

# MAXIMUM LOCAL PARTIAL LIKELIHOOD ESTIMATORS FOR THE COUNTING PROCESS INTENSITY FUNCTION AND ITS DERIVATIVES

Feng Chen

*University of New South Wales*

*Abstract:* We propose estimators for the counting process intensity function and its derivatives by maximizing the local partial likelihood. We prove the consistency and asymptotic normality of the proposed estimators. In addition to the computational ease, a nice feature of the proposed estimators is the automatic boundary bias correction property. We also discuss the choice of the tuning parameters in the definition of the estimators. An effective and easy-to-calculate data-driven bandwidth selector is proposed. A small simulation experiment is carried out to assess the performance of the proposed bandwidth selector and the estimators.

*Key words and phrases:* Asymptotic normality, automatic boundary correction, composite likelihood estimator, consistency, counting process, intensity function, local likelihood, local polynomial methods, martingale, maximum likelihood estimator, multiplicative intensity model, partial likelihood, point process,  $Z$ -estimator.

## 1. Introduction

The counting process is a useful statistical model that is extensively applied in the analysis of data arising from such fields as medicine and public health, biology, finance, insurance, and social sciences. The evolution of a counting process over time is (stochastically) governed by its intensity process, and by its intensity function in the case of multiplicative intensity process models. Therefore, to understand the behavior of a multiplicative intensity counting process, the estimation of the intensity function is an important question. There is a sizable literature devoted to this question. Ramlau-Hansen (1983) applied the kernel smoothing method to estimate the intensity function and established the consistency and local asymptotic normality of the proposed estimator. Ramlau-Hansen also proposed to estimate the derivatives of the intensity function by differentiating the estimator of the intensity function. Karr (1987) used the sieve method and established the strong consistency and the asymptotic normality of the resultant estimator for the intensity function. Antoniadis (1989) proposed a penalized maximum likelihood estimator for the intensity function and proved its consistency and asymptotic normality. Patil and Wood (2004) used the wavelet

method to estimate the intensity function and calculated the mean integrated squared error of the estimator. Chen et al. (2008) proposed to estimate the intensity function and its derivatives using a local polynomial estimator based on martingale estimating equations, and proved the consistency and asymptotic normality of the estimators.

The roughness penalty method and the sieve method are computationally expensive because of the potentially high dimensional optimization involved. Also, formal methods for derivative estimation based on these methods or the wavelet thresholding method seem lacking. The kernel smoothing method is computationally very appealing, but requires extra effort in modifying the estimators for boundary points to reduce the boundary or edge effects. The biased martingale estimating equation method of Chen et al. (2008) provides estimators for the intensity function and its derivatives in a unified manner and enjoys both the computational ease and the automatic boundary correction property of local polynomial type methods (Chen, Yip, and Lam (2009)) but, from a theoretic point of view, the construction of the estimating equations is somewhat arbitrary. In this paper we propose an estimation procedure for the intensity function and its derivatives based on the idea of *local partial likelihood*, which falls under the umbrella of the more general concept of *composite likelihood* (Lindsay (1988)). The resultant estimators are fast to compute, and are consistent and asymptotically normally distributed under mild regularity conditions. Moreover, similar to the biased estimating equation estimators, they also enjoy the automatic boundary correction property. The practically important issue of bandwidth selection is also discussed.

In Section 2 we introduce a multiplicative intensity counting process and give a heuristic derivation of the partial likelihood that is used later. We present the maximum local partial likelihood estimator for the intensity function and its derivatives in Section 3. The properties of the estimators are considered in Section 4. A data-driven bandwidth selector is proposed in Section 5. Section 6 reports the results of a small scale numerical experiment in verifying the properties of the proposed estimators and assessing the finite sample numerical performance of the bandwidth selector as well as the estimators. Finally, some concluding remarks and discussions are given in Section 7. A lemma used in proving the consistency of the estimators is given in the Appendix.

## 2. Multiplicative Intensity Counting Process

The multiplicative intensity counting process is a counting process  $N(t)$  that has an intensity process with respect to a filtration  $\{\mathcal{F}_t : t \in [0, 1]\}$  of the multiplicative form  $Y(t)\alpha(t)$ ,  $t \in [0, 1]$ . Here  $Y$  is a nonnegative predictable process called the *exposure process*, and  $\alpha$  a positive deterministic function, called the *intensity function*. Note the compensated counting process  $M(t) =$

$N(t) - \int_0^t Y(s)\alpha(s)ds$  is a local square integrable martingale. Suppose we can observe the processes  $N(t)$  and  $Y(t)$  over  $[0, 1]$  and are interested in estimating  $\alpha(t)$  and its derivatives.

Using the chain rule for conditional probabilities, we can informally write the likelihood of the data as

$$\begin{aligned}
 & \Pr(N(t), Y(t); 0 \leq t \leq 1) \\
 &= \Pr(N_0, Y_0) \times \Pr(dN(t), dY(t); 0 < t \leq 1 | \mathcal{F}_0) \\
 &= \Pr(N_0, Y_0) \times \prod_{t \in (0,1]} \Pr(dN(t) | \mathcal{F}_{t-}) \Pr(dY(t) | \mathcal{F}_{t-}) \\
 &= \left\{ \prod_{t \in (0,1]} (Y(t)\alpha(t)dt)^{dN(t)} (1 - Y(t)\alpha(t)dt)^{1-dN(t)} \right\} \\
 & \quad \times \left\{ \Pr(N_0, Y_0) \prod_{t \in (0,1]} \Pr(dY(t) | \mathcal{F}_{t-}) \right\}. \tag{2.1}
 \end{aligned}$$

To specify the full likelihood we have to specify the distributions of  $dY(t)$ . To avoid this usually awkward task, we base our inference about  $\alpha$  on the *partial* likelihood obtained by discarding the terms in the second pair of braces in (2.1), and neglecting a term  $dt^{N(1)}$  from the first pair of braces. Thus,

$$\begin{aligned}
 L(\alpha) &= \prod_{t \in (0,1]} (Y(t)\alpha(t))^{dN(t)} (1 - Y(t)\alpha(t)dt)^{1-dN(t)} \\
 &= \prod_{i=1}^{N(1)} (Y(T_i)\alpha(T_i))^{\Delta N(T_i)} \exp\left\{-\int_0^1 Y(s)\alpha(s)ds\right\}, \tag{2.2}
 \end{aligned}$$

where the  $T_i$  are the jump times of the counting process  $N$ . With nonparametric conditions on  $\alpha$  such as differentiability over the interval  $[0, 1]$ ,  $L(\alpha)$  can be made arbitrarily large by suitably choosing  $\alpha$ . Therefore, the fully nonparametric maximum likelihood estimator for  $\alpha$  does not exist. To overcome this difficulty, the penalized maximum likelihood estimator (MLE) and the sieve MLE for  $\alpha$  have been proposed by Antoniadis (1989) and Karr (1987), respectively. In this paper, we consider the computationally appealing alternative of maximum local partial likelihood estimator.

### 3. Definition of the Local Partial Likelihood and the Estimator

Consider the estimation of  $\alpha(t)$  for any fixed  $t \in [0, 1]$ . We define a reasonable local log-partial likelihood function for the purpose of estimating  $\alpha(t)$ . First we

note the *global* log-partial likelihood function can be written, again informally, as

$$\log L = \int_0^1 [\{\log(Y(s)\alpha(s))\}dN(s) + \{1 - dN(s)\} \log(1 - Y(s)\alpha(s)ds)]. \quad (3.1)$$

Introducing a weighting process  $W_t(s)$  to modulate the likelihood contributions in the infinitesimal intervals  $(s, s + ds]$ , and replacing  $\alpha(s)$  by a local version  $\tilde{\alpha}_t(s)$ , we arrive at the localized log-partial likelihood

$$\begin{aligned} & \int_0^1 W_t(s) [\{\log(Y(s)\tilde{\alpha}_t(s))\}dN(s) + \{1 - dN(s)\} \log(1 - Y(s)\tilde{\alpha}_t(s)ds)] \\ &= \int_0^1 \{\log(Y(s)\tilde{\alpha}_t(s))\}W_t(s)dN(s) - \int_0^1 Y(s)\tilde{\alpha}_t(s)W_t(s)ds, \end{aligned} \quad (3.2)$$

where the equality follows from  $\log(1 - Y(s)\tilde{\alpha}_t(s)ds) = -Y(s)\tilde{\alpha}_t(s)ds$  and the fact that the set  $\{s : dN(s) \neq 0\}$  has Lebesgue measure 0.

The obvious choice of the local version of  $\alpha$  at  $t$  seems to be the truncated Taylor series expansion

$$\tilde{\alpha}_t(s) = \alpha(t) + \alpha'(t)(s - t) + \cdots + \frac{\alpha^{(p)}(t)}{p!}(s - t)^p = \mathbf{g}_p(s - t)^\top \boldsymbol{\theta}_t, \quad (3.3)$$

where  $\mathbf{g}_p(x) = (1, x, \dots, x^p/p!)^\top$  and  $\boldsymbol{\theta}_t = (\alpha(t), \alpha'(t), \dots, \alpha^{(p)}(t))^\top$ .

The weighting process  $W_t(s)$  is expected to give less weight to the likelihood contribution of  $dN(s)$  if  $s$  is further away from  $t$ , or if  $dN(s)$  has larger (conditional) variance. As the conditional variance of  $dN(s)$  given  $\mathcal{F}_{s-}$  is  $Y(s)\alpha(s)ds$  and is proportional to  $Y(s)$ , we let  $W_t(s)$  be inversely proportional to  $Y(s)$ . The above considerations motivate the choice of weighting process

$$W_t(s) = \frac{1}{b} K\left(\frac{s-t}{b}\right) \frac{J(s)}{Y(s)} = K_b(s-t) \frac{J(s)}{Y(s)}, \quad (3.4)$$

where  $b$  is a positive tuning parameter, called the *bandwidth* or *window size*, that controls the size of the local neighborhood,  $K(x)$  is a nonnegative function called the *kernel function*, that is nonnegative and integrates to unity, and  $J(s) = 1\{Y(s) > 0\}$  is the indicator of  $\{Y(s) > 0\}$  that, together with the convention that  $0/0 = 0$ , prevents the weighting process from giving infinite weights.

Substituting (3.3) and (3.4) into (3.2) gives a form of the local log-likelihood that can be used to estimate  $\alpha$  and its derivatives up to order  $p$  at  $t$ :

$$\begin{aligned} \ell(\boldsymbol{\theta}_t) &= \int_0^1 \{\log(Y(s)\mathbf{g}_p(s-t)^\top \boldsymbol{\theta}_t)\} K_b(s-t) \frac{J(s)}{Y(s)} dN(s) \\ &\quad - \int_0^1 \mathbf{g}_p(s-t)^\top \boldsymbol{\theta}_t J(s) K_b(s-t) ds. \end{aligned} \quad (3.5)$$

The MLPLE for  $\boldsymbol{\theta}_t$  comes as a maximizer of  $\ell(\boldsymbol{\theta}_t)$ . To maximize  $\ell(\boldsymbol{\theta}_t)$ , we need the local score and (observed) information matrix for  $\boldsymbol{\theta}_t$ . These follow from differentiating  $\ell(\boldsymbol{\theta}_t)$  with respect to  $\boldsymbol{\theta}_t$ :

$$\begin{aligned} \mathbf{s}(\boldsymbol{\theta}_t) &= \frac{\partial \ell(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t} \\ &= \int_0^1 \frac{\mathbf{g}_p(s-t)K_b(s-t)J(s)}{\mathbf{g}_p(s-t)^\top \boldsymbol{\theta}_t Y(s)} dN(s) - \int_0^1 \mathbf{g}_p(s-t)J(s)K_b(s-t) ds, \end{aligned} \quad (3.6)$$

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta}_t) &= -\frac{\partial^2 \ell(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t \partial \boldsymbol{\theta}_t^\top} = \int_0^1 \frac{\mathbf{g}_p(s-t)\mathbf{g}_p(s-t)^\top J(s)K_b(s-t)}{\{\mathbf{g}_p(s-t)^\top \boldsymbol{\theta}_t\}^2 Y(s)} dN(s) \\ &= \sum_{i=1}^{N(1)} \frac{\mathbf{g}_p(T_i-t)\mathbf{g}_p(T_i-t)^\top J(T_i)K_b(T_i-t)}{\{\mathbf{g}_p(T_i-t)^\top \boldsymbol{\theta}_t\}^2 Y(T_i)} \Delta N(T_i). \end{aligned} \quad (3.7)$$

The MLPLE  $\hat{\boldsymbol{\theta}}_t$  of  $\boldsymbol{\theta}_t$  should solve the local score equation  $\ell(\boldsymbol{\theta}_t) = \mathbf{0}$ . Except for the case  $p = 0$ , where the MLPLE  $\hat{\boldsymbol{\theta}}_t$  is one-dimensional and is given by

$$\hat{\alpha}^{(0)}(t) = \frac{\int_0^1 K_b(s-t)J(s)/Y(s) dN(s)}{\int_0^1 K_b(s-t)J(s) ds}, \quad (3.8)$$

a closed-form solution for the score equation is generally not anticipated and we have to resort to numerical procedures. As we have the information matrix at hand, the obvious choice of the numerical procedure is the Newton-Raphson iteration

$$\boldsymbol{\theta}_t^{(new)} = \boldsymbol{\theta}_t^{(old)} + \mathbf{I}(\boldsymbol{\theta}_t^{(old)})^{-1} \mathbf{s}(\boldsymbol{\theta}_t^{(old)}). \quad (3.9)$$

The initial value to start the iteration can be chosen as  $\boldsymbol{\theta}_t^{(0)} = \hat{\alpha}_t^{(0)} \mathbf{e}_{0,p}^\top$ , where  $\hat{\alpha}_t^{(0)}$  is given by (3.8) and  $\mathbf{e}_{i,p}$  is the  $(p+1)$ -dimensional unit vector with the  $(i+1)$ st component being 1.

#### 4. Properties of the Estimator

As we are interested in asymptotic properties, we consider a sequence of multiplicative models, indexed by  $n = 1, 2, \dots$ . Throughout, convergence is as  $n$  tends to infinity, unless indicated otherwise. For each  $n$ , with respect to the filtration  $\mathcal{F}^{(n)} = \{\mathcal{F}_t^{(n)} : t \in [0, 1]\}$ , the counting process  $N^{(n)}$  has an intensity process  $Y^{(n)}\alpha(t)$ , with  $Y^{(n)}$  being predictable with respect to the filtration  $\mathcal{F}^{(n)}$ . Note here the intensity function  $\alpha$  is the same in all models. In the definition of the MLPLE for  $\theta_t$ , we let the kernel function  $K$  and the order  $p$  of the local polynomials be fixed, but allow the bandwidth parameter  $b$  to vary with

$n$ . Then, under regularity conditions on the model and on the choices of the bandwidth and of the kernel function, we can prove that the MLPLE is consistent and asymptotically normally distributed. To avoid repetitious statements of regularity conditions, and for ease of reference, we list all the conditions below.

- C1. (Positive and differentiable intensity). The true intensity is positive and  $p$  times continuously differentiable at  $t$ . When  $t = 0$  or  $1$ , derivatives are interpreted as right or left derivatives accordingly.
- C2. (Explosive exposure process). The exposure process  $Y^{(n)} \rightarrow \infty$  in probability, uniformly in a neighborhood of  $t$ .
- C3. (Slowly shrinking bandwidth). The bandwidth  $b_n \searrow 0$ , but  $b_n^{2p+1}Y^{(n)} \rightarrow \infty$  in probability, uniformly in a neighborhood of  $t$ .
- C4. (Positive definite kernel moment matrix). The kernel function  $K$  is bounded, is supported by a compact interval, say  $[-1, 1]$ , and is such that the matrix

$$\mathbf{A}_t = \int_{-\min(1,t/0)}^{\min(1,(1-t)/0)} \mathbf{g}_p^*(x)^{\otimes 2} K(x) dx \quad (4.1)$$

is positive definite, where we follow the convention  $0/0 = 0$  as before, and use  $\mathbf{g}_p^*$  to denote the vector-valued function  $\mathbf{g}_p^*(x) = (1, x, \dots, x^p)^\top$  and  $\mathbf{x}^{\otimes 2} = \mathbf{x}\mathbf{x}^\top$  to denote the outer product of a vector  $\mathbf{x}$ .

- C5. (Explosive exposure process). There exists a sequence of positive constants  $a_n \nearrow \infty$ , and a deterministic function  $y$  that is positive and continuous at  $t$  such that  $Y^{(n)}/a_n^2$  converges in probability to  $y$ , uniformly in a neighborhood of  $t$ .
- C6. (Higher order differentiable intensity). The true intensity is  $(p+1)$  times continuously differentiable at  $t$ .

We use  $\xrightarrow{P}$  to indicate convergence in probability, and  $\xrightarrow{d}$  convergence in distribution. We take  $\|\mathbf{x}\| = (\sum_i x_i^2)^{1/2}$  for a vector  $\mathbf{x}$ ,  $\mathbf{D}_p$  to denote the  $(p+1) \times (p+1)$  diagonal matrix with diagonal elements  $D_{i+1,i+1} = b^i/i!$ ,  $i = 0, \dots, p$ , and  $\boldsymbol{\theta}_t^0$  to denote the true value of  $\boldsymbol{\theta}_t$ .

#### 4.1. Consistency and asymptotic normality

**Theorem 4.1 (Consistency of  $\hat{\boldsymbol{\theta}}_t$ ).** *Under C1–C4,  $\hat{\boldsymbol{\theta}}_t \xrightarrow{P} \boldsymbol{\theta}_t^0$ .*

By definition, our estimator is an  $M$ -estimator as well as a  $Z$ -estimator. Therefore it is natural to try to establish its consistency using the general theory about these types of estimators, such as Theorem 5.7 and Theorem 5.9 of van der Vaart (1998). However, these two theorems cannot be directly applied here,

due to the lack of a useful *fixed* limit of the criterion function. To overcome this difficult, we generalize Theorem 5.9 of van der Vaart (1998) to accommodate situations in which a useful fixed limit of the criterion function does not exist but a deterministic approximating sequence of the criterion function can be constructed. This generalization is stated as Lemma A.1 in the Appendix.

**Proof of Theorem 4.1.** In the sequel we suppress the subscript or superscript  $p$  and  $n$  from various notations for simplicity. Define  $\mathbf{s}_0$ , a vector-valued function of  $\boldsymbol{\theta}_t$ , by

$$\mathbf{s}_0(\boldsymbol{\theta}_t) = \int_0^1 \mathbf{g}(s-t) \frac{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t^0}{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t} K_b(s-t) ds - \int_0^1 \mathbf{g}(s-t) K_b(s-t) ds.$$

Note that  $\mathbf{s}_0$  depends on  $n$  through the bandwidth  $b$  and has a zero  $\boldsymbol{\theta}_t^0$ . A change of variable in the integrals involved shows that

$$\mathbf{D}^{-1} \mathbf{s}_0(\boldsymbol{\theta}_t) = \int_{-\min(1,t/0)}^{\min(1,(1-t)/0)} \mathbf{g}^*(x) \left\{ \frac{\mathbf{g}^*(x)^\top \mathbf{D} \boldsymbol{\theta}_t^0}{\mathbf{g}^*(x)^\top \mathbf{D} \boldsymbol{\theta}_t} - 1 \right\} K(x) dx$$

when  $n$  is large. The derivative matrix of  $\mathbf{D}^{-1} \mathbf{s}_0(\boldsymbol{\theta}_t)$  evaluated at  $\boldsymbol{\theta}_t^0$  is given by

$$-\mathbf{D} \int_{-\min(1,t/0)}^{\min(1,(1-t)/0)} \frac{\mathbf{g}^*(x)^{\otimes 2} K(x)}{\mathbf{g}^*(x)^\top \mathbf{D} \boldsymbol{\theta}_t^0} dx$$

which, by C1 and C4 is positive definite and has smallest eigenvalue larger than  $\lambda b^p/p!$  for some constant  $\lambda > 0$ , when  $n$  is large. Thus, a linearization of  $\mathbf{D}^{-1} \mathbf{s}_0$  shows that when  $n$  is large,  $\boldsymbol{\theta}_t^0$  is a suitably separated solution of  $\mathbf{s}_0(\boldsymbol{\theta}_t) = \mathbf{0}$  in the sense that

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_t^0\| \geq \epsilon \Rightarrow \|\mathbf{D}^{-1} \mathbf{s}_0(\boldsymbol{\theta}_t)\| \geq \lambda b^p \epsilon / (2p!), \quad \forall \epsilon > 0 \text{ small enough.} \quad (4.2)$$

If we can show that  $\sup_{\boldsymbol{\theta}_t} \|b^{-p} \mathbf{D}^{-1} \{\mathbf{s}(\boldsymbol{\theta}_t) - \mathbf{s}_0(\boldsymbol{\theta}_t)\}\| \xrightarrow{P} 0$ , then Lemma A.1 guarantees  $\hat{\boldsymbol{\theta}}_t \xrightarrow{P} \boldsymbol{\theta}_t^0$ .

To complete the proof, it remains to show that  $\sup_{\boldsymbol{\theta}_t} \|b^{-p} \mathbf{D}^{-1} \{\mathbf{s}(\boldsymbol{\theta}_t) - \mathbf{s}_0(\boldsymbol{\theta}_t)\}\| \xrightarrow{P} 0$ . With  $\alpha_0$  denoting the true intensity, and  $M(t) = N(t) - \int_0^t Y(s) \alpha_0(s) ds$  the compensated counting process, we can write  $\mathbf{s}(\boldsymbol{\theta}_t) - \mathbf{s}_0(\boldsymbol{\theta}_t)$  as the sum of the random functions

$$\begin{aligned} \boldsymbol{\Delta}_1(\boldsymbol{\theta}_t) &= \int_0^1 \mathbf{g}(s-t) \frac{J(s)}{Y(s)} \frac{1}{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t} K_b(s-t) dM(s), \\ \boldsymbol{\Delta}_2(\boldsymbol{\theta}_t) &= \int_0^1 \mathbf{g}(s-t) J(s) \frac{1}{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t} K_b(s-t) \{\alpha_0(s) - \mathbf{g}(s-t)^\top \boldsymbol{\theta}_t^0\} ds, \end{aligned}$$

$$\begin{aligned}\Delta_3(\boldsymbol{\theta}_t) &= \int_0^1 \mathbf{g}(s-t)\{J(s)-1\} \frac{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t^0}{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t} K_b(s-t) ds, \\ \Delta_4(\boldsymbol{\theta}_t) &= \int_0^1 \mathbf{g}(s-t)\{1-J(s)\} K_b(s-t) ds.\end{aligned}$$

We only need to show  $b^{-p} \sup_{\boldsymbol{\theta}_t} \|\mathbf{D}^{-1} \Delta_i(\boldsymbol{\theta}_t)\| \xrightarrow{P} 0$ , for  $i = 1, 2, 3, 4$ . By C2, with probability tending to 1,  $|1 - J(s)|$  is zero in a neighborhood of  $t$ . Since  $K_b$  is supported by  $[-b, b]$ , and thus the effective intervals of integrations in the  $\Delta_i$  are  $[t - b, t + b]$  and fall inside the neighborhood of  $t$  for  $n$  large, we have  $b^{-p} \mathbf{D}^{-1} \Delta_i(\boldsymbol{\theta}_t)$  equals  $\mathbf{0}$  with probability tending to 1 for  $i = 3, 4$ . This implies  $b^{-p} \sup_{\boldsymbol{\theta}_t} \|\mathbf{D}^{-1} \Delta_i(\boldsymbol{\theta}_t)\| \xrightarrow{P} 0$ ,  $i = 3, 4$ .

By C1 we can assume the local parameter space  $\Theta_t$  of  $\boldsymbol{\theta}_t$  is a compact rectangle in  $(0, \infty) \times \mathbb{R}^p$ . Therefore, when  $n$  is sufficiently large,  $\sup_{\boldsymbol{\theta}_t} \sup_{s \in [t-b, t+b]} |1/\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t|$  is uniformly (in  $n$ ) bounded, say by  $C < \infty$ . This implies, when  $n$  is large and with the notations  $\leq$ ,  $\sup$ , and  $|\cdot|$  understood component-wise for vectors,

$$\begin{aligned}& b^{-p} \sup_{\boldsymbol{\theta}_t} |\mathbf{D}^{-1} \Delta_2(\boldsymbol{\theta}_t)| \\ &= \sup_{\boldsymbol{\theta}_t} \left| \int_{-\min(1, t/0)}^{\min(1, (1-t)/0)} \mathbf{g}^*(x) \frac{J(t+bx)}{\mathbf{g}(bx) \boldsymbol{\theta}_t} K(x) \frac{\alpha_0(t+bx) - \mathbf{g}_p(bx)^\top \boldsymbol{\theta}_t^0}{b^p} dx \right| \\ &\leq C \int_{-\min(1, t/0)}^{\min(1, (1-t)/0)} |\mathbf{g}^*(x)| K(x) \left| \frac{\alpha_0(t+bx) - \mathbf{g}_p(bx)^\top \boldsymbol{\theta}_t^0}{b^p} \right| dx \rightarrow \mathbf{0},\end{aligned}$$

where the convergence to 0 is by dominated convergence and the differentiability condition C1. Now some simple algebra shows the component-wise convergence  $b^{-p} \sup_{\boldsymbol{\theta}_t} |\mathbf{D}^{-1} \Delta_2(\boldsymbol{\theta}_t)| \rightarrow \mathbf{0}$  implies  $b^{-p} \sup_{\boldsymbol{\theta}_t} \|\mathbf{D}^{-1} \Delta_2(\boldsymbol{\theta}_t)\| \rightarrow 0$ .

Now we need only show  $b^{-p} \sup_{\boldsymbol{\theta}_t} \|\mathbf{D}^{-1} \Delta_1(\boldsymbol{\theta}_t)\| \xrightarrow{P} 0$ . Thanks to the boundedness of  $\sup_{\boldsymbol{\theta}_t} \sup_{s \in [t-b, t+b]} |1/\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t|$ , it is sufficient to show  $b^{-p} \left| \mathbf{D}^{-1} \int_0^1 \mathbf{g}(s-t)(J(s)/Y(s)) K_b(s-t) dM(s) \right| \xrightarrow{P} \mathbf{0}$ , which is equivalent to the component-wise convergences

$$b^{-p} \left| \mathbf{e}_j \mathbf{D}^{-1} \int_0^1 \mathbf{g}(s-t) \frac{J(s)}{Y(s)} K_b(s-t) dM(s) \right| \xrightarrow{P} 0, \quad j = 0, \dots, p. \quad (4.3)$$

For  $j = 0, \dots, p$ , define the process

$$\widetilde{M}(u) = \int_0^u b^{-p} \mathbf{e}_j \mathbf{D}^{-1} \mathbf{g}(s-t) \frac{J(s)}{Y(s)} K_b(s-t) dM(s), \quad u \in [0, 1].$$



By the theory of counting process stochastic integrals (Aalen (1978); Andersen et al. (1993); Fleming and Harrington (1991)),  $\widetilde{M}(u)$  is a local square integrable martingale with predictable variation process

$$\langle \widetilde{M} \rangle(u) = \int_0^u b^{-2p} \{ \mathbf{e}_j \mathbf{D}^{-1} \mathbf{g}(s-t) \}^2 \frac{J(s)}{Y(s)^2} K_b(s-t)^2 Y(s) \alpha_0(s) ds.$$

By a change of variables and C3,

$$\langle \widetilde{M} \rangle(1) = \int_{-\min(1,t/0)}^{\min(1,(1-t)/0)} x^{2j} \frac{J(t+bx)}{b^{2p+1} Y(t+bx)} K(x)^2 \alpha_0(t+bx) dx \xrightarrow{P} 0. \quad (4.4)$$

By Lenglar's inequality (Andersen et al. (1993); Fleming and Harrington (1991)),

$$\Pr \left( \sup_{u \in [0,1]} |\widetilde{M}(u)| > \epsilon \right) \leq \frac{\delta}{\epsilon^2} + \Pr \left( \langle \widetilde{M} \rangle(1) > \delta \right), \quad \forall \epsilon, \delta > 0. \quad (4.5)$$

Taking lim sup on both sides of (4.5), we have by (4.4) that

$$\limsup_{n \rightarrow \infty} \Pr \left( \sup_{u \in [0,1]} |\widetilde{M}(u)| > \epsilon \right) \leq \frac{\delta}{\epsilon^2}.$$

The arbitrariness of  $\delta$  guarantees  $\lim \Pr(\sup_{u \in [0,1]} |\widetilde{M}(u)| > \epsilon)$  exists and equals 0. So  $\sup_{u \in [0,1]} |\widetilde{M}(u)|$ , and therefore  $|\widetilde{M}(1)|$ , converge in probability to 0. This concludes the proof of (4.3) and the theorem.

**Theorem 4.2 (Asymptotic normality).** *Under Conditions C1–C6, if the bandwidth is chosen such that  $\limsup a_n^2 b_n^{2p+3} < \infty$ , then we have  $a_n b_n^{1/2} \mathbf{D} \{ \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t^0 - b_n^{p+1} \mathbf{D}^{-1} \mathbf{m}_t \} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{A}_t^{-1} \mathbf{V}_t \mathbf{A}_t^{-1})$ , where  $\mathbf{A}_t$  is given in (4.1) and*

$$\mathbf{m}_t = \frac{\alpha_0^{(p+1)}(t)}{(p+1)!} \mathbf{A}_t^{-1} \int_{-\min(1,t/0)}^{\min(1,(1-t)/0)} \mathbf{g}^*(x) x^{p+1} K(x) dx, \quad (4.6)$$

$$\mathbf{V}_t = \frac{\alpha_0(t)}{y(t)} \int_{-\min(1,t/0)}^{\min(1,(1-t)/0)} \mathbf{g}^*(x)^{\otimes 2} K(x)^2 dx. \quad (4.7)$$

Moreover, if  $\hat{\mathbf{S}} = \int_0^1 \frac{\mathbf{g}(s-t)^{\otimes 2}}{\{ \mathbf{g}(s-t)^\top \hat{\boldsymbol{\theta}}_t \}^2} \frac{J(s)}{Y(s)^2} K_b(s-t)^2 dN(s)$ , then

$$a_n^2 b_n \mathbf{D} \mathbf{I}(\hat{\boldsymbol{\theta}}_t)^{-1} \hat{\mathbf{S}} \mathbf{I}(\hat{\boldsymbol{\theta}}_t)^{-1} \mathbf{D} \xrightarrow{P} \mathbf{A}_t^{-1} \mathbf{V}_t \mathbf{A}_t^{-1}.$$

This theorem implies that, when  $n$  is large, the distribution of  $\hat{\boldsymbol{\theta}}_t$  is approximately normal and its variance-covariance matrix can be estimated by the sandwich-type estimator

$$\widehat{\text{var}}(\hat{\boldsymbol{\theta}}_t) = \mathbf{I}(\hat{\boldsymbol{\theta}}_t)^{-1} \hat{\mathbf{S}} \mathbf{I}(\hat{\boldsymbol{\theta}}_t)^{-1}. \quad (4.8)$$

**Proof of Theorem 4.2.** For the local score function  $\mathbf{s}$ , we have

$$\mathbf{s}(\hat{\boldsymbol{\theta}}_t) = \mathbf{s}(\boldsymbol{\theta}_t^0) - \mathbf{I}(\boldsymbol{\theta}_t^*) (\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t^0),$$

where  $\boldsymbol{\theta}_t^*$  is on the line segment joining  $\hat{\boldsymbol{\theta}}_t$  and  $\boldsymbol{\theta}_t^0$ . Since  $\mathbf{s}(\hat{\boldsymbol{\theta}}_t) = \mathbf{0}$ ,

$$\mathbf{D}^{-1} \mathbf{I}(\boldsymbol{\theta}_t^*) \mathbf{D}^{-1} a_n b_n^{1/2} \mathbf{D} (\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t^0) = a_n b_n^{1/2} \mathbf{D}^{-1} \mathbf{s}(\boldsymbol{\theta}_t^0).$$

Mimicking the proof of Theorem 4.1, we can show C1–C4 imply  $\mathbf{D}^{-1} \mathbf{I}(\hat{\boldsymbol{\theta}}_t) \mathbf{D}^{-1} - \mathbf{A}_t \alpha_0(t) / \{\mathbf{e}_0^\top \hat{\boldsymbol{\theta}}_t\}^2 \xrightarrow{P} \mathbf{0}^{\otimes 2}$ , with  $\mathbf{A}_t$  given by (4.1). As convergence in probability is preserved by continuous mappings,  $\mathbf{A}_t \alpha_0(t) / \{\mathbf{e}_0^\top \hat{\boldsymbol{\theta}}_t\}^2 \xrightarrow{P} \mathbf{A}_t / \alpha_0(t)$ . It follows that  $\mathbf{D}^{-1} \mathbf{I}(\hat{\boldsymbol{\theta}}_t) \mathbf{D}^{-1} \xrightarrow{P} \mathbf{A}_t / \alpha_0(t)$ , from which we also have  $\mathbf{D}^{-1} \mathbf{I}(\boldsymbol{\theta}_t^*) \mathbf{D}^{-1} \xrightarrow{P} \mathbf{A}_t / \alpha_0(t)$ , since  $\boldsymbol{\theta}_t^*$  is sandwiched by  $\hat{\boldsymbol{\theta}}_t$  and  $\boldsymbol{\theta}_t^0$ .

We next prove  $a_n b_n^{1/2} \{\mathbf{D}^{-1} \mathbf{s}(\boldsymbol{\theta}_t^0) - b_n^{p+1} \mathbf{A}_t \mathbf{m}_t / \alpha_0(t)\} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}_t / \alpha_0(t)^2)$ , so that by Slutsky's Theorem,

$$\begin{aligned} & a_n b_n^{1/2} \mathbf{D} (\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t^0 - b_n^{p+1} \mathbf{D}^{-1} \mathbf{m}_t) \\ &= a_n b_n^{1/2} \mathbf{D} \mathbf{I}(\boldsymbol{\theta}_t^*)^{-1} \mathbf{D} \{\mathbf{D}^{-1} \mathbf{s}(\boldsymbol{\theta}_t^0) - b_n^{p+1} \frac{\mathbf{A}_t \mathbf{m}_t}{\alpha_0(t)}\} \\ & \quad + a_n b_n^{p+3/2} \{\mathbf{D} \mathbf{I}(\boldsymbol{\theta}_t^*)^{-1} \mathbf{D} \frac{\mathbf{A}_t \mathbf{m}_t}{\alpha_0(t)} - \mathbf{m}_t\} \\ & \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{A}_t^{-1} \mathbf{V}_t \mathbf{A}_t^{-1}). \end{aligned}$$

Define a vector-valued process  $\widetilde{\mathbf{M}}(u)$ ,  $u \in [0, 1]$ , and a vector-valued random variable  $\mathbf{R}$  by

$$\begin{aligned} \widetilde{\mathbf{M}}(u) &= \int_0^u a_n b_n^{1/2} \frac{\mathbf{D}^{-1} \mathbf{g}(s-t) J(s)}{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t^0 Y(s)} K_b(s-t) dM(s) \\ &\triangleq \int_0^u \mathbf{H}(s) dM(s), \\ \mathbf{R} &= \int_0^1 \frac{a_n b_n^{1/2} \mathbf{D}^{-1} \mathbf{g}(s-t) J(s) K_b(s-t)}{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t^0} \{\alpha_0(s) - \mathbf{g}(s-t)^\top \boldsymbol{\theta}_t^0\} ds, \end{aligned}$$

so we can write  $a_n b_n^{1/2} \mathbf{D}^{-1} \mathbf{s}(\boldsymbol{\theta}_t^0) = \widetilde{\mathbf{M}}(1) + \mathbf{R}$ . By stochastic integration theory,  $\widetilde{\mathbf{M}}(u)$  is a local square integrable martingale with predictable variation process given by

$$\begin{aligned} \langle \widetilde{\mathbf{M}} \rangle(u) &= \int_0^u \mathbf{H}(s)^{\otimes 2} Y(s) \alpha_0(s) ds \\ &= \int_0^u \frac{\mathbf{g}^*((s-t)/b)^{\otimes 2} J(s)}{\{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t^0\}^2 Y(s) / a_n^2 b} \frac{1}{b} K\left(\frac{s-t}{b}\right)^2 \alpha_0(s) ds \end{aligned}$$

$$= \int_{\max\{-1, -t/b\}}^{\min\{1, (u-t)/b\}} \frac{\mathbf{g}^*(x)^{\otimes 2} \alpha_0(t+bx)}{\{\mathbf{g}^*(bx)^\top \boldsymbol{\theta}_t^0\}^2} \frac{J(t+bx)}{Y(t+bx)/a_n^2} K(x)^2 dx.$$

By C5, this converges (point-wise) in probability to

$$\int_{\max\{-1, -t/0\}}^{\min\{1, (u-t)/0\}} \frac{\mathbf{g}^*(x)^{\otimes 2}}{\alpha_0(t)} \frac{1}{y(t)} K(x)^2 dx. \quad (4.9)$$

Moreover, for any  $\epsilon > 0$  and  $j = 0, \dots, p$ , that

$$1\left\{\left|\mathbf{e}_j^\top \mathbf{H}(s)\right| > \epsilon\right\} = 1\left\{\left|\frac{(s-t)^j}{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t^0} \frac{J(s)}{Y(s)/a_n^2} K\left(\frac{s-t}{b_n}\right)\right| > a_n b_n^{1/2+j} \epsilon\right\}$$

converges to 0 uniformly in  $s$  as  $a_n^2 b_n^{2p+1} \rightarrow \infty$  implies  $a_n b_n^{1/2+j} \rightarrow \infty$ . It follows that

$$\int_0^1 \{\mathbf{e}_j^\top \mathbf{H}(s)\}^2 1\left\{\left|\mathbf{e}_j^\top \mathbf{H}(s)\right| > \epsilon\right\} ds \xrightarrow{P} 0.$$

By the Martingale Central Limit Theorem (Aalen, Borgan and Gjessing (2008); Andersen et al. (1993); Fleming and Harrington (1991)), the process  $\widetilde{\mathbf{M}}(u)$  converges in distribution (in the Skorohod space) to a  $(p+1)$ -dimensional zero-mean Gaussian martingale with predictable variation variance process given by (4.9). This means  $\widetilde{\mathbf{M}}(1)$  converges in distribution to a zero mean normal vector with variance covariance matrix  $\int_{-\min(1, t/0)}^{\min(1, (1-t)/0)} (\mathbf{g}^*(x)^{\otimes 2} / \alpha_0(t)) (1/y(t)) K(x)^2 dx = \mathbf{V}_t / \alpha_0(t)^2$ . By Taylor's expansion and the assumption that  $\limsup a_n^2 b_n^{2p+3} < \infty$ , it can be shown that  $\mathbf{R} - a_n b_n^{p+3/2} \mathbf{A}_t \mathbf{m}_t / \alpha_0(t) \xrightarrow{P} \mathbf{0}$ . So, by Slutsky's Theorem,

$$\begin{aligned} a_n b_n^{1/2} \left\{ \mathbf{D}^{-1} \mathbf{s}(\boldsymbol{\theta}_t^0) - b_n^{p+1} \frac{\mathbf{A}_t \mathbf{m}_t}{\alpha_0(t)} \right\} &= \widetilde{\mathbf{M}}(1) + \mathbf{R} - a_n b_n^{p+3/2} \frac{\mathbf{A}_t \mathbf{m}_t}{\alpha_0(t)} \\ &\xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{\mathbf{V}_t}{\alpha_0(t)^2}\right), \end{aligned}$$

which concludes the proof of the asymptotic normality of  $\hat{\boldsymbol{\theta}}_t$ .

Finally, by considering a matrix-valued processes indexed by  $\boldsymbol{\theta}_t$ ,

$$\int_0^u a_n^2 b_n \frac{\mathbf{D}^{-1} \mathbf{g}(s-t)^{\otimes 2} \mathbf{D}^{-1}}{\{\mathbf{g}(s-t)^\top \boldsymbol{\theta}_t\}^2} \frac{J(s)}{Y(s)^2} K_b(s-t)^2 dN(s), \quad u \in [0, 1],$$

we can show  $a_n^2 b_n \mathbf{D}^{-1} \hat{\mathbf{S}} \mathbf{D}^{-1} \xrightarrow{P} \mathbf{V}_t / \alpha_0(t)^2$  by an argument similar to the one leading to  $\mathbf{D}^{-1} \mathbf{I}(\hat{\boldsymbol{\theta}}_t) \mathbf{D}^{-1} \xrightarrow{P} \mathbf{A}_t / \alpha_0(t)$ . Therefore

$$a_n^2 b_n \mathbf{D} \mathbf{I}(\hat{\boldsymbol{\theta}}_t)^{-1} \hat{\mathbf{S}} \mathbf{I}(\hat{\boldsymbol{\theta}}_t)^{-1} \mathbf{D}$$

$$\begin{aligned}
&= \{\mathbf{D}^{-1}\mathbf{I}(\hat{\boldsymbol{\theta}}_t)\mathbf{D}^{-1}\}^{-1}\{a_n^2 b_n \mathbf{D}^{-1}\hat{\mathbf{S}}\mathbf{D}^{-1}\}\{\mathbf{D}^{-1}\mathbf{I}(\hat{\boldsymbol{\theta}}_t)\mathbf{D}^{-1}\}^{-1} \\
&\xrightarrow{P} \mathbf{A}_t^{-1}\mathbf{V}_t\mathbf{A}_t^{-1}.
\end{aligned}$$

This concludes the proof of the theorem.

#### 4.2. Asymptotic bias and variance and the choice of the order of the local polynomial

As implied by Theorem 4.2, the MLPLE estimator  $\hat{\boldsymbol{\theta}}_t$  for  $\boldsymbol{\theta}_t$  is generally biased, unless we know a priori that the intensity function to be estimated is a polynomial. It is important to study the bias and variance of the estimator in order to strike a balance between them by suitable choice of the bandwidth. Due to the lack of a closed-form expression for the estimator, the exact bias and variance are not easy to work out, and in situations where the exact expressions of the bias and variance of a biased estimator are easily available, they are typically intractable. Therefore, it is common to work with suitable asymptotic versions of the bias and variance. In the sequel, the *asymptotic bias*, *asymptotic variance-covariance*, and *asymptotic mean squared error* are in the sense of Shao (2003), which are not necessarily the leading terms in the exact versions of the corresponding quantities. As a corollary to Theorem 4.2, we have the following result concerning the asymptotic bias, variance, and mean square error of the MLPLE  $\hat{\boldsymbol{\theta}}_t = (\hat{\alpha}(t), \hat{\alpha}'(t), \dots, \hat{\alpha}^{(p)}(t))^\top$ .

**Corollary 4.3.** *Assume that the conditions of Theorem 4.2 hold. Let  $\mathbf{m}_t$  and  $\mathbf{V}_t$  be given by (4.6) and (4.7). Then, the asymptotic bias and variance-covariance matrix of the MLPLE  $\hat{\boldsymbol{\theta}}_t$  are given, respectively, by  $b_n^{p+1}\mathbf{D}^{-1}\mathbf{m}_t$  and  $(a_n^2 b_n)^{-1}\mathbf{D}^{-1}\mathbf{A}_t^{-1}\mathbf{V}_t\mathbf{A}_t^{-1}\mathbf{D}^{-1}$ .*

It is possible to show that under different conditions on  $Y$  and  $b$ , the asymptotic bias and variance-covariance given here are also the leading terms of the bias and variance-covariance of  $\hat{\boldsymbol{\theta}}_t$ , and therefore the *approximate* bias and variance. However the conditions are very cumbersome, and the condition on the shrinking rate of  $b$  seems incompatible with that assumed by Theorem 4.2. We consent ourselves with the asymptotic bias and variance.

In applications the commonly used kernel functions, such as the normal density kernel  $K(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ , the Epanechnikov kernel  $K(x) = (3/4)(1 - x^2)_+$ , and the general beta kernel  $K(x) = [\Gamma(\lambda + 3/2)/\{\Gamma(1/2)\Gamma(\lambda + 1)\}](1 - x^2)_+^\lambda$ ,  $\lambda \in [0, \infty)$ , are symmetric about 0. As symmetric kernels have zero odd-order moments, for  $t \in (0, 1)$ ,  $\boldsymbol{\mu}_t(K) = \int_{-\min(1,t/0)}^{\min(1,(1-t)/0)} \mathbf{g}_p^*(x)x^{p+1}dx$  is independent of  $t$  with the form

$$(\dots, \star, 0, \star, 0)^\top, \quad (4.10)$$

and the matrices  $\mathbf{A}_t$  and  $\mathbf{A}_t^{-1}$  are independent of  $t$  and have the common form

$$\begin{bmatrix} \star & 0 & \star & \cdots & & \\ 0 & \star & 0 & \cdots & & \\ \star & 0 & \star & \cdots & & \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ & & & & \star & 0 \\ \cdots & & & \cdots & 0 & \star \end{bmatrix}. \tag{4.11}$$

Therefore,  $\mathbf{m}_t \propto \mathbf{A}_t^{-1} \boldsymbol{\mu}_t$  also takes the form (4.10). This means the asymptotic bias of  $\hat{\alpha}^{(\nu)}(t) = \mathbf{e}_{\nu,p}^\top \hat{\boldsymbol{\theta}}_t$  for  $\alpha^{(\nu)}(t)$  is of order  $b^{p+1-\nu}$  (unless  $\alpha^{(\nu)}(t) = 0$ ) when  $p - \nu$  is odd, and is 0 when  $p - \nu$  is even. However, at the boundary points  $t = 0$  and 1, the vector  $\boldsymbol{\mu}_t$  does not take the form (4.10), nor does the matrix  $\mathbf{A}_t$  take the form (4.11). Therefore,  $\mathbf{m}_t$  is not generally of the special form (4.10), and the asymptotic bias of  $\hat{\alpha}^{(\nu)}(0)$  and  $\hat{\alpha}^{(\nu)}(1)$  is of order  $b^{p+1-\nu}$  no matter  $p - \nu$  is even or odd. This implies that when  $p - \nu$  is even, the estimator for  $\hat{\alpha}^{(\nu)}(t)$  converges to  $\alpha^{(\nu)}(t)$  more slowly at a boundary  $t$  than at an interior  $t$  and suffers from *boundary effects*. As the Ramlau-Hansen estimator for the intensity function is essentially the MLPLE with  $p = \nu = 0$ , it suffers from boundary effects as noted, for instance, by Andersen et al. (1993) and Nielsen and Tanggaard (2001). At the same time, if  $p - \nu$  is odd, then the rate of convergence of  $\hat{\alpha}^{(\nu)}(t)$  to  $\alpha^{(\nu)}(t)$  is the same for all  $t \in [0, 1]$  and, in this sense, the MLPLE for  $\alpha^{(\nu)}(t)$  has an *automatic boundary bias correction* property.

The asymptotic variance of  $\hat{\alpha}^{(\nu)}(t)$  is  $(a^2 b^{2\nu+1})^{-1} \nu!^2 \mathbf{e}_\nu^\top \mathbf{A}_t^{-1} \mathbf{V}_t \mathbf{A}_t^{-1} \mathbf{e}_\nu$ , which is of order  $(a^2 b^{2\nu+1})^{-1}$  for all  $t \in [0, 1]$ . The observations about the behavior of the asymptotic bias and variance of the MLPLE estimator imply that, if we are interested in estimating  $\alpha^{(\nu)}$  using the MLPLE and a symmetric kernel is to be used, it is advisable to choose the order of the local polynomials to be  $\nu$  plus an odd integer. Meanwhile, as choosing large  $p$  contradicts the principle of local modeling and tends to reduce the computational attractiveness of the estimator, the practically relevant advice is to choose  $p = \nu + 1$  or, if data are abundant,  $\nu + 3$ . In the sequel, we assume  $p - \nu$  is odd if the kernel  $K$  in question is symmetric.

### 4.3. Asymptotically optimal bandwidth

From Corollary 4.3, it is clear that in order to reduce the asymptotic bias of the MLPLE we should use a smaller bandwidth, but to reduce the asymptotic variance, large bandwidths are preferred. An asymptotically optimal bandwidth that takes into account both accuracy and precision can be obtained by minimizing the asymptotic mean square error (AMSE) or the asymptotic integrated

mean square error (AIMSE). Here, the AMSE of the estimator is defined as the squared asymptotic bias plus the asymptotic variance, and the AIMSE is defined as the integral of the AMSE. As direct consequences of Corollary 4.3 and Theorem 4.2 we have the following.

**Corollary 4.4.** *Fix  $t \in [0, 1]$ . Assume the conditions of Theorem 4.2 hold. Then the AMSE of the MLPLE  $\hat{\alpha}^{(\nu)}(t)$  for  $\alpha^{(\nu)}(t)$  is*

$$\begin{aligned} \text{amse}\{\hat{\alpha}^{(\nu)}(t)\} &= \left\{ \frac{\alpha_0^{(p+1)}(t)\nu!}{(p+1)!} \right\}^2 \mathbf{e}_\nu^\top \mathbf{A}_t^{-1} \boldsymbol{\mu}_t^{\otimes 2} \mathbf{A}_t^{-1} \mathbf{e}_\nu \times b^{2(p+1-\nu)} \\ &\quad + \frac{\alpha_0(t)}{y(t)} \nu!^2 \mathbf{e}_\nu^\top \mathbf{A}_t^{-1} \boldsymbol{\Sigma}_t \mathbf{A}_t^{-1} \mathbf{e}_\nu \times (a_n^2 b_n^{2\nu+1})^{-1}, \end{aligned} \quad (4.12)$$

where  $\boldsymbol{\mu}_t = \int_{-\min(1,t/0)}^{\min(1,(1-t)/0)} \mathbf{g}^*(x) x^{p+1} K(x)^2 dx$ ,  $\boldsymbol{\Sigma}_t = \int_{-\min(1,t/0)}^{\min(1,(1-t)/0)} \mathbf{g}^*(x)^{\otimes 2} K(x) dx$ . If  $\alpha^{(p+1)}(t) \neq 0$ ,  $t \in [0, 1]$ , the asymptotically optimal bandwidth for  $\hat{\alpha}^{(\nu)}(t)$  is

$$\begin{aligned} b_{n,\text{opt}}(t) &= \left[ \frac{\mathbf{e}_\nu^\top \mathbf{A}_t^{-1} \boldsymbol{\Sigma}_t \mathbf{A}_t^{-1} \mathbf{e}_\nu}{\mathbf{e}_\nu^\top \mathbf{A}_t^{-1} \boldsymbol{\mu}_t^{\otimes 2} \mathbf{A}_t^{-1} \mathbf{e}_\nu} \right]^{1/(2p+3)} \left[ \frac{(p+1)!^2 (2\nu+1)}{2(p+1-\nu)} \right]^{1/(2p+3)} \\ &\quad \times \left[ \frac{\alpha_0(t)}{\{\alpha_0^{(p+1)}(t)\}^2 y(t)} \right]^{1/(2p+3)} \times a_n^{-2/(2p+3)}. \end{aligned} \quad (4.13)$$

**Corollary 4.5.** *Assume the conditions of Theorem 4.2 hold for all  $t \in (0, 1)$ . Then the AIMSE of the MLPLE  $\hat{\alpha}^{(\nu)}(t)$  for  $\alpha^{(\nu)}(t)$  is*

$$\begin{aligned} \text{aimse}\{\hat{\alpha}^{(\nu)}\} &= \int_0^1 \text{amse}\{\hat{\alpha}^{(\nu)}(t)\} w(t) dt \\ &= \left\{ \frac{\nu!}{(p+1)!} \right\}^2 \left\{ \mathbf{e}_\nu^\top \mathbf{A}^{-1} \boldsymbol{\mu}^{\otimes 2} \mathbf{A}^{-1} \mathbf{e}_\nu \right\} \int_0^1 \{\alpha_0^{(p+1)}(t)\}^2 w(t) dt \\ &\quad \times b^{2(p+1-\nu)} \\ &\quad + \nu!^2 \mathbf{e}_\nu^\top \mathbf{A}^{-1} \boldsymbol{\Sigma} \mathbf{A}^{-1} \mathbf{e}_\nu \int_0^1 \frac{\alpha_0(t) w(t)}{y(t)} dt \times (a_n^2 b_n^{2\nu+1})^{-1}, \end{aligned} \quad (4.14)$$

where  $w(t)$  is a bounded weighting function, and  $\boldsymbol{\mu}$ ,  $\boldsymbol{\Sigma}$ , and  $\mathbf{A}$  are, respectively, the  $\boldsymbol{\mu}_t$ ,  $\boldsymbol{\Sigma}_t$  and  $\mathbf{A}_t$  in Corollary 4.4 evaluated at any  $t \in (0, 1)$ . If  $\int_0^1 \{\alpha^{(p+1)}(t)\}^2 w(t) dt \neq 0$ , the asymptotically optimal global bandwidth for  $\hat{\alpha}^{(\nu)}$  is

$$\begin{aligned} b_{n,\text{opt}}^* &= \left[ \frac{\mathbf{e}_\nu^\top \mathbf{A}^{-1} \boldsymbol{\Sigma} \mathbf{A}^{-1} \mathbf{e}_\nu}{\mathbf{e}_\nu^\top \mathbf{A}^{-1} \boldsymbol{\mu}^{\otimes 2} \mathbf{A}^{-1} \mathbf{e}_\nu} \right]^{1/(2p+3)} \left[ \frac{(p+1)!^2 (2\nu+1)}{2(p+1-\nu)} \right]^{1/(2p+3)} \\ &\quad \times \left[ \frac{\int_0^1 \alpha_0(t) w(t) / y(t) dt}{\int_0^1 \{\alpha_0^{(p+1)}(t)\}^2 w(t) dt} \right]^{1/(2p+3)} \times a_n^{-2/(2p+3)}. \end{aligned} \quad (4.15)$$

### 5. Practical Bandwidth Selection

The performance of the MLPLE is highly influenced by the choice of the bandwidth parameter. The explicit expressions for the optimal bandwidths presented in Section 4.3 are not directly applicable in practice, since they involve the unknown intensity function and its derivatives that we want to estimate. However, these explicit expressions allow us to construct plug-in type estimators for the optimal global or local bandwidths by replacing the unknown quantities by their estimators. Depending on how we choose the initial estimators for the unknown quantities, we can have different estimators for the optimal bandwidths. Here we propose a crude global bandwidth selector which is similar in spirit to the *rule of thumb* bandwidth selector for the local polynomial estimator of a regression function (Fan and Gijbels (1996)).

Let  $U_1 = \int_0^1 \alpha_0(t)w(t)/y(t)dt$  and  $U_2 = \int_0^1 \{\alpha_0^{(p+1)}(t)\}^2 w(t)dt$  be the unknown quantities in (4.15). If we assume the conditions of Corollary 4.5, then by considering the zero mean local square integrable martingale

$$\left\{ \int_0^u \frac{J(t)w(t)}{Y(t)^2} dM(t) \right\}, \quad u \in [0, 1],$$

and by using Lengart's inequality, we note a consistent estimator for  $U_1$  is  $\hat{U}_1 = a_n^2 \int_0^1 [J(t)w(t)/Y(t)^2] dN(t)$ .

To estimate  $U_2$ , it seems that we have to estimate  $\alpha^{(p+1)}$ , and do a numerical integration. While there are clearly many possibilities for a pilot estimate of  $\alpha^{(p+1)}$ , we simply estimate  $\alpha$  using a parametric method and then use the  $(p+1)$ th derivative of the estimate as an estimator of  $\alpha^{(p+1)}$ . More specifically, we impose a global polynomial form for  $\alpha$ ,  $\alpha(t) = \sum_{i=0}^{p+q} \gamma_i t^i$ , with  $q \geq 3$ , insert it into the modified global partial log-likelihood,

$$\begin{aligned} \tilde{l} &= \int_0^1 \{\log(Y(t)\alpha(t))\} \frac{J(t)}{Y(t)} dN(t) - \int_0^1 \alpha(t)J(t)dt \\ &\approx \int_0^1 \{\log(Y(t)\alpha(t))\} \frac{J(t)}{Y(t)} dN(t) - \int_0^1 \alpha(t)dt, \end{aligned}$$

do the maximization with respect  $(\gamma_0, \dots, \gamma_{p+q})^\top$  to obtain the estimates  $\hat{\gamma}_i$  of the coefficients, and estimate  $\alpha^{(p+1)}$  by

$$\hat{\alpha}^{(p+1)}(t) = \sum_{i=1}^q \frac{(p+i)! \hat{\gamma}_{p+i}}{(i-1)!} t^{i-1}. \tag{5.1}$$

Now the estimator for  $U_2$  is

$$\hat{U}_2 = \int_0^1 \{\hat{\alpha}^{(p+1)}(t)\}^2 w(t)dt.$$

Plugging  $\hat{U}_1$  and  $\hat{U}_2$  into (4.15), we have an estimator for  $b_{n,\text{opt}}^*$ . We refer to it as the *rule of thumb* bandwidth selector, and denote it by  $\hat{b}_{\text{ROT}}^*$ , so

$$\hat{b}_{\text{ROT}}^* = \left[ \frac{\mathbf{e}_\nu^\top \mathbf{A}^{-1} \boldsymbol{\Sigma} \mathbf{A}^{-1} \mathbf{e}_\nu}{\mathbf{e}_\nu^\top \mathbf{A}^{-1} \boldsymbol{\mu}^{\otimes 2} \mathbf{A}^{-1} \mathbf{e}_\nu} \right]^{1/(2p+3)} \left[ \frac{(p+1)!^2 (2\nu+1)}{2(p+1-\nu)} \right]^{1/(2p+3)} \\ \times \left[ \frac{\int_0^1 J(t)w(t)/Y(t)^2 dN(t)}{\int_0^1 \{\hat{\alpha}^{(p+1)}(t)\}^2 w(t) dt} \right]^{1/(2p+3)}, \quad (5.2)$$

with  $\hat{\alpha}^{(p+1)}$  given by (5.1). It is worth mentioning that  $\hat{b}_{\text{ROT}}^*$  is free of  $a_n$ , which means it is not necessary to work out the normalizing sequence  $a_n$  of the exposure process  $Y^{(n)}$  before we can use the ROT bandwidth selector.

In obtaining the pilot estimate for  $\alpha^{(p+1)}$ , we have chosen to use the modified global partial log-likelihood instead of the unmodified partial likelihood (3.1). The main motivation for this choice is computational stability. If we work with (3.1), we need to evaluate the integral  $\int_0^1 Y(t)\alpha(t)dt$ , that is numerically unstable when  $Y(t)$  is a jump process as in survival analysis.

## 6. Numerical Studies

In this section, we use simulated data to verify the automatic boundary correction property of the MLPLE and to assess the performance of the variance estimator (4.8) and the ROT bandwidth selector.

### 6.1. Automatic boundary bias correction

We simulated a Poisson process on  $[0, 1]$  with intensity process  $500\alpha(t)$ , where  $\alpha(t) = 1 + e^{-t} \cos(4\pi t)$ , and estimated  $\alpha$  and its derivative using the MLPLE with the Epanechnikov kernel. The simulation and estimation was repeated  $N = 100$  times. The bandwidth used in the intensity estimation was  $b = 0.088$  and, in the derivative estimation,  $b = 0.13$ . The order of the local polynomial was chosen to be  $p = \nu + 1$ . The average of estimated intensity curves is graphed in Figure 1, together with the true intensity curve and the average of the estimates obtained from the Ramlau-Hansen estimator. The average of the estimated derivative curves is shown in Figure 2, together with the true derivative. From these two figures we note the MLPLE of  $\alpha^{(\nu)}(t)$  with  $p = \nu + 1$  had the automatic boundary bias correction property, but with  $p = \nu$  it suffered from the boundary effects.

### 6.2. Variance estimator

In the simulations reported in Section 6.1, we also estimated the standard error of the MLPLE estimators using the formula

$$\widehat{\text{var}}\{\hat{\alpha}^{(\nu)}(t)\} = \mathbf{e}_\nu^\top \widehat{\text{var}}(\hat{\boldsymbol{\theta}}_t) \mathbf{e}_\nu,$$



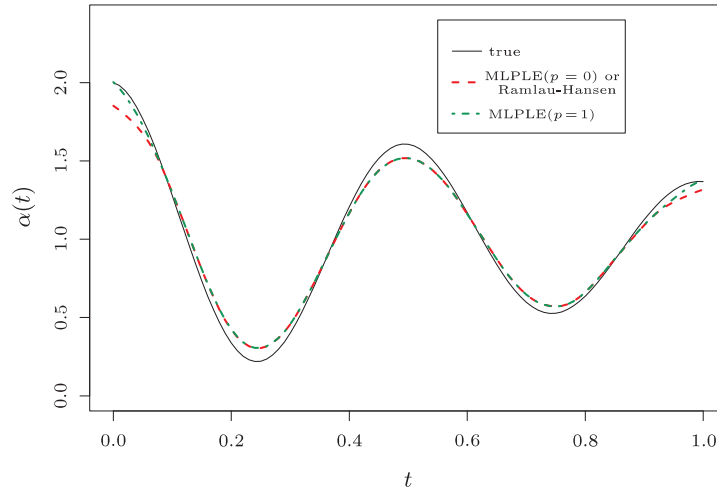


Figure 1. True intensity function and the averages of the estimates using the MLPLE with  $p = 1$  and the Ramlau-Hansen estimator. The MLPLE with  $p = 1$  did not suffer from the boundary effects, but the Ramlau-Hansen estimator did.

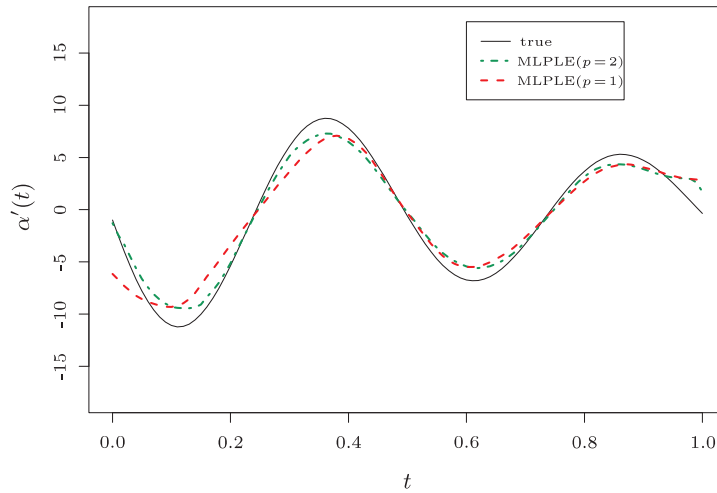


Figure 2. Derivative of the intensity function and the averages of the estimated derivative curves obtained from the MLPLE with  $p = \nu = 1$  and  $p = \nu + 1 = 2$ . The MLPLE with  $p = \nu + 1$  did not suffer from the boundary effects while it did with  $p = \nu$ .

where  $\widehat{\text{var}}(\hat{\theta}_t)$  is given by (4.8). The average of estimated standard error of  $\hat{\alpha}^{(\nu)}(t)$ , and its true standard error, approximated by the standard deviation of the estimated value of  $\hat{\alpha}^{(\nu)}(t)$ , are plotted in Figure 3. The ratio of the estimated standard error to the true standard error is shown in Figure 4. From these figures

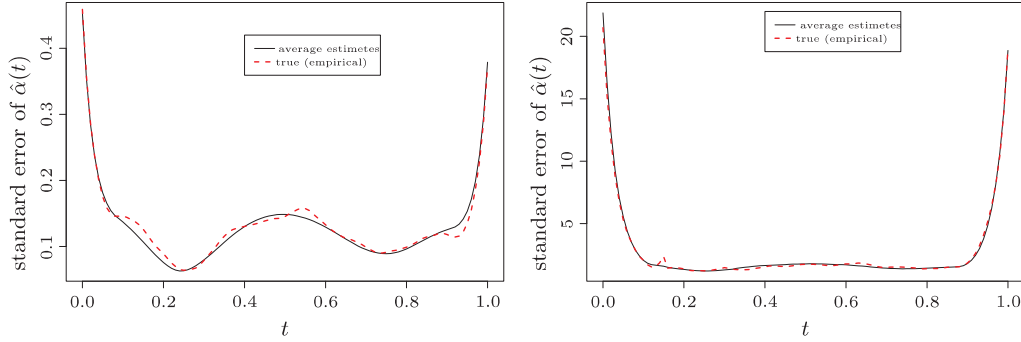


Figure 3. The average of the estimated standard error and the empirical standard error of  $\hat{\alpha}^{(\nu)}(t)$ . Left,  $\nu = 0$ ; right,  $\nu = 1$ .

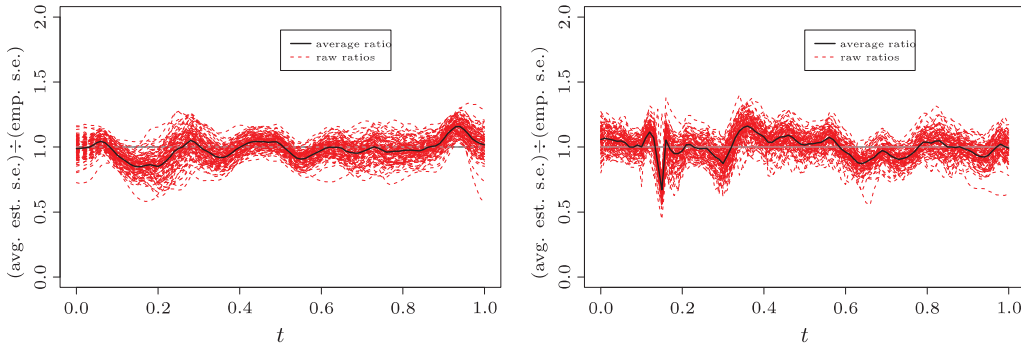


Figure 4. The ratio of the estimated standard error of  $\hat{\alpha}^{(\nu)}(t)$  to the empirical standard error: left,  $\nu = 0$ ; right,  $\nu = 1$ .

Table 1. ROT bandwidths for the MLPLE  $\hat{\alpha}^{(\nu)}$ .

$\nu$	Summary of the 100 simulated $\hat{b}_{\text{ROT}}^*$						$b_{\text{opt}}^*$
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	
0	0.07374	0.07828	0.08051	0.08129	0.08456	0.09092	0.08835
1	0.13170	0.14350	0.14720	0.14780	0.15280	0.16560	0.13280

we can see the variance estimator had satisfactory performance.

### 6.3. ROT bandwidth selector

For each of the 100 sample paths of the Poisson process simulated in Section 6.1, we calculated the ROT bandwidth using (5.2), with the tuning parameters chosen as  $p = \nu + 1$ ,  $q = 3$ ,  $w(t) \equiv 1$ . The kernel function used was the Epanechnikov kernel as before. The ROT bandwidths for the MLPLE  $\hat{\alpha}^{(\nu)}$  is summarized in Table 1.

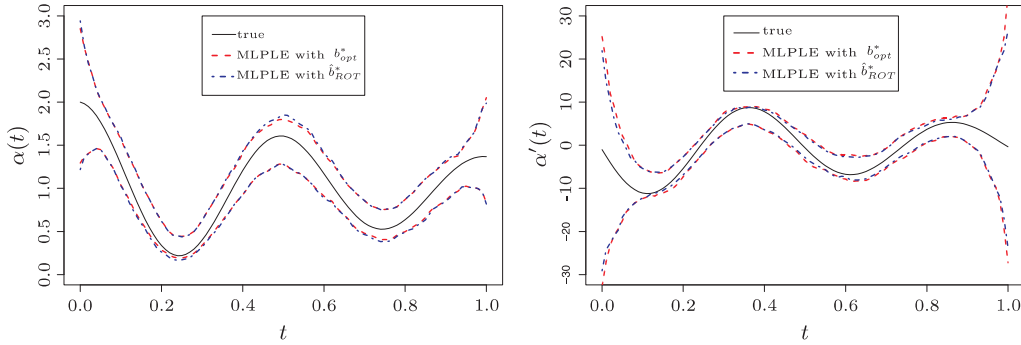


Figure 5. True intensity function and its derivative, and the lower and upper 0.025 quantiles of their estimates obtained using the MLPLE with the optimal and the ROT bandwidths.

Comparing the ROT bandwidths with the corresponding asymptotically optimal bandwidths, we might conclude that the ROT bandwidth selector as an estimator of the theoretically optimal bandwidth is generally biased and the direction of the bias seems unpredictable. The performance of the ROT bandwidth selector depends on how well the unknown intensity function can be approximated by a low-order polynomial. However, the range of the ratio of the ROT bandwidth to the optimal bandwidth was fairly close to 1,  $[0.8347, 1.0290]$  for  $\nu = 0$  and  $[0.9913, 1.2470]$  for  $\nu = 1$ . Moreover, the estimator with the ROT bandwidth might have similar or even better finite sample performance than with the optimal bandwidth. For instance, in our simulations, the empirical IMSE of  $\hat{\alpha}(t)$  was 0.0234 if the optimal bandwidth was used, and was 0.0243 if the ROT bandwidths were used. The IMSE of  $\hat{\alpha}'(t)$  was 42.01 if the optimal bandwidth was used, and is a slightly better 39.48 if the ROT bandwidths were used. We have plotted in Figure 5 the true derivative curve and the point-wise lower and upper 0.025 quantiles of the estimates obtained with the optimal and the ROT bandwidths, which also implies the MLPLE with the ROT bandwidth selector had better or very similar finite sample performances to that with the asymptotically optimal bandwidth. We conclude that the performance of the ROT bandwidth selector was satisfactory.

## 7. Conclusion and Discussion

We have proposed estimators for the intensity function of a counting process and its derivatives through using the method of maximum local partial likelihood. The resultant estimators are easy and fast to evaluate numerically, and they possess an automatic boundary bias correction property if the tuning parameters in the definition of the estimators are suitably chosen. We have also proved the consistency and asymptotic normality of the estimators under mild regularity

conditions. We have calculated the asymptotic bias, variance and mean square errors of the estimators and based on these derived the explicit formulas of the theoretically optimal bandwidths that minimize the asymptotic mean square error or integrated mean square error. For practical bandwidth selection, we have proposed an easy-to-evaluate data-driven bandwidth selector which has favorable numerical performance.

Along the way of proving the consistency of the estimators, we proposed and proved a lemma which is potentially useful in establishing the consistency of general  $Z$  estimators constructed from biased estimating equations, which is a generalization of Theorem 5.9 of van der Vaart (1998).

It is worth noting that the automatic boundary bias correction property of the local polynomial estimators in the contexts of nonparametric regression and density function estimation have been long noticed and well understood, and the same advice has been made for the choice of the order of the local polynomials; see Fan and Gijbels (1996) and the references therein for details. In the context of hazard rate function estimation, or more generally the counting process intensity function estimation, there has been relatively little effort in investigating the automatic boundary bias correction property of the local polynomial type estimators for the counting process intensity function and its derivatives except, for example, Jiang and Doksum (2003) and Chen, Yip, and Lam (2009).

Under same regularity conditions the asymptotic properties of the MLPLE estimators are exactly the same as the biased martingale estimating equation based local polynomial estimators proposed in Chen et al. (2008) and further investigated in Chen, Yip, and Lam (2009). These two estimators are asymptotically equivalent, and both enjoy the nice features of the local polynomial estimators in the contexts of regression function estimation. However, our experience is that the MLPLE is roughly five times quicker to evaluate than the local polynomial estimators based on the martingale estimating equations.

From the construction of the local partial log-likelihood (2.1)–(3.2), it can be seen that the local partial likelihood considered in this paper is essentially a continuous-observation version of the composite conditional likelihood. The consistency and asymptotic normality of the MLPLE echo those same properties of the MCCLEs (maximum composite conditional likelihood estimators), although as estimators for the intensity function and its derivatives, the MLPLEs are not efficient or even root- $n$  consistent as the MCCLEs for finite dimensional parameters (see e.g., Liang (1984)).

The ROT global bandwidth selector proposed in Section 5 is of ad hoc nature, and are not expected to work satisfactorily when the intensity function to be estimated has highly inhomogeneous shape. In fact in these situations any constant bandwidth cannot be expected to work very well, and variable/local bandwidths should be used instead.

### Acknowledgements

The author would like to thank the referee, an AE and the Co-Editors for constructive comments that have led to an improved presentation. This work is partly supported by a UNSW SFRGP/ECR grant.

### Appendix

#### A. Consistency of $Z$ -estimators

The following lemma is handy in proving the consistency of  $Z$ -estimators in situations where a usable fixed limit of the estimating equations does not exist. It is a generalization of Theorem 5.9 of van der Vaart (1998).

**Lemma A.1 (Consistency of  $Z$ -estimators).** *Let  $f_n$  be a sequence of random functions and  $g_n$  a sequence of nonrandom functions defined on a metric space  $(\Theta, d)$  and taking values in a metric space  $(Y, \rho)$ . If there exist a  $\theta_0 \in \Theta$ , a  $y_0 \in Y$ , a constant  $\eta > 0$ , a function  $\delta$  from  $(0, \eta)$  to  $(0, \eta)$ , and a sequence of positive constants  $c_n \rightarrow 0$ , such that*

- (i)  $g_n(\theta_0) = y_0$ , for  $n$  large;
- (ii)  $\inf_{\theta: d(\theta, \theta_0) \geq \epsilon} \rho(g_n(\theta), y_0) \geq \delta(\epsilon)c_n$ ,  $\forall \epsilon \in (0, \eta)$ , for  $n$  large; and
- (iii)  $\sup_{\theta \in \Theta} \rho(f_n(\theta), g_n(\theta)) = o_P(c_n)$ ,

then any sequence of estimators  $\hat{\theta}_n$  such that  $\rho(f_n(\hat{\theta}_n), y_0) = o_P(c_n)$  converges in probability to  $\theta_0$ .

**Proof.** By the definition of convergence in probability, we only need to show  $\Pr(d(\hat{\theta}, \theta_0) \geq \epsilon) \rightarrow 0$ , for any  $\epsilon > 0$  small enough. However, for  $\epsilon \in (0, \eta)$ ,

$$\begin{aligned} \Pr(d(\hat{\theta}, \theta_0) \geq \epsilon) &\leq P(c_n^{-1} \rho(g_n(\hat{\theta}), y_0) \geq \delta(\epsilon)) \\ &\leq \Pr(c_n^{-1} \{\rho(g_n(\hat{\theta}), f_n(\hat{\theta})) + \rho(f_n(\hat{\theta}), y_0)\} \geq \delta(\epsilon)) \\ &\leq \Pr(c_n^{-1} \{\sup_{\theta} \rho(g_n(\theta), f_n(\theta)) + \rho(f_n(\theta), y_0)\} \geq \delta(\epsilon)). \end{aligned}$$

By assumption,  $c_n^{-1} \sup_{\theta \in \Theta} \rho(f_n(\theta), g_n(\theta)) \xrightarrow{P} 0$  and  $c_n^{-1} \rho(f_n(\hat{\theta}_n), y_0) \xrightarrow{P} 0$ . Thus the last term in the preceding display converges to 0, which completes the proof.

### References

- Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6**, 701-726.

- Aalen, O. O., Borgan, Ø. and Gjessing, H. K. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer, New York.
- Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.
- Antoniadis, A. (1989). A penalty method for nonparametric estimation of the intensity function of a counting process. *Ann. Inst. Statist. Math.* **41**, 781-807.
- Chen, F., Huggins, R. M., Yip, P. S. F. and Lam, K. F. (2008). Nonparametric estimation of multiplicative counting process intensity functions with an application to the Beijing SARS epidemic. *Comm. Statist. Theory Methods* **37**, 294-306.
- Chen, F., Yip, P. S. F. and Lam, K. F. (2009). On the local polynomial estimation of the counting process intensity function and its derivatives. Submitted.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.
- Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. Wiley, New York.
- Jiang, J. and Doksum, K. A. (2003). On local polynomial estimation of hazard functions and their derivatives under random censoring. In *Mathematical Statistics and Applications: Festschrift for Constance van Eeden* (Edited by M. Moore, S. Froda and C. Léger), IMS Lecture Notes and Monograph Series **42**, 463-482. Institute of Mathematical Statistics, Bethesda, MD.
- Karr, A. F. (1987). Maximum likelihood estimation in the multiplicative intensity model via sieves. *Ann. Statist.* **15**, 473-490.
- Liang, K.-Y. (1984). The asymptotic efficiency of conditional likelihood methods. *Biometrika* **71**, 305-313.
- Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221-239.
- Nielsen, J. P. and Tanggaard, C. (2001). Boundary and bias correction in kernel hazard estimation. *Scand. J. Statist.* **28**, 675-698.
- Patil, P. N. and Wood, A. T. (2004). Counting process intensity estimation by orthogonal wavelet methods. *Bernoulli* **10**, 1-24.
- Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel function. *Ann. Statist.* **11**, 453-466.
- Shao, J. (2003). *Mathematical Statistics*. Second edition. Springer, New York.
- van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press, New York.

Department of Statistics, University of New South Wales, Sydney, Australia.

E-mail: feng.chen@unsw.edu.au

(Received September 2009; accepted March 2010)