Stochastic Processes and their Applications 38 (1991) 157-165
North-Holland

Asymptotic optimality of the least-squares cross-validation bandwidth for kernel estimates of intensity functions

Maria Mori Brooks and J. Stephen Marron*

Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, USA

Received 12 November 1989
Revised 25 July 1990 and 21 August 1990

In this paper, kernel function methods are considered for estimating the intensity function of a non-homogeneous Poisson process. A least-squares cross-validation bandwidth for the kernel intensity estimator is introduced, and it is proven that this bandwidth is asymptotically optimal for kernel intensity estimation.

bandwidth selection * intensity function * cross-validation bandwidth * kernel estimation * nonstationary Poisson processes

1. Introduction

Let $X_1, X_2, \ldots, X_N$ be ordered observations on the interval $[0, T]$ from a non-stationary Poisson process with intensity function $\lambda(x)$. In this paper, we consider the estimation of $\lambda(x)$. N, the number of observations that occur in the interval $[0, T]$, has a Poisson distribution with $E[N]=\int_0^T \lambda(u)\,du$. See Cox and Isham (1980) and Diggle (1983) for further information regarding point processes.

A natural kernel estimator for $\lambda(x)$ is:

$$\hat{\lambda}_h(x)=\sum_{i=1}^{N} K_h(x-X_i), \quad x\in[0, T], \tag{1.1}$$

where $K_h(x)=h^{-1}K(x/h)$. The kernel function, $K(\cdot)$, is assumed here to be a probability density function, and the smoothing parameter, h, also known as the bandwidth, quantifies the smoothness of $\hat{\lambda}_h(x)$ (see, for example, Silverman, 1986). The kernel intensity estimator differs from the typical kernel density estimator in two ways. First, $\hat{\lambda}_h(x)$ doesn't include a normalization factor, n^{-1}, since $\int_s^t \lambda(x)$ is

* Research partially supported by NSF Grant DMS-8902973.

the expected number rather than the expected proportion of observations between s and t. Second, N, the number of observations, is a random variable in the intensity estimation setting. Theoretical properties of the kernel intensity estimator have been developed by Leadbetter and Wold (1983), Ramlau-Hansen (1983) and Diggle (1985).

It is accepted that the choice of the smoothing parameter is usually far more important than the choice of the kernel function for kernel estimators. The tradeoff between smaller and larger values of h can be seen in the mean square error (MSE) of $\hat{\lambda}_h(x)$. Rosenblatt (1971) showed that a small value of h results in high variance; $\hat{\lambda}_h(x)$ is affected by individual observations and hence is more variable. On the other hand, a large value of h results in high bias; $\hat{\lambda}_h(x)$ is very smooth but does not include minor features of the true intensity function. Thus, it is desirable to find a data-based bandwidth that balances the effects of the bias and the variance of the estimate.

We use a simple multiplicative intensity model to put a mathematical structure on this problem. The general multiplicative intensity model, introduced by Aalen (1978), is frequently used to model counting processes. See Anderson and Borgan (1985) for an overview of these models.

For the related setting of kernel density estimation, Hall (1983), Burman (1985), and Stone (1984) have proven that the least-squares cross-validation bandwidth is asymptotically optimal. Hall (1983) showed that this bandwidth is optimal in probability by using asymptotic expansions of the cross-validation score function and the integrated square error (ISE) of the kernel estimator. This method requires the density function to have two continuous bounded derivatives. Burman (1985) used methods from Shibata (1980) to prove that the cross-validation bandwidth is optimal in probability. In his proof, there are few restrictions on the kernel function, and the density is a bounded L^2 function. By assuming that the kernel function is compactly supported and Hölder continuous, Stone (1984) was able to show using Poissonization methods that the cross-validation bandwidth is optimal almost surely for a continuous set of bandwidths. This was the first paper to show that the optimality result applied to all bounded density functions.

In this paper, we show that the least-squares cross-validation method is asymptotically optimal almost surely for intensity estimation. The additional difficulty in the intensity setting is that the number of observations is now a random variable. The intensity function is required to have two continuous bounded derivatives. Moreover, the kernel function is a bounded compactly supported probability density function. We use arguments similar to the martingale methods employed by Härdle, Marron and Wand (1990) to prove the asymptotic optimality of density derivatives. This method is based on a martingale inequality given by Burkholder (1973).

In Section 2, we discuss the mathematical model for the intensity function. Section 3 contains the main result regarding the asymptotic optimality of the least-squares cross-validation bandwidth. Finally, the proof of the theorem is presented in Section 4.

2. The simple multiplicative intensity model

The simple multiplicative intensity model is a specific form of Aalen's (1978) multiplicative intensity model. Suppose that $X_1, X_2, \ldots, X_N$ are observations from a nonhomogeneous Poisson process with intensity

$$\lambda_c(x) = c\alpha(x), \quad x \in [0, T], \tag{2.1}$$

where c is a positive constant, and $\alpha(x)$ is an unknown nonnegative deterministic function with $\int_0^T \alpha(x)\,\mathrm{d}x = 1$. Given N, the occurrence times, $X_1, X_2, \ldots, X_N$, have the same distribution as the order statistics corresponding to N independent random variables with probability density function $\alpha(x)$ on the interval $[0, T]$. The kernel estimate of $\lambda_c(x)$ is given in (1.1). Under this model, N is a Poisson random variable that has expected value equal to c. In order to avoid boundary effects, we assume a circular design such that $\lambda_c(0) = \lambda_c(T)$, $\lambda_c'(0) = \lambda_c'(T)$ and $\lambda_c''(0) = \lambda_c''(T)$.

Asymptotic analysis provides a powerful tool for understanding the behavior of the kernel intensity estimator. Letting $T \to \infty$ is not appropriate since this results in all of the new observations occurring at the right endpoint. In the simple multiplicative intensity model, letting $c \to \infty$ has the desirable effect of adding observations everywhere on the interval $[0, T]$ and not changing the relative shape of the target function $\lambda_c(x)$ in the limiting process. In other words, $c^{-1}\lambda_c(x)$ is a fixed function as $c \to \infty$.

Since we will consider convergence properties for sequences of random variables, we require the values of c to come from the sequence of positive real numbers $\{c_s\}_{s=1}^{\infty}$ such that $c_s/s \to \tau$ for some constant $\tau > 0$ as $s \to \infty$. Thus, for the sequence of intensity functions $\lambda_{c_s}(x) = c_s\alpha(x)$ indexed by s, we construct a corresponding sequence of kernel estimators $\hat{\lambda}_h^s(x)$. For these kernel estimators, the bandwidth h is dependent on s. It follows from Ramlau-Hansen (1983) that $\hat{\lambda}_h^s(x)$ is uniformly consistent and asymptotically normal as $c_s \to \infty$, $h \to 0$ and $hc_s \to \infty$.

3. Asymptotic optimality of the cross-validation bandwidth

We are interested in finding a data based bandwidth that approximately minimizes the integrated square error (ISE) of $\hat{\lambda}_h$ where

$$\begin{aligned}\mathrm{ISE}_\lambda(h) &= \int_0^T [\hat{\lambda}_h(x) - \lambda(x)]^2\,\mathrm{d}x \\ &= \int_0^T \hat{\lambda}_h(x)^2\,\mathrm{d}x - 2\int_0^T \hat{\lambda}_h(x)\lambda(x)\,\mathrm{d}x + \int_0^T \lambda(x)^2\,\mathrm{d}x.\end{aligned} \tag{3.1}$$

For kernel density estimates, Rudemo (1982) and Bowman (1984) suggested using the method of least-squares cross-validation for selecting the bandwidth. In the intensity estimation setting, the cross-validation score function is defined as:

$$\mathrm{CV}_\lambda(h) = \int_0^T \hat{\lambda}_h(x)^2\,\mathrm{d}x - 2\sum_{i=1}^{N} \hat{\lambda}_{hi}(X_i) \tag{3.2}$$

where $\hat{\lambda}_{hi}(x)$ is the leave-one-out estimator,

$$\hat{\lambda}_{hi}(x)=\sum_{\substack{j=1\\ j\neq i}}^{N} K_h(x-X_j). \tag{3.3}$$

Since $\sum_{i=1}^{N} \hat{\lambda}_{hi}(X_i)$ is a method of moments estimator of $\int_0^T \hat{\lambda}_h(x)\lambda(x)\,\mathrm{d}x$, and $\int_0^T \lambda(x)^2\,\mathrm{d}x$ is independent of h, $\mathrm{CV}_\lambda(h)$ is a reasonable unbiased estimate of the terms in $\mathrm{ISE}_\lambda(h)$ that depend on h. Therefore, the bandwidth that minimizes $\mathrm{CV}_\lambda(h)$ should be close to the bandwidth that minimizes $\mathrm{ISE}_\lambda(h)$.

Let $\hat{h}_o$ be any bandwidth that minimizes $\mathrm{ISE}_\lambda(h)$ and $\hat{h}_{cv}$ any bandwidth that minimizes $\mathrm{CV}_\lambda(h)$ (these minima always exist since $\mathrm{ISE}_\lambda(h)$ and $\mathrm{CV}_\lambda(h)$ are continuous and bounded functions). Assume that:

(a) The kernel function, $K(\cdot)$, is a bounded compactly supported probability density function.

(b) The true intensity function, $\lambda(\cdot)$, has two continuous bounded derivatives.

(c) The bandwidths under consideration come from a set H_s where for each s and some constants ρ, $\delta>0$,

$$\#(H_s)=\{\text{the number of elements in } H_s\}\leq c_s^{\rho},$$

and for $h\in H_s$, $c_s^{-1+\delta}\leq h\leq c_s^{-\delta}$.

(d) For some constant $\tau>0$, $c_s/s\to\tau$ as $s\to\infty$.

Assumption (b) is a common technical assumption on $\lambda(x)$ which allows Taylor expansion methods to be used. With this assumption, the true intensity function can have any amount of underlying smoothness. Assumption (c) can be weakened so that H_s is a continuous interval by using a continuity argument found in Härdle and Marron (1985). This set of possible bandwidths nearly covers the range of consistent bandwidths.

Under these assumptions, the ISE obtained with the cross-validation bandwidth converges almost surely to the minimum ISE. In this sense, the least-squares cross-validation bandwidth is asymptotically optimal for kernel intensity estimation under the simple multiplicative intensity model. This result is stated in Theorem 1.

Theorem 1. *If assumptions* (a), (b), (c) *and* (d) *hold, then, under the simple multiplicative intensity model,*

$$\frac{\mathrm{ISE}_\lambda(\hat{h}_{cv})}{\mathrm{ISE}_\lambda(\hat{h}_o)}\to 1 \quad a.s. \quad as\ s\to\infty. \tag{3.4}$$

The mean integrated square error (MISE),

$$\mathrm{MISE}_\lambda(h)=E\left[\int_0^T (\hat{\lambda}_h(x)-\lambda(x))^2\,\mathrm{d}x\right], \tag{3.5}$$

is another error criterion that is used to evaluate bandwidth selection procedures. Let h_o be the bandwidth that minimizes $\mathrm{MISE}_\lambda(h)$. By Lemma 1 in Section 4, the ISE and MISE are essentially the same for large s. As a result of Theorem 1, $\hat{h}_{cv}$ is also asymptotically optimal with respect to MISE in the sense that

$$\frac{\mathrm{MISE}_\lambda(\hat{h}_{\mathrm{cv}})}{\mathrm{MISE}_\lambda(h_{\mathrm{o}})} \to 1 \quad \text{a.s.} \quad \text{as } s \to \infty. \tag{3.6}$$

Details of the proof of Theorem 1 are presented in Section 4.

4. Proof of Theorem 1

In this section, we outline the proof of Theorem 1.

The mean integrated square error (MISE) of $\hat{\lambda}_h$ can be decomposed into a variance term and a squared bias term. Using Taylor expansion methods similar to those in Silverman (1986, pp. 39–40), it is straightforward to show that the MISE of $\hat{\lambda}_h$ is

$$\mathrm{MISE}_\lambda(h) = h^{-1}c_s\left(\int K^2\right) + h^4c_s^2\left[\tfrac{1}{2}\int u^2K\right]^2 \int_0^T [\alpha''(x)]^2 + \mathrm{o}(h^{-1}c_s + h^4c_s^2) \tag{4.1}$$

as $h \to 0$, $c_s \to \infty$ and $hc_s \to \infty$. Thus, the asymptotic mean integrated square error (AMISE) is

$$\mathrm{AMISE}_\lambda(h) = h^{-1}c_s\left(\int K^2\right) + h^4c_s^2\left[\tfrac{1}{2}\int u^2K\right]^2 \int_0^T [\alpha''(x)]^2. \tag{4.2}$$

The two lemmas below are used to prove statement (3.4).

Lemma 1.

$$\sup_{h\in H_s}\left|\frac{\mathrm{ISE}_\lambda(h) - \mathrm{AMISE}_\lambda(h)}{\mathrm{AMISE}_\lambda(h)}\right| \to 0 \quad \textit{a.s.} \quad \textit{as } s \to \infty.$$

Lemma 2.

$$\sup_{h,b\in H_s}\left|\frac{\mathrm{CV}_\lambda(h) - \mathrm{ISE}_\lambda(h) + [\mathrm{CV}_\lambda(b) - \mathrm{ISE}_\lambda(b)]}{\mathrm{AMISE}_\lambda(h) + \mathrm{AMISE}_\lambda(b)}\right| \to 0 \quad \textit{a.s.} \quad \textit{as } s \to \infty.$$

Lemma 1 says that the ISE and the AMISE of $\hat{\lambda}_h(x)$ are asymptotically equivalent, and the two lemmas together imply that

$$\sup_{h,b\in H_s}\left|\frac{\mathrm{CV}_\lambda(h) - \mathrm{ISE}_\lambda(h) - [\mathrm{CV}_\lambda(b) - \mathrm{ISE}_\lambda(b)]}{\mathrm{ISE}_\lambda(h) + \mathrm{ISE}_\lambda(b)}\right| \to 0 \quad \text{a.s.} \quad \text{as } s \to \infty. \tag{4.3}$$

Since $\mathrm{ISE}(\hat{h}_{\mathrm{o}}) \leq \mathrm{ISE}(\hat{h}_{\mathrm{cv}})$ and $\mathrm{CV}(\hat{h}_{\mathrm{cv}}) \leq \mathrm{CV}(\hat{h}_{\mathrm{o}})$, Theorem 1 follows for the simple multiplicative intensity model.

Now, we must prove Lemma 1 and Lemma 2. The details of the proof of Lemma 2 are given below; Lemma 1 is proven using similar martingale methods.

Proof of Lemma 2. Let $g(1, 2, \ldots, N) \to (1, 2, \ldots, N)$ be a random permutation of the numbers 1, 2, …, N. Define $Y_i = X_{g(i)}$. Essentially, the Y_i's are the 'unordered' X_i's. Since the X_i's are observations from a nonhomogeneous Poisson process with intensity $\lambda(x)$, then given N, the Y_i's are i.i.d. random variables with density

$\alpha(x)I_{[0,T]}(x)$. As a result, kernel density methods developed by Härdle, Marron and Wand (1990) can be used to study $\alpha(x)$. Define:

$$\hat{\alpha}_h(x)=c_s^{-1}\hat{\lambda}_h(x),$$
$$\hat{\alpha}_{hi}(x)=c_s^{-1}\hat{\lambda}_{hi}(x),$$
$$\mathrm{CV}_\alpha(h)=\int_0^T \hat{\alpha}_h(x)^2\,\mathrm{d}x-2c_s^{-1}\sum_{i=1}^N \hat{\alpha}_{hi}(X_i)=c_s^{-2}\mathrm{CV}_\lambda(h).$$

Hence, $\mathrm{ISE}_\alpha(h)=c_s^{-2}\,\mathrm{ISE}_\lambda(h)$, and $\mathrm{MISE}_\alpha(h)=c_s^{-2}\,\mathrm{MISE}_\lambda(h)$. It is not difficult to show that

$$\begin{aligned}&\sup_{h,b\in H_s}\left|\frac{\mathrm{CV}_\lambda(h)-\mathrm{ISE}_\lambda(h)-[\mathrm{CV}_\lambda(b)-\mathrm{ISE}_\lambda(b)]}{\mathrm{AMISE}_\lambda(h)+\mathrm{AMISE}_\lambda(b)}\right|\\ &=\sup_{h,b\in H_s}\left|\frac{\mathrm{CV}_\alpha(h)-\mathrm{ISE}_\alpha(h)-[\mathrm{CV}_\alpha(b)-\mathrm{ISE}_\alpha(b)]}{\mathrm{AMISE}_\alpha(h)+\mathrm{AMISE}_\alpha(b)}\right|\\ &\leqslant 2\sup_{h\in H_s}\left|\frac{c_s^{-1}\sum_{i=1}^N\hat{\alpha}_{hi}(X_i)-\int_0^T\hat{\alpha}_h\alpha-G}{\mathrm{AMISE}_\alpha(h)}\right|\end{aligned}$$

where $G=[2(N-1)/c_s-1]c_s^{-1}\sum_{i=1}^N\alpha(X_i)-[N(N-1)/c_s^2]\int_0^T\alpha^2$.

Thus, it suffices to prove the following:

$$\sup_{h\in H_s}\left|\frac{c_s^{-1}\sum_{i=1}^N\hat{\alpha}_{hi}(X_i)-\int_0^T\hat{\alpha}_h\alpha-G}{\mathrm{AMISE}_\alpha(h)}\right|\to 0\quad\text{a.s.}\quad\text{as } s\to\infty. \tag{4.4}$$

Define:

$$U_{ij}(h)=K_h(Y_i-Y_j)-\int K_h(y-Y_j)\alpha(y)\,\mathrm{d}y-\alpha(Y_i)+\int_0^T\alpha^2(y)\,\mathrm{d}y,$$

$$\begin{aligned}V_i(h)&=E(U_{ij}\mid Y_i)\\ &=\int K_h(Y_i-y)\alpha(y)\,\mathrm{d}y-\int\int K_h(y-z)\alpha(y)\alpha(z)\,\mathrm{d}y\,\mathrm{d}z-\alpha(Y_i)\\ &\quad+\int_0^T\alpha^2(y)\,\mathrm{d}y,\end{aligned}$$

$$\begin{aligned}W_{ij}(h)&=U_{ij}-V_i\\ &=K_h(Y_i-Y_j)-\int K_h(y-Y_j)\alpha(y)\,\mathrm{d}y-\int K_h(Y_i-y)\alpha(y)\,\mathrm{d}y\\ &\quad+\int\int K_h(y-z)\alpha(y)\alpha(z)\,\mathrm{d}y\,\mathrm{d}z,\end{aligned}$$

$$\begin{aligned}R_i(h)=[(N-1)/c_s-1]\Big[2\int K_h(Y_i-y)\alpha(y)\,\mathrm{d}y&\\ -\int\int K_h(y-z)\alpha(y)\alpha(z)\,\mathrm{d}y\,\mathrm{d}z&\\ -2\alpha(Y_i)+\int_0^T\alpha^2(y)\,\mathrm{d}y\Big].&\end{aligned}$$

Note that $E(V_i \mid N)=0$ and $E(W_{ij} \mid Y_i, N)=E(W_{ij} \mid Y_j, N)=0$ for $i, j=1, 2, \ldots, N$.

For sums over $i=1, 2, \ldots, N$, Y_i (the unordered observation) can be replaced by X_i (the ordered observation) in the summand. Thus,

$$c_s^{-1} \sum_{i=1}^{N} V_i + c_s^{-2} \sum_{i=1}^{N} \sum_{j \neq i} W_{ij} + c_s^{-1} \sum_{i=1}^{N} R_i = c_s^{-1} \sum_{i=1}^{N} \hat{\alpha}_{hi}(X_i) - \int_0^T \hat{\alpha}_h \alpha - G.$$

Hence, statement (4.4) holds when (4.5), (4.6) and (4.7) hold.

$$\sup_{h \in H_s} \left| \frac{c_s^{-1} \sum_{i=1}^{N} V_i(h)}{\text{AMISE}_\alpha(h)} \right| \to 0 \quad \text{a.s.} \quad \text{as } s \to \infty, \tag{4.5}$$

$$\sup_{h \in H_s} \left| \frac{c_s^{-2} \sum_{i=1}^{N} \sum_{j \neq i} W_{ij}(h)}{\text{AMISE}_\alpha(h)} \right| \to 0 \quad \text{a.s.} \quad \text{as } s \to \infty, \tag{4.6}$$

$$\sup_{h \in H_s} \left| \frac{c_s^{-1} \sum_{i=1}^{N} R_i(h)}{\text{AMISE}_\alpha(h)} \right| \to 0 \quad \text{a.s.} \quad \text{as } s \to \infty. \tag{4.7}$$

Therefore, in order to prove Lemma 2, it is sufficient to prove statements (4.5), (4.6) and (4.7).

Conditional on N, $\{\sum_{i=1}^{k} V_i\}_{k=1}^{N}$ and $\{\sum_{i=1}^{k} \sum_{j=1}^{i-1} W_{ij}\}_{k=1}^{N}$ are martingales with respect to the σ-fields generated by $\{Y_1, Y_2, \ldots, Y_k\}$. Burkholder's (1973, p. 40) inequality implies that for some constant A:

$$E\left[\left(\sup_{k=1,\ldots,N} \sum_{i=1}^{k} V_i\right)^{2m} \middle| N\right] \leq AE\left[\left(\sum_{i=1}^{N} E[V_i^2 \mid N]\right)^m \middle| N\right] + A \sum_{i=1}^{\infty} E[|V_i|^{2m} \mid N], \tag{4.8}$$

$$E\left[\left(\sup_{k=1,\ldots,N} \sum_{i=1}^{k} \sum_{j=1}^{i-1} W_{ij}\right)^{2m} \middle| N\right]$$
$$\leq AE\left[\left(\sum_{i=1}^{N} E\left[\left(\sum_{j=1}^{i-1} W_{ij}\right)^2 \middle| Y_1, Y_2, \ldots, Y_{i-1}, N\right]\right)^m \middle| N\right]$$
$$+ A \sum_{i=1}^{\infty} E\left[\left|\sum_{j=1}^{i-1} W_{ij}\right|^{2m} \middle| N\right]. \tag{4.9}$$

Since N is a Poisson random variable with mean c_s, $E[N^m] \leq A_m c_s^m$ where A_m is constant for each m. Hence,

$$E\left[r_s^{-1} \sum_{i=1}^{N} V_i\right]^{2m} \leq c_s^{-2m} E\left[E\left[\left(\sup_{k=1,\ldots,N} \sum_{i=1}^{k} V_i\right)^{2m} \middle| N\right]\right]$$
$$\leq c_s^{-2m} E[A_m N^m h^{4m} + A_m N] \quad \text{(by (4.8))}$$
$$\leq A_m (c_s^{-m} h^{4m} + c_s^{-2m+1})$$

and

$$E\left[c_s^{-2}\sum_{i=1}^{N}\sum_{j=1}^{i-1}W_{ij}\right]^{2m}\leqslant c_s^{-4m}E[A_mN^{2m}h^{-m}+A_mN^{m+2}h^{-2m}]\quad \text{(by (4.9))}$$

$$\leqslant A_m(c_s^{-2m}h^{-m}+c_s^{-3m+2}h^{-2m}).$$

As a result, for $h\in H_s$, m sufficiently large and some $\gamma>0$,

$$E\left[\frac{c_s^{-1}\sum_{i=1}^N V_i}{\text{AMISE}_\alpha(h)}\right]^{2m}\leqslant\frac{A_m(c_s^{-m}h^{4m}+c_s^{-2m+1})}{(c_s^{-1}h^{-1}+h^4)^{2m}}\leqslant A_m[h^m+c_sh^{2m}]\leqslant A_mc_s^{-\gamma m}. \tag{4.10}$$

Using Chebychev's theorem,

$$\sup_{h\in H_s}P\left[\left|c_s^{-1}\sum_{i=1}^N V_i\right|>c_s^{-\gamma/4}\,\text{AMISE}_\alpha(h)\right]\leqslant A_mc_s^{-(\gamma/2)m}. \tag{4.11}$$

Recall that $\#(H_s)\leqslant c_s^\rho$, and choose m such that $m>2(\rho+2)/\gamma$. By assumption (d), one can show that for the sequence $\{c_s\}$,

$$\sum_{s=1}^{\infty}P\left[\sup_{h\in H_s}\left|\frac{c_s^{-1}\sum_{i=1}^N V_i}{\text{AMISE}_\alpha(h)}\right|>\varepsilon\right]$$

$$\leqslant\sum_{s=1}^{\infty}\#(H_s)\sup_{h\in H_s}P\left[\left|c_s^{-1}\sum_{i=1}^N V_i\right|>\varepsilon\,\text{AMISE}_\alpha(h)\right]$$

$$<\infty.$$

Thus, the Borel-Cantelli Lemma implies that

$$\sup_{h\in H_s}\left|\frac{c_s^{-1}\sum_{i=1}^N V_i}{\text{AMISE}_\alpha(h)}\right|\to 0\quad\text{a.s.}\quad\text{as } s\to\infty. \tag{4.12}$$

This proves (4.5). Following a similar procedure, statement (4.6) can be verified.

Finally, consider (4.7). Since N is Poisson(c_s), it is known that $[(N-c_s)/\sqrt{c_s}]$ is asymptotically normal. Using this fact and a Taylor expansion of R_i, one can show that for any $0<\varepsilon<\frac{1}{2}$,

$$c_s^{1/2}h^{-2+\varepsilon}\left[c_s^{-1}\sum_{i=1}^N R_i\right]\overset{\text{p}}{\to}0.$$

Hence, for each m, there exists a constant c_0 such that whenever $c_s>c_0$,

$$E\left[\frac{c_s^{-1}\sum_{i=1}^N R_i}{\text{AMISE}_\alpha(h)}\right]^{2m}\leqslant\frac{A_mc_s^{-m}h^{4m-2\varepsilon m}}{(c_s^{-1}h^{-1}+h^4)^{2m}}\leqslant A_mh^{(1-2\varepsilon)m}\leqslant A_mc_s^{-\gamma m} \tag{4.13}$$

for $h\in H_s$ and some $\gamma>0$. (4.7) follows from (4.13) as seen above. Therefore, we have proven Lemma 2. □

References

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

P.K. Andersen and O. Borgan, Counting process models for life history data: A review, Scand. J. Statist. 12 (1985) 97-158.

A.W. Bowman, An alternative of cross-validation for the smoothing of density estimates, Biometrika 71 (1984) 353-360.

D.L. Burkholder, Distribution function inequalities for martingales, Ann. Probab. 1 (1973) 19-42.

P. Burman, A data dependent approach to density estimation, Z. Wahrsch. Verw. Gebiete 69 (1985) 609-628.

D.R. Cox, and V. Isham, Point Processes (Chapman and Hall, London, 1980).

L. Devroye and L. Györfi, Nonparametric Density Estimation: The L_1 View (Wiley, New York, 1984).

P.J. Diggle, Statistical Analysis of Spatial Point Patterns (Academic Press, London, 1983).

P. Diggle, A kernel method for smoothing point process data, Appl. Statist. 34 (1985) 138-147.

W. Härdle and J.S. Marron, Optimal bandwidth selection in nonparametric regression estimation, Ann. Statist. 13 (1985) 1465-1481.

W. Härdle, J.S. Marron and M.P. Wand, Bandwidth choice for density derivatives, J. Roy. Statist. Soc. Ser. B 52 (1990) 223-232.

P. Hall, Large sample optimality of least squares cross-validation in density estimation, Ann. Statist. 11 (1983) 1156-1174.

M.R. Leadbetter and D. Wold, On estimation of point process intensities, in: P.K. Sen, ed., Contributions to Statistics: Essays in Honour of Norman L. Johnson (North-Holland, Amsterdam, 1983) pp. 299-312.

J.S. Marron and W. Härdle, Random approximations to some measures of accuracy in nonparametric curve estimation, J. Multivariate Anal. 20 (1986) 91-113.

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

M. Rosenblatt, Curve estimates, Ann. Math. Statist. 42 (1971) 1815-1842.

M. Rudemo, Empirical choice of histograms and kernel density estimators, Scand. J. Statist. 9 (1982) 65-78.

R. Shibata, Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, Ann. Statist. 8 (1980) 147-164.

B.W. Silverman, Density estimation for statistics and data analysis (Chapman and Hall, New York, 1986).

C. Stone, An asymptotically optimal window selection rule for kernel density estimates, Ann. Statist. 12 (1984) 1285-1297.