4)$ only. If we take 'the parametric model is roughly adequate' to mean (5.4), then indeed the dynamic likelihood estimator is always better than (5.1) under such circumstances, for each h and each K.

For the two-parametric running Gompertz estimator (4.4) the criterion is

$$0 \le \alpha'(s)^2/\{\alpha(s)\alpha''(s)\} \le 2, \tag{5.5}$$

and the ratio is 1 exactly for Gompertz hazards $a\exp(\beta s)$. If for example $\alpha(s) = a + be^{cs}$ is of Gompertz–Makeham form, then the ratio is $be^{cs}/(a+be^{cs})$ and well inside $(0, 2)$, showing that (4.4) will be better than (5.1) for all such hazards. Similarly, for the two-parametric running Weibull estimator of 4.3, the new estimator is always better than (5.1) in regions where

$$0 \le \alpha'(s)^2/\{\alpha(s)\alpha''(s)\} - \alpha'(s)/\{s\alpha''(s)\} \le 2, \tag{5.6}$$

and the function appearing in the middle is equal to 1 exactly for Weibull hazards. And finally the criterion for when the estimator of 4.4 is always better than (5.1) is

$$0 \le \alpha'(s)^2/\{\alpha(s)\alpha''(s)\} \le 1. \tag{5.7}$$

5.3. VICINITY OF PARAMETRIC MODEL. As explained above one can expect the methods developed to perform better than the traditional (5.1), and surely also better than other purely nonparametric estimators, if only the parametric model used is roughly adequate. So far this statement has been referring to a comparison when the two methods have used the same window width h. But in these cases we would really expect the new methods to perform not only better but much better, by carefully choosing a good window width. When the bias is smaller we can select a larger window and be rewarded with smaller variability, cf. the mean squared error calculations of 6.1.

It should also be possible to improve on the convergence rate if the true hazard lies suitably close to the parametric family. A mathematical framework to make this notion more precise could be as follows. There is a sequence of experiments where at stage n there are data (X_i, δ_i) on n individuals coming from a distribution with true hazard $\alpha(.) = \alpha_n(.)$. Suppose this is such that the bias factor is $b(s) = b_0(s)n^{-\epsilon}$, for some $\epsilon \in [0, \frac{1}{2})$. Then the best achievable mean squared error is of size proportional to $n^{-(4-2\epsilon)/5}$, and this happens with h chosen as a suitable $h_0 = cY_n(s)^{-(1-2\epsilon)/5}$. A cross validation or other clever h selection scheme will pick this up, cf. the following section. The best nonparametric convergence rate is $n^{-4/5}$, for both point-wise and integrated mean squared error, and these calculations show that this can be improved upon for alternatives in the vicinity of the parametric model. If $\alpha''(s) - \alpha_0''(s) = O(n^{-1/4})$, for example, then the mean squared error is $O(n^{-9/10})$, and alternatives lying almost $O(n^{-1/2})$ away, in the above sense, are estimated with almost full parametric $O(n^{-1})$ precision. The point of comparison is that the traditional (5.1) estimator will still only accomplishes $O(n^{-4/5})$ precision for these hazard rates.

6. Choosing the smoothing parameter. We have defined $\widehat{\alpha}(s) = \alpha(s, \widehat{\theta}(s))$ for given parametric family $\alpha(s, \theta)$ and kernel K. The most decisive influence on the estimator is due to the smoothing parameter h.

6.1. Mean squared error calculations. By (3.10) and (4.7) the approximate mean squared error is of the form

$$\mathrm{E}\{\widehat{\alpha}(s) - \alpha(s)\}^2 \doteq \frac{1}{4}h^4\beta_K^2 b(s)^2 + \frac{\gamma_K}{nh}\frac{\alpha(s)}{y(s)},$$

where $b(s)$ is the appropriate bias factor stemming from the parametric recipe used. The mean squared error is minimised for

$$h_0(s) = \left\{\frac{\gamma_K}{\beta_K^2}\frac{\alpha(s)}{b(s)^2}\right\}^{1/5}\frac{1}{\{ny(s)\}^{1/5}}. \tag{6.1}$$

The resulting minimal mean squared error is $\frac{5}{4}(\beta_K\gamma_K^2)^{2/5}\alpha(s)^{4/5}b(s)^{2/5}/\{ny(s)\}^{4/5}$. Different choices of reasonable kernels give about the same result, but the best choice, managing to minimise $\beta_K\gamma_K^2$ among kernels on $[-\frac{1}{2}, \frac{1}{2}]$ with integral 1, is the Bartlett–Yepanechnikov kernel $K_0(u) = \frac{3}{2}(1 - 4u^2)$ on $[-\frac{1}{2}, \frac{1}{2}]$. The resulting $\alpha(s, \widehat{\theta}(s))$ estimator is continuous in s but its derivative will have discontinuities at points s where $s \pm \frac{1}{2}h$ hits an observed failure time.

We have seen that the new methods can outperform the traditional ones by reducing the bias, say from $b_{\mathrm{trad}}(s) = \alpha''(s)$ of (5.1) to possibly smaller $b(s) = \alpha''(s) - \alpha_0''(s)$ for those of Section 4. It is therefore of interest to note that the squared bias makes up 20% and the variance 80% of the approximate mean squared error, so bias reduction can perhaps not be expected to give dramatic gains. If $b(s) = \frac{1}{2}b_{\mathrm{trad}}(s)$, for example, then the best theoretical window width becomes $h_0 = 1.32\, h_{0,\mathrm{trad}}$, and the mean squared error is reduced with 24%.

The h_0 formula cannot be put to direct use since it depends on the hazard rate itself, but the rate of convergence to zero of mean squared error becomes the optimal $n^{-4/5}$ when h is chosen proportional to $n^{-1/5}$. The formula indicates that h should be chosen proportional to $Y(s)^{-1/5}$ in practice. One possibility is to use $h_n = cY(s)^{-1/5}$ and try to minimise a global criterion like $\mathrm{E}\int_0^T w(s)\{\widehat{\alpha}(s)-\alpha(s)\}^2\,ds$ w.r.t. c. The result is a local variable kernel smoothed likelihood estimator with

$$h_n = \Big\{\frac{\gamma_K}{\beta_K^2}\frac{\int_0^T w(s)y(s)^{-4/5}\alpha(s)\,ds}{\int_0^T w(s)y(s)^{-4/5}b(s)^2\,ds}\Big\}^{1/5}\frac{1}{Y(s)^{1/5}}. \tag{6.2}$$

We might for example choose weight function $w(s)=y(s)^{4/5}$ here, this being inversely proportional to the optimal mean squared error, and this simplifies (6.2). The nominator integral can be estimated with $n^{1/2}$-precision. Some pilot estimate $\widehat{\alpha}_{\rm pil}$, like the (5.1) estimator with an overall twice differentiable kernel K_2 and a somewhat large h_2, can be used to estimate the denominator integral. A final adjustment is needed since $\int_0^T \widehat{b}(s)^2\,ds$ will be biased. Working out expressions for the bias of $\int_0^T \widehat{b}(s)^2\,ds$ as an estimator of $\int_0^T b(s)^2\,ds$ takes some efforts, for the most interesting estimators of Sections 3 and 4, but is within comfortable reach of Ramlau-Hansen's (1983) methods. In the end this produces a practical algorithm of 'plug-in' type.

The discussion above is valid for one-parameter families and also for the class of multi-parameter families considered in Section 4, since the approximate variance of $\widehat{\alpha}(s)$ also in these situations turned out to be of the form $\gamma_K(nh)^{-1}\alpha(s)/y(s)$. In the models of Section 4 there is a 'local position' parameter θ and a 'local slope' parameter β. Note that the local slope estimate $\widehat{\beta}(s)$ has quite larger variance than the local position estimate $\widehat{\theta}(s)$. In the running Gompertz case the slope estimate has variance proportional to $n^{-1}h^{-3}/\{y(s)\alpha(s)\}$, for example. The best window size for β estimation is proportional to $Y(s)^{-1/7}$, but the best size for θ estimation, which is our primary concern, is still proportional to $Y(s)^{-1/5}$. These quantitative results are perhaps as expected, in view of similar results from density estimation and nonparametric regression. They also suggest that the $\widehat{\beta}(s)$ that is inserted in $\widehat{\theta}(s,\beta)$ of (4.3) to produce the final (4.4) can be quite variable if produced from $Y(s)^{-1/5}$-windows, and it may be advantageous to use a separate estimation scheme for estimation of this parameter, with somewhat larger windows. See also Remarks 7B and 7G.

The reasoning that led to (6.1) and (6.2) is also pertinent for the problem of choosing h in the Ramlau-Hansen–Yandell estimator (5.1), since the bias and variance structure are of the same type, only with $b(s)=\alpha''(s)$ instead. We also note that there are other ways of obtaining a data-driven $h_n(s)$, like cross validation or bootstrapping, but these are not discussed further here. References to cross validation techniques for the (5.1) estimator are Nielsen (1990) and Grégoire (1993), and these techniques should carry over at least to the (3.12) estimator.

6.2. Local goodness of fit testing. If the parametric model doesn't fit well the dynamic likelihood hazard estimator is still reasonable, and resembles the nonparametric Nelson–Aalen smoother (5.1) in performance. At the same time our method is able to outperform (5.1) as well as other purely nonparametric methods in cases where the parametric family $\alpha(s,\theta)$ used is only roughly acceptable, as explained in Section 5. In such cases the size of the bias is small, which by (6.2) suggests using quite a large bandwidth h, which in its turn almost amounts to using an ordinary parametric method.

A natural but somewhat elaborate strategy is to choose $h=\widehat{h}(s)$ to be the smallest h for which some convenient goodness of fit criterion rejects the parametric model on $s\pm\frac{1}{2}h$. The ultimate case is of course no detectable departure from the model over the full range $[0,T]$, which then leads to using $h=\infty$, i.e. ordinary parametric estimation $\alpha(s,\widehat{\theta}_{[0,T]})$, say.

Hjort (1985, 1990) has developed classes of goodness of fit tests for general parametric counting process models, and these are indeed presented there as tests of validity over the full range $[0,T]$. Similar mathematical techniques can however be used to construct procedures that check model adequacy over a general $[a,b]$ interval, and some such are presented next. This apparatus would then be used with $[a, b] = [s - \frac{1}{2}h, s + \frac{1}{2}h]$, mostly, but to get the running estimator started one would look for model adequacy over $[0, b]$ intervals first, cf. Remark 7A.

6.3. ONE-PARAMETER FAMILIES. Consider dynamic smoothing of an arbitrary one-dimensional parametric family $\alpha(u, \theta)$. Let $\widehat{\theta}_{[a,b]}$ be the local maximum likelihood estimator using only $[a, b]$ information, i.e. it solves $\int_a^b \psi(u, \theta)\{\mathrm{d}N(u) - Y(u)\alpha(u, \theta)\,\mathrm{d}u\} = 0$. Let

$$D_n(t) = n^{-1/2} \int_a^t \psi(u, \widehat{\theta}_{[a,b]})\{\mathrm{d}N(u) - Y(u)\alpha(u, \widehat{\theta}_{[a,b]})\,\mathrm{d}u\} \quad \text{for } t \in [a, b].$$

It uses the 'basic martingale' $\mathrm{d}N(u) - \alpha(u, \theta)\,\mathrm{d}u$ and is able to pick up departures from the parametric model. Notice that $D_n(.)$ starts and ends at zero. Methods of Hjort (1990) can be used to prove that $D_n(.)$, if indeed the model holds on $[a, b]$, converges to a zero-mean Gaußian process $D(.)$ with covariance function $\mathrm{cov}\{D(t_1), D(t_2)\} = \tau^2(b)\{p(t_1 \wedge t_2) - p(t_1)p(t_2)\}$, in which $\tau^2(t) = \int_a^t y(u)\psi(u, \theta)^2\alpha(u, \theta)\,\mathrm{d}u$ and $p(t) = \tau^2(t)/\tau^2(b)$. But this shows that $D(.)$ is distributed as a scaled and time-transformed Brownian bridge, $\tau(b)W^0(p(.))$. Consequently $\max_{a \le t \le b} |D_n(t)|/\widehat{\tau}(b)$ is asymptotically distributed as $\|W^0\| = \max_{0 \le s \le 1} |W^0(s)|$, where $\widehat{\tau}^2(b) = \int_a^b n^{-1}Y(u)\psi(u, \widehat{\theta}_{[a,b]})^2 \alpha(u, \widehat{\theta}_{[a,b]})\,\mathrm{d}u$ estimates $\tau^2(b)$. A natural procedure is therefore to stretch the $[a, b] = [s - \frac{1}{2}h, s + \frac{1}{2}h]$ interval until

$$\Big\{\int_a^b Y(u)\psi(u, \widehat{\theta}_{[a,b]})^2\alpha(u, \widehat{\theta}_{[a,b]})\,\mathrm{d}u\Big\}^{-1/2} \max_{a \le t \le b} \Big|\int_a^t \psi(u, \widehat{\theta}_{[a,b]})\{\mathrm{d}N(u) - Y(u)\alpha(u, \widehat{\theta}_{[a,b]})\,\mathrm{d}u\}\Big| \ge 1.225, \tag{6.3}$$

say, 1.225 being the upper 10% point of the distribution of $\|W^0\|$. One might opt for 1.359 instead, the upper 5% point. Observe that the maximum value must be attained at one of the points x_i of x_i-, with $a \le x_i \le b$, so the continuous maximum is really only a finite maximum, and is perfectly feasible to compute efficiently, for given $s \pm \frac{1}{2}h$ window.

When choosing window sizes for the (3.12) estimator, for example, which uses local constants, the windows should be stretched until

$$N[a, b]^{-1/2} \max_{a \le t \le b} \Big|N[a, t] - \int_a^t Y(u)\widehat{\theta}_{[a,b]}\,\mathrm{d}u\Big| \ge 1.225, \tag{6.4}$$

where in this case $\widehat{\theta}_{[a,b]} = N[a, b]/\int_a^b Y(u)\,\mathrm{d}u$.

6.4. LOCAL MODEL ADEQUACY FOR MULTI-PARAMETER HAZARD RATES. Next turn attention to dynamic likelihood smoothing of a multi-parametric class of hazards, with $p \ge 2$ parameters. Let this time

$$D_n(t) = n^{-1/2}\Big\{N[a, t] - \int_a^t Y(u)\alpha(u, \widehat{\theta}_{[a,b]})\,\mathrm{d}u\Big\} \quad \text{for } t \in [a, b],$$

with a view towards using the maximal absolute value as a test for model adequacy. Here $\widehat{\theta}_{[a,b]}$ is the local maximum likelihood estimator using $[a, b]$-information, i.e. solving the p equations $\int_a^b \psi(u, \theta)\{\mathrm{d}N(u) - Y(u)\alpha(u, \theta)\,\mathrm{d}u\} = 0$. Techniques of Hjort (1990) can be used to demonstrate process convergence of $D_n(.)$ towards

$$D(t) = V[a, t] - \Big(\int_a^t y(u)\psi(u, \theta)\alpha(u, \theta)\,\mathrm{d}u\Big)'\Sigma^{-1} \int_a^b \psi(u, \theta)\,\mathrm{d}V(s),$$

where $V(.)$ is a Gaußian martingale with noise level $\mathrm{Var}\,\mathrm{d}V(u) = y(u)\alpha(u, \theta)\,\mathrm{d}u$, and where $\Sigma = \int_a^b y(u)\psi(u, \theta)\psi(u, \theta)'\alpha(u, \theta)\,\mathrm{d}u$. The $D_0(t) = V[a, t] = \int_a^t \mathrm{d}V(u)$ process is quite simple, it has independent increments and hence is a scale- and time-transformed Brownian motion process. The point is now that if one considers the $D_0(.)$ process conditioned on the p events $\int_a^b \psi(u,\theta)\,dV(u) = 0$, then Gaußianeity and covariance calculations can be furnished to demonstrate that this is exactly distributed as the $D(.)$ process; cf. Remark 7F in Hjort (1990). This makes it possible to bound the distribution of $\|D\| = \max_{a\le t\le b}|D(t)|$, even though the exact distribution might be too difficult to obtain.

We now specialise to a class of hazards of the form $\alpha(u,\theta) = \theta\gamma(u,\beta)$, cf. (4.6), in which case the $\psi(.)$ function has first component $1/\theta$ and second component $\phi(u,\beta)$, say. In this case the limit process $D(.)$ is distributed as $D_0(.)$, tied down first with $D_0(b) = 0$ and then with $\int_a^b \phi(u,\beta)\,dV(u) = 0$. Letting $D^*(.)$ be the result of tying down $D_0(.)$ with only the first requirement, covariance calculations show that $D^*(.) =_d \tau(b)W^0(p(.))$, this time with $\tau^2(t) = \int_a^t y(u)\theta\gamma(u,\beta)\,du$ and $p(t) = \tau^2(t)/\tau^2(b)$. So the distribution of $D(.)$ is that of tying down $D^*(.)$ further, and it can be seen that the distribution of $\|D\|$ is stochastically smaller than the distribution of $\|D^*\|$, just as the distribution of a maximal absolute Brownian bridge is stochastically smaller than the distribution of a maximal absolute Brownian motion. Here $\tau^2(b)$ is estimated consistently with $\int_a^b n^{-1}Y(u)\widehat\theta_{[a,b]}\gamma(u,\widehat\beta_{[a,b]})\,du$, and we also have $\widehat\theta_{[a,b]} = N[a,b]/\int_a^b Y(u)\gamma(u,\widehat\beta_{[a,b]})\,du$. The end result is to use

$$N[a,b]^{-1/2}\max_{a\le t\le b}\left|N[a,t] - \int_a^t Y(u)\widehat\theta_{[a,b]}\gamma(u,\widehat\beta_{[a,b]})\,du\right| \ge 1.225 \tag{6.5}$$

as a conservative 10% level test criterion for rejecting $\theta\gamma(u,\beta)$ as a model for the hazard on $[a,b]$.

It is worth noting that the difference between the distributions of $\|D\|$ and $\|D^*\|$ is small when the $[a,b]$ interval is not large, provided the model being tested has the local reparametrisation form (4.6). This can be shown after expanding the Σ^{-1} matrix here in a way similar to that for J_W^{-1} in Sections 4.1 and 4.2. Thus 1.225 above is meant to be a conservative value but actually also an approximation to the real 0.90 point of the null distribution. Of course this approximation cannot be expected to be overly precise, and some experimentation with the 1.225 rejection limit would be needed. On the computational side we point out that the maximum again must be attained for one of $t = x_i$ or x_i- with $a \le x_i \le b$. Furthermore,

$$\int_a^{x_i} Y(u)\widehat\theta_{[a,b]}\gamma(u,\widehat\beta_{[a,b]})\,du = \sum_{j:x_j\ge a}\widehat\theta_{[a,b]}\{G(x_i\wedge x_j,\widehat\beta_{[a,b]}) - G(a,\widehat\beta_{[a,b]})\}, \tag{6.6}$$

where $G(t,\beta) = \int_0^t \gamma(u,\beta)\,du$.

6.5. Other tests for model adequacy on intervals. There are naturally other possible goodness of fit tests for intervals, see Hjort (1990) for other $D_n(.)$ type functions and for classes of chi squared type tests and Hjort and Lumley (1993) for normalised local hazard plots. Chi squared methods would be awkward to implement in a general way here, since the $[a,b]$ intervals would often be short. We record a couple of potentially useful variations on the $D_n(.)$ theme, however, with a view towards quick calculations and decisions, since tests are to be carried out on slowly expanding $[s-\frac12 h, s+\frac12 h]$ intervals for each s.

(6.3)–(6.5) arose as maxima of the $D_n(.)$ process, and utilised convergence to suitably scaled and time-transformed Brownian bridges, as with Kolmogorov–Smirnov type tests. Martingale techniques for the counting process N can be used to show

$$\begin{aligned}\int_a^b |D_n(t)|^q\,dN(t)/n &\to_d \int_a^b |D(t)|^q y(t)\theta\gamma(t,\beta)\,dt\\ &\le_d \int_a^b |\tau(b)W^0(p(t))|^q\tau^2(b)\,dp(t) = \tau(b)^{2+q}\int_0^1 |W^0(s)|^q\,ds,\end{aligned}$$

where '$\le_d$' means 'stochastically smaller than'. For $q=2$ we have a Cramér–von Mises type test, with rejection criterion

$$\begin{aligned}\widehat{\tau}(b)^{-4}\int_a^b D_n(t)^2\,\mathrm{d}N(t)/n &= \frac{n^{-1}\sum_{a\le x_i\le b} D_n(x_i)^2\delta_i}{(n^{-1}N[a,b])^2} \\ &= \frac{\sum_{a\le x_i\le b}\{N[a,x_i]-\int_a^{x_i} Y(u)\widehat{\theta}_{[a,b]}\gamma(u,\widehat{\beta}_{[a,b]})\,\mathrm{d}u\}^2\delta_i}{N[a,b]^2}\ge 0.347\end{aligned}\tag{6.7}$$

on the 10% significance level. With wished for 5% level we would use 0.461 instead, the 0.95 quantile of the $\int_0^1 W^0(s)^2\,\mathrm{d}s$ distribution. The second variation is for $q=1$, where we use

$$\begin{aligned}\widehat{\tau}(b)^{-3}\int_a^b |D_n(t)|\,\mathrm{d}N(t)/n &= \frac{n^{-1}\sum_{a\le x_i\le b} |D_n(x_i)|\delta_i}{(n^{-1}N[a,b])^{3/2}} \\ &= \frac{\sum_{a\le x_i\le b}\left|N[a,x_i]-\int_a^{x_i} Y(u)\widehat{\theta}_{[a,b]}\gamma(u,\widehat{\beta}_{[a,b]})\,\mathrm{d}u\right|}{N[a,b]^{3/2}}\ge 0.499\end{aligned}\tag{6.8}$$

for intended 10% significance level, and 0.582 for intended 5% significance level. 0.499 and 0.582 are upper quantiles of the $\int_0^1 |W^0(s)|\,\mathrm{d}s$ distribution. There are simpler one-parameter analogues to (6.7) and (6.8), essentially as in these formulae but with $\gamma(u,\beta)=1$. Note that (6.6) can be used when computing any of these test statistics.

Some experimentation with these $\widehat{h}=\widehat{h}(s)$ selectors is necessary. One should avoid using too small windows since this would lead to too irregular local estimates. We should therefore only search for acceptable windows $s\pm\frac12 h$ with h at least as large as some suitably determined $h_0(s)$. One possibility is to demand at least k observed x_i's in the window, say with $k=10$. Hence the (6.3)–(6.5) and (6.7)–(6.8) stopping criteria are to be used with such a modification. Secondly the realised $\widehat{h}(s)$ could be somewhat irregular as a function of s. A natural modification is to smooth this curve first, before finally computing the local likelihood estimate $\alpha(s,\widehat{\theta}_{[s-\widehat{h}(s)/2,s+\widehat{h}(s)/2]})$.

7. Supplementing remarks.

7A. Starting the estimator. We have defined $\widehat{\alpha}(s)=\alpha(s,\widehat{\theta}(s))$ with parameter estimate obtained from $s\pm\frac12 h$ data, which also means that a separate definition is required for $s\le\frac12 h$. One natural strategy is to use the model adequacy on intervals methods of Sections 6.3–6.5 to find the smallest b for which the model is rejected on $[0,b]$, and then use $\widehat{\alpha}(s)=\alpha(s,\widehat{\theta}_{[0,b_0]})$ for $s\in[0,b_0]$, with a somewhat smaller b_0 than b. Another possibility is to use $\alpha(s,\widehat{\theta}(\frac12 h))$ on $[0,\frac12 h]$.

7B. Post-smoothing of parameter estimates. The basic estimator is $\widehat{\alpha}(s)=\alpha(s,\widehat{\theta}(s))$ where $\widehat{\theta}(s)$ uses only $s\pm\frac12 h$ information. It is useful in practical applications to display not only the final $\widehat{\alpha}(s)$ but also the parameter estimate function or functions $\widehat{\theta}(s)$. Sometimes this function has discontinuities, cf. (3.12) and the requirements on K noted there to give smoothness. A general alternative is to post-smooth the parameter estimates, before plugging in to give $\widehat{\alpha}(s)$. Comments in 5.1, for example, suggest using post-smoothing of $\widehat{\beta}(s)$ in (4.3) and (4.4).

7C. Density estimation with dynamic likelihood. When $\alpha(.)$ is estimated one can of course also estimate other quantities depending on $\alpha(.)$. The local likelihood methods of Sections 3 and 4 therefore apply to nonparametric or semiparametric density estimation as well, via the $f(t)=\alpha(t)\exp\{-A(t)\}$ connection. Methods given there can be used to obtain a locally estimated normal density of the type $\widehat{f}(t) = N\{\widehat{\mu}(t), \widehat{\sigma}(t)^2\}(t)$, for example. There are at least two general immediate possibilities, namely

$$\widehat{f}_1(t) = \alpha(t, \widehat{\theta}(t)) \exp\{-A(t, \widehat{\theta}(t)\} \quad \text{and}$$
$$\widehat{f}_2(t) = \Big[\prod_{[0,t)} \{1 - \alpha(s, \widehat{\theta}(s))\, \mathrm{d}s\}\Big] \alpha(t, \widehat{\theta}(t)) = \exp\Big\{-\int_0^t \alpha(s, \widehat{\theta}(s))\, \mathrm{d}s\Big\} \alpha(t, \widehat{\theta}(t)).$$

The simplest case would again be that of a locally constant hazard, for which

$$\widehat{f}_1(t) = \widehat{\theta}(t) \exp\{-\widehat{\theta}(t)\, t\} \quad \text{and} \quad \widehat{f}_2(t) = \exp\Big\{-\int_0^t \widehat{\theta}(s)\, \mathrm{d}s\Big\}\, \widehat{\theta}(t).$$

These are somewhat cumbersome density estimators. There are better schemes more directly geared towards the density estimation problem, but still with the same local likelihood characteristics, see Hjort and Jones (1993).

7D. REGRESSION MODELS. Methods of this paper can be made to work in situations with covariate information. Consider the Cox regression model where individual i has hazard rate of the form

$$\alpha_i(s) = \alpha_0(s) \exp(\beta' z_i) \quad \text{for } s \in [0, T] \text{ and } i = 1, \ldots, n.$$

The $\alpha_0(.)$ function is the hazard rate for individuals with covariate vector $z = 0$, and is left unspecified. This baseline hazard function can now be estimated using dynamic likelihood. If we fit a local constant on window $W = s \pm \frac{1}{2}h$ the recipe is to maximise the kernel smoothed log-likelihood

$$\sum_{i=1}^n \int_W K(h^{-1}(t-s))\{(\log\theta + \beta' z_i)\, \mathrm{d}N_i(t) - Y_i(t)\theta \exp(\beta' z_i)\, \mathrm{d}t\},$$

where $\mathrm{d}N_i(t) = I\{x_i \in [t, t + \mathrm{d}t], \delta_i = 1\}$ and $Y_i(t) = I\{x_i \geq t\}$ are the 0–1 counting process and at risk process for individual i. This gives

$$\widehat{\alpha}_0(s) = \frac{\sum_{i=1}^n \int_W K(h^{-1}(t-s))\, \mathrm{d}N_i(t)}{\sum_{i=1}^n \int_W K(h^{-1}(t-s)) Y_i(t) \exp(\widehat{\beta}' z_i)\, \mathrm{d}t}.$$

Here $\widehat{\beta}$ could be evaluated only locally, but if one trusts the Cox model then β remains constant over the $[0, T]$ range, and we should accordingly use the same $\widehat{\beta}$ regardless of s. But this is the same as smoothing the traditional Breslow estimator. One can similarly construct a nonparametric $\alpha_0(.)$ estimator by fitting a running Weibull $\theta\gamma s^{\gamma-1}$, for example. The result is of the form

$$\widehat{\alpha}_0(s) = \frac{\sum_{i=1}^n \int_W K(h^{-1}(t-s))\, \mathrm{d}N_i(t)}{\sum_{i=1}^n \int_W K(h^{-1}(t-s)) Y_i(t) \widehat{\gamma}(s) t^{\widehat{\gamma}(s)-1} \exp(\widehat{\beta}' z_i)\, \mathrm{d}t}.$$

Dynamic likelihood methods can also be developed in Aalen's linear hazard rate regression model, by local parametric modelling of the hazard factor functions. See Hjort (1993a).

7E. MODERATELY INCORRECT PARAMETRIC MODELS. A parametric model does not have to be fully perfect in order for the methods based on it to be better than more conservative ones. In Hjort (1993b) a 'tolerance distance' is calculated from a moderately incorrect model to a wider and correct one; inside the tolerance radius estimators based on the incorrect model are better than those based on the correct model. For an example, suppose the true model is the gamma one, with hazard function inherited from the density $f(s,\theta,\gamma) = \{\theta^\gamma/\Gamma(\gamma)\}\, s^{\gamma-1}\exp(-\theta s)$. Then estimators based on the incorrect assumption of a constant rate (which corresponds to $\gamma = 1$) are better than the two-parameter methods if $|\gamma - 1| \le 1.245/\sqrt{n}$ (assuming no censoring). This can be seen as yet another argument for not giving up simple parametric methods, even though the underlying models might be wrong.

7F. Counting process models. Methods and results of this paper can be generalised in various directions. They could be developed for Aalen's general multiplicative intensity model for counting processes, and hence be used to estimate hazard transition rates in time-inhomogeneous Markov chains, for example. There will then be a more complicated expression for the M_W matrix of (2.6), but otherwise there will be few complications. In another direction our results could be extended to the full halfline $[0,\infty)$ with appropriate extra assumptions on the censoring mechanism.

7G. More theory. In our presentation we have concentrated on the perhaps most immediate aspects of the dynamic likelihood estimation method. There are further natural questions to ask and further natural results to prove. (i) One can prove uniform consistency of the (3.12) estimator without too much work, for example. One can more generally establish $\max_{a\le s\le b}|\alpha(s,\widehat\theta(s)) - \alpha(s)| \to_p 0$ under natural conditions. (ii) And the approximate size and distribution of this maximal deviation quantity are also of interest. (iii) It is not difficult to establish that $(nh)^{1/2}\{\widehat\alpha(s) - \alpha(s) - \frac{1}{2}\beta_K\alpha''(s)\}$ has a limiting zero-mean normal distribution, when $h \to 0$ and $nh \to \infty$. This can also be used to construct point-wise approximate confidence band for the $\alpha(.)$ function, for example incorporating a bias correction $-\frac{1}{2}\beta_K\widehat\alpha''(s)$. (iv) One should work out a reliable cross validation method for minimising a nearly unbiased estimate of $\int_0^T w(s)\{\widehat\alpha(s) - \alpha(s)\}^2\,ds$, say, as a function of the window width h, or as a function of c in $h = cY_n(s)^{-1/5}$. The crux is to estimate $\int_0^T w(s)\widehat\alpha(s)\alpha(s)\,ds$. (v) Theory can also be worked out for estimation of derivatives of the hazard function, as touched on in 6.1. Taking the derivative of (3.12) to define $\widehat\alpha'(s)$, with a smooth kernel function K, one can show that the bias is proportional to a $h^2 b_1(s)$ and that the variance is proportional to $n^{-1}h^{-3}\alpha(s)/y(s)$, but with a $b_1(s)$ function different from that of the derivative of the Ramlau-Hansen–Yandell estimator.

7H. Questions. A simulation study comparing the various estimators would be welcome. Some of the questions to answer include: How much better are the new estimators than the purely nonparametric ones when the true hazard is in the vicinity of the parametric model used? How much do they lose to the parametric ones on the latter's home turf? Are there significant advantages to using multi-parameter models for the dynamic likelihood methods of Sections 3 and 4? What are the most useful ways of choosing window width $h = h_n(s)$?

References

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

Borgan, Ø. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scand. J. Statist.* **11**, 1–16. Corrigendum, *ibid.* p. 275.

Grégoire, C. (1993). Least squares cross validation for counting process intensities. *Scand. J. Statist.*, to appear.

Hjort, N.L. (1985). Contribution to the discussion of Andersen and Borgan's 'Counting process models for life history data: a review'. *Scand. J. Statist.* **12**, 141–150.

Hjort, N.L. (1990). Goodness of fit tests in models for life history data based on cumulative hazard rates. *Ann. Statist.* **18**, 1221–1258.

Hjort, N.L. (1991). Semiparametric estimation of parametric hazard rates. In *Survival Analysis: State of the Art*, Kluwer, Dordrecht, pp. 211–236. Proceedings of the *NATO Advanced Study Workshop on Survival Analysis and Related Topics*, Columbus, Ohio, eds. P.S. Goel and J.P. Klein.

Hjort, N.L. (1992). On inference in parametric survival data models. *Int. Statist. Review* **60**, 355–387.

Hjort, N.L. (1993a). Efficiency of three estimators in Aalen's linear hazard rate regression model. Submitted for publication.

Hjort, N.L. (1993b). Estimation in moderately misspecified models. Submitted for publication.

Hjort, N.L. and Jones, M.C. (1993). Locally parametric nonparametric density estimation. Manuscript in progress.

Hjort, N.L. and Lumley, T. (1993). Normalised local hazard plots. Submitted for publication.

Nielsen, J.P. (1990). Kernel estimation of densities and hazards: a counting process approach. PhD. dissertation, Department of Statistics, University of Berkeley, California.

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

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

Yandell, B.S. (1983). Nonparametric inference for rates and densities with censored serial data. *Ann. Statist.* **11**, 1119–1135.