

BANDWIDTH SELECTION METHODS FOR KERNEL ESTIMATORS  
OF THE INTENSITY FUNCTION  
OF A NONHOMOGENEOUS POISSON PROCESS

by


Maria Mori Brooks


A Dissertation submitted to the faculty of  
The University of North Carolina at Chapel Hill  
in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy in the Department of Statistics

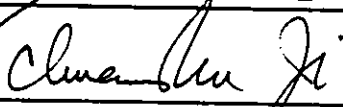
Chapel Hill

1991

Approved by:

  
\_\_\_\_\_  
Advisor

  
\_\_\_\_\_  
Reader

  
\_\_\_\_\_  
Reader

## ABSTRACT

Kernel function methods are appealing tools for estimating the intensity function of a nonhomogeneous Poisson process. The bandwidth, which quantifies the smoothness of the kernel estimator, dramatically affects the accuracy of the estimator. Therefore, one would like to find a data based bandwidth selection method that performs well for kernel intensity estimation in some sense. We consider two models for putting a mathematical structure on the intensity function of the point process: a simple multiplicative intensity model and a stationary Cox process model.

Asymptotic analysis allows one to see the basic structure of an estimator. In this dissertation, we show that the least-squares cross-validation bandwidth is asymptotically optimal for intensity estimation under the simple multiplicative intensity model. We also show that given the stationary Cox process model, this bandwidth is not asymptotically optimal. Once asymptotic optimality of an estimator has been established, it is important to find the rate of convergence for that estimator. We determine the convergence rate for the cross-validation bandwidth by finding the asymptotic distribution of this bandwidth under the simple multiplicative intensity model. Finally, we include an example with coffee purchase data; this example demonstrates how kernel estimation and the cross-validation bandwidth perform on a real data set.

## PREFACE

I would like to express my great appreciation to Steve Marron, my dissertation advisor. His guidance, many valuable insights and continuous encouragement have helped me immensely throughout the past few years. In addition, the entire dissertation committee has provided useful comments and advice regarding my research.

With heartfelt gratitude, I thank my parents and sister for their constant support and thoughtfulness.

Finally, I would like to thank my husband Rob for his big bright smiles and his love - they allow me to find happiness in statistics and in life.

## TABLE OF CONTENTS

	Page
LIST OF FIGURES.....	v
 <b>Chapter</b>	
I. LITERATURE REVIEW AND INTRODUCTION.....	1
1.a. Kernel Density Estimation Literature Review.....	2
1.b. Kernel Intensity Estimation Literature Review.....	9
1.c. Models for Intensity Estimation.....	11
1.d. Results for the Least-Squares Cross-Validation Bandwidth.....	14
II. ASYMPTOTIC OPTIMALITY OF THE LEAST-SQUARES CROSS-VALIDATION BANDWIDTH UNDER THE SIMPLE MULTIPLICATIVE INTENSITY MODEL.....	18
2.a. The Model and the Kernel Estimator.....	19
2.b. Results.....	23
2.c. Proofs.....	26
III. THE LEAST-SQUARES CROSS-VALIDATION BANDWIDTH UNDER THE STATIONARY COX PROCESS MODEL.....	45
3.a. The Model.....	45
3.b. Differences Between the Stationary Cox Process Model and the Simple Multiplicative Intensity Model.....	47
3.c. Results.....	51
3.d. Proofs.....	52
IV. THE ASYMPTOTIC DISTRIBUTION OF THE LEAST-SQUARES CROSS- VALIDATION BANDWIDTH UNDER THE SIMPLE MULTIPLICATIVE INTENSITY MODEL.....	56
4.a. Assumptions and Notation.....	57
4.b. Results and Discussion.....	59
4.c. Proofs of the Theorems.....	66
4.d. The Lemmas and Their Proofs.....	69
V. AN EXAMPLE OF KERNEL INTENSITY ESTIMATION WITH THE LEAST- SQUARES CROSS-VALIDATION BANDWIDTH APPLIED TO COFFEE SALES DATA.....	99
5.a. Kernel Estimation of the Coffee Data.....	99
5.b. The Least-Squares Cross-Validation Bandwidth.....	107
REFERENCES.....	116

## LIST OF FIGURES

Figure 4.1:	Intensity Function #1.....	65
Figure 4.2:	Intensity Function #2.....	65
Figure 5.1:	Brand 4, $h=2.82$ .....	103
Figure 5.2:	Brand 4, $h=.75$ .....	103
Figure 5.3:	Brand 4, $h=10.57$ .....	104
Figure 5.4:	Brand 4 Sales and Promotions.....	106
Figure 5.5:	Brand 4 Cross-Validation Curve.....	109
Figure 5.6:	Brand 29 Cross-Validation Curve.....	109
Figure 5.7:	Brand 29, $h=4.08$ .....	111
Figure 5.8:	Brand 29, $h=31.13$ .....	111
Figure 5.9:	Brand 20 Cross-Validation Curve.....	113
Figure 5.10:	Brand 20, $h=55.56$ .....	113
Figure 5.11:	Brand 20, $h=4.00$ .....	114

## CHAPTER 1

### LITERATURE REVIEW AND INTRODUCTION

The goal of this dissertation is to evaluate the least-squares cross-validation bandwidth for the kernel estimator of the intensity function of a nonstationary Poisson process. This chapter contains a literature review for kernel estimation and bandwidth selection methods. The literature related to kernel intensity estimation comes from two fields of research. First of all, many of the ideas and methods developed for kernel density estimation can be applied to the intensity estimation context. Secondly, the work in intensity estimation provides several models for the intensity function and a variety of methods for studying this function. Therefore, the literature review is divided into two parts; section 1.a. covers kernel density estimation methods, and section 1.b reviews intensity estimation methods.

The last two sections of this chapter deal with the work contained in the following chapters. In section 1.c, we discuss the mathematical frameworks that we will assume for the point processes. Finally, section 1.d includes a summary and a brief discussion of the main results found in this dissertation.

1.a. Kernel Density Estimation Literature Review.

In the density setting, assume that  $X_1, X_2, \dots, X_n$  are independent identically distributed observations from an unknown probability density  $f(x)$ . A delta sequence density estimator is a general type of estimate of  $f(x)$  which is written in the form

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^n \delta_h(x, X_i)$$

where  $h \in \mathbb{R}^+$  is the smoothing parameter. The general formulation of this nonparametric estimator was introduced by Watson and Leadbetter (1963) and then extended by Foldes and Revesz (1974). By changing the definition of  $\delta_h(x, X_i)$ , delta estimators include orthogonal series estimators, spline estimators, histogram estimators and kernel estimators. See Walter and Blum (1979) for examples of various delta estimators.

The kernel density estimator is a specific type of delta sequence density estimator. It is defined as

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i) \quad (1.1)$$

where  $K_h(x) = h^{-1} K(x/h)$ .  $K(\cdot)$ , known as the kernel function, is the "shape" of the weight that is placed on each data point;  $h$ , called the bandwidth or the smoothing parameter, is the quantity that controls the amount of smoothing done by  $\hat{f}_h(x)$ . In general, the bandwidth depends on the sample size such that as  $n$  increases,  $h \rightarrow 0$  and  $nh \rightarrow \infty$ . The kernel density estimator takes on large values where the data are dense and small values where the data are sparse. Due to its intuitive appeal as well as its mathematical tractability, the kernel density estimator has been widely studied over the past thirty years.

Early work in kernel density estimation was done by Rosenblatt (1956), Whittle (1958), Parzen (1962), and Watson and Leadbetter (1963). Rosenblatt (1956) introduced the kernel density estimator with a uniform kernel. He also found expressions for the mean squared error (MSE) and the mean integrated square error (MISE) of the first kernel estimator. Whittle (1958) considered the problem of density estimation with a random number of observations on a finite interval. Among other things, he calculated the MISE for the kernel density estimator in this more complicated setting.

Parzen (1962) generalized Rosenblatt's kernel estimator by allowing  $K(\cdot)$  to be any function that satisfies relatively nonrestrictive conditions. When  $K(\cdot)$  satisfies the specified conditions, he showed that  $\hat{f}_h(x)$  is a consistent estimate of  $f(x)$  and that  $\hat{f}_h(x)$  is asymptotically normal. Watson and Leadbetter (1963) presented further consistency results for the kernel estimator. More recently, Castellana and Leadbetter (1986) proved that kernel estimators for the marginal probability density function of a stationary sequence or a continuous parameter process are also consistent and asymptotically normal.

In order to fully determine the kernel density estimator, the kernel function and the bandwidth must be specified. Epanechnikov (1969) found the kernel function,  $K(\cdot)$ , that minimizes the asymptotic MSE among density functions with unit variance. Parzen (1962) and Watson and Leadbetter (1963) first suggested that the kernel estimator could be improved asymptotically by using higher order kernels, that is, the kernel function can assume negative values. Later, Schucanny

and Sommers (1977) and Casser, Mueller and Mammitzsch (1985) used simulation studies to verify that higher order kernels are beneficial for finite data sets, and Hall and Marron (1987b) proposed a method of choosing the kernel order. However, it is believed that the choice of the smoothing parameter is far more important than the choice of the kernel function; see, for example, Silverman (1986, pp. 15-17) for a discussion and examples illustrating this point. Therefore, for convenience, the kernel function is often chosen to be a symmetric probability density function.

An important goal in kernel estimation is to find the bandwidth that minimizes the "distance" between the estimate  $\hat{f}_h(x)$  and the true density  $f(x)$ . A criterion that is often used to measure this error is the mean integrated square error where

$$\text{MISE}(h) = E[\int(\hat{f}_h(x) - f(x))^2 dx] . \quad (1.2)$$

Rosenblatt (1971) presented an intuitive description for the tradeoff between smaller and larger values of the bandwidth. The MSE of  $\hat{f}_h(x)$  is the sum of the squared bias and the variance of  $\hat{f}_h(x)$ . Rosenblatt showed that a small value of  $h$  results in high variance (i.e.  $\hat{f}_h(x)$  is affected by individual observations and hence is more variable), while a large value of  $h$  results in high bias (i.e.  $\hat{f}_h(x)$  is very smooth but may not include minor features of the true distribution). Thus, it is desirable to find a bandwidth that balances the effects of the bias and the variance of the estimate.

A number of methods have been proposed for choosing  $h$ . The "plug-in" method, introduced by Woodroffe (1970) and Scott, Tapia and Thompson (1977), uses the concept that

$$h_o \sim \left[ \frac{C(K) f(x)}{f''(x)^2} \right]^{1/5} n^{-1/5} \quad (1.3)$$

where  $h_o$  is the bandwidth that minimizes MSE. Therefore, they estimate  $f$  and  $f''$  with kernel estimators, and then substitute these estimates into equation (1.3) to obtain a value for  $h$ . With this procedure, a smoothing parameter still must be chosen in order to estimate  $f(x)$  and  $f''(x)$ ; in particular,  $f''(x)$  is even more difficult to estimate than  $f(x)$ . Moreover, equation (1.3) is an asymptotic equation, and thus it may not give a good solution for  $h$  when  $n$  is finite.

The Kullback-Leibler or pseudo-likelihood method was introduced by Habbema, Hermans and van den Broek (1974) and Duin (1976), and later improved by Hall (1987a). The motivation for this method is that a function which is similar to the likelihood function is maximized over  $h$ , and consequently, the solution is analogous to the maximum likelihood solution. The Kullback-Leibler method maximizes the function

$$KL(h) = \prod_{i=1}^n \hat{f}_{hi}(X_i) \quad (1.4)$$

where

$$\hat{f}_{hi}(x) = (n-1)^{-1} \sum_{\substack{j=1 \\ j \neq i}}^n K_h(x-X_j) \quad (1.5)$$

$\hat{f}_{hi}(\cdot)$  is known as the leave-one-out estimator. If  $\hat{f}_h(x)$ , as in (1.1), were used in place of  $\hat{f}_{hi}(x)$  then the trivial solution  $h=0$  would maximize equation (1.4).

Rudemo (1982) and Bowman (1984) suggested using the method of least-squares cross-validation for selecting the bandwidth. Their goal was to find a bandwidth that minimizes the integrated square error

(ISE) of  $\hat{f}_h(x)$  where

$$\begin{aligned} \text{ISE}(h) &= \int [\hat{f}_h(x) - f(x)]^2 dx \\ &= \int \hat{f}_h(x)^2 dx - 2 \int \hat{f}_h(x) f(x) dx + \int f(x)^2 dx \end{aligned} \quad (1.6)$$

Rudemo (1982) and Bowman (1984) proposed choosing  $h$  to be the minimizer of the cross validation score:

$$\text{CV}(h) = \int \hat{f}_h(x)^2 dx - 2n^{-1} \sum_{i=1}^n \hat{f}_{hi}(X_i) \quad (1.7)$$

where  $\hat{f}_{hi}(x)$  is the leave-one-out estimator as in (1.5). Since  $n^{-1} \sum_{i=1}^n \hat{f}_{hi}(X_i)$  is a method of moments estimator of  $\int \hat{f}_h f$ , and  $\int f(x)^2 dx$  is independent of  $h$ ,  $\text{CV}(h)$  is a reasonable estimate of the terms in the ISE that depend on  $h$ . Therefore, the bandwidth that minimizes  $\text{CV}(h)$  should be close to the bandwidth that minimizes the ISE of  $\hat{f}_h(x)$ .

The optimal rate of convergence for  $\hat{f}_h(x)$  was first considered by Farrell (1972) and Stone (1980) in order to evaluate different bandwidth selection procedures. Let  $C$  be a given set of density functions with two existing derivatives, then Farrell (1972) showed that for  $f(x) \in C$ ,  $r=2/5$  is the smallest value of  $r$  such that

$$\lim_{a \rightarrow \infty} \liminf_{n \rightarrow \infty} \inf_{f \in C} P\{|\hat{f}_h(x) - f(x)| \leq an^{-r}\} = 1.$$

In other words,  $n^{-2/5}$  is the best possible rate of convergence for  $|\hat{f}_h(x) - f(x)|$ . Furthermore, he showed that when  $p$  is the number of existing derivatives for the density  $f(\cdot)$ ,  $n^{-p/(2p+1)}$  is the optimal rate of convergence for  $|\hat{f}_h(x) - f(x)|$ . For a clear presentation of this result as well as other optimal rate results, see Stone (1980).

Asymptotic rates of convergence were also used to study various bandwidths and their error criteria, ISE or MISE. Let:

$h_0$  be the bandwidth that minimizes  $\text{MISE}(h)$ ,

$\hat{h}_0$  be the bandwidth that minimizes  $ISE(h)$ .

$\hat{h}_{cv}$  be the bandwidth that minimizes  $CV(h)$ .

$\hat{h}_{kl}$  be the bandwidth that minimizes  $KL(h)$ .

Under the assumption that  $f(x)$  has a continuous second derivative, Hall (1983) proved that  $\hat{h}_{cv}$  is asymptotically equivalent to  $\hat{h}_0$  in the sense that

$$\hat{h}_{cv} / \hat{h}_0 \rightarrow 1 \text{ in probability and}$$
$$\frac{ISE(\hat{h}_{cv})}{ISE(\hat{h}_0)} \rightarrow 1 \text{ in probability.}$$

Stone (1984) and Burman (1985) showed that under the weaker assumption that  $f(x)$  is continuous

$$\frac{ISE(\hat{h}_{cv})}{ISE(\hat{h}_0)} \rightarrow 1 \text{ a.s. and}$$
$$\frac{MISE(\hat{h}_{cv})}{MISE(\hat{h}_0)} \rightarrow 1 \text{ a.s.}$$

Therefore, the ISE and MISE obtained with the cross-validation bandwidth converge to the minimum ISE and MISE respectively. The results of Hall (1983), Stone (1984) and Burman (1985) imply that the least-squares cross-validation method is asymptotically optimal for choosing  $h$  to estimate a density with any amount of underlying smoothness. Marron and Härdle (1986) showed that ISE and MISE are essentially the same for large  $n$  by proving that under reasonable assumptions

$$\limsup_{n \rightarrow \infty} \sup_{h \in H_n} \left| \frac{ISE(h) - MISE(h)}{MISE(h)} \right| = 0 \text{ a.s.}$$

where  $H_n$  is a finite set that depends on  $n$ . It has been proven that if

$\hat{h}$  is either  $\hat{h}_{cv}$  or  $\hat{h}_{kl}$  then

$$\lim_{n \rightarrow \infty} \frac{D(\hat{h})}{\inf_{h \in H_n} D(h)} = 1 \quad \text{a.s.}$$

where  $D$  is either ISE or MISE; see for example Marron (1987). However, Hall (1987a) and Marron (1987) concluded that the least-squares cross-validation method has several advantages over the Kullback-Leibler method for choosing  $h$ . These advantages include weaker assumptions needed and fewer sources of error for the cross-validation method.

The rate of convergence for the least-squares cross-validation bandwidth was introduced by Hall and Marron (1987). Since  $h_0$  depends on the unknown density function  $f(\cdot)$  while  $\hat{h}_0$  is a function of the sample observations, they suggested that one should aim to minimize ISE. Hall and Marron (1987a) proved that  $\hat{h}_{cv}$  performs as well as  $h_0$ , an unattainable value, when  $\hat{h}_0$  is the standard of comparison. This is seen from the following results:

$$n^{3/10} (\hat{h}_0 - h_0) \xrightarrow{D} N(0, a_3^{-2} \sigma_0^2),$$

$$n^{3/10} (\hat{h}_{cv} - \hat{h}_0) \xrightarrow{D} N(0, a_3^{-2} \sigma_{cv}^2),$$

$$n [ISE(h_0) - ISE(\hat{h}_0)] \xrightarrow{D} (a_3^{-1} \sigma_0^2 / 2) \chi_1^2,$$

$$n [ISE(\hat{h}_{cv}) - ISE(\hat{h}_0)] \xrightarrow{D} (a_3^{-1} \sigma_{cv}^2 / 2) \chi_1^2$$

where  $a_3, \sigma_0^2, \sigma_{cv}^2$  are constants depending on  $f(\cdot)$  and  $K(\cdot)$ .

Finally, boundary adjustments are very important for estimating the intensity function near the endpoints of the interval. Results relating to boundary adjustments in the density estimation context can be found in Rice (1984), Schuster (1985), Gasser, Mueller, and Mammitzsch (1985), and Cline and Hart (1986).

### 1.b. Kernel Intensity Estimation Literature Review.

Let  $X_1, X_2, \dots, X_N$  be all of the observations on the interval  $[0, T]$  from a point process with intensity function  $\lambda(x)$ . That is, the number of occurrences in  $(0, t]$  has expected value equal to  $\int_0^t \lambda(u) du$ . See Cox and Isham (1980), Ripley (1981), and Diggle (1983) for further information regarding point processes. The intensity function is very similar to a density function in that we expect to have more observations when the intensity function takes on high values and have fewer observations when the intensity function takes on low values. Hence, a natural estimate for  $\lambda(x)$  is the kernel estimator:

$$\hat{\lambda}_h(x) = \frac{1}{N} \sum_{i=1}^N K_h(x - X_i) \quad \text{for } x \in [0, T] \quad (1.8)$$

where  $K_h(x) = h^{-1} K(x/h)$ . Again, the kernel function,  $K(\cdot)$ , is generally a symmetric probability density function, and the bandwidth,  $h$ , is the smoothing parameter for  $\hat{\lambda}_h(x)$ . In the intensity estimation setting,  $N$ , the number of observations in  $[0, T]$ , is a random variable. Note that the kernel density estimate,  $\hat{f}_h(x)$ , includes a normalization factor of  $n^{-1}$  which makes  $\hat{f}_h(x)$  a probability density function; this adjustment is not needed for estimating an intensity function.

Theoretical properties of the kernel intensity estimator have been developed by Devroye and Györfi (1984), Leadbetter and Wold (1983), and Ellis (1986). In particular, Leadbetter and Wold (1983) proved that  $\hat{\lambda}_h(x)$  is almost sure pointwise consistent, almost sure uniformly consistent and asymptotically normal under specified conditions.

A variety of approaches have been taken to study kernel intensity estimation. The multiplicative intensity model, introduced by Aalen (1978), is frequently used to model counting processes. This model

assumes that the intensity function takes the form

$$\lambda(x) = \alpha(x) Y(x) \quad \text{for } x \in [0, T]$$

where  $\alpha(x)$  is an unknown non-negative deterministic function, and  $Y(x)$  is a nonnegative observable stochastic process such that  $Y(x)$  is known just prior to time  $x$ . Often  $\alpha(x)$  can be interpreted as an individual intensity for making an occurrence, and  $Y(x)$  as the number of individuals that are at risk at time  $x$ . In order to estimate  $\alpha(x)$ , Ramlau-Hansen (1983) proposed the following estimator:

$$\hat{\alpha}(x) = \frac{\sum_{i=1}^n K_h(x - X_i)}{Y(X_i)}$$

He proved that this estimator is uniformly consistent and asymptotically normal as  $n \rightarrow \infty$ ,  $h \rightarrow 0$  and  $nh \rightarrow \infty$ . See Anderson and Borgan (1985) for an overview of the multiplicative intensity model and the above kernel estimator  $\hat{\alpha}(x)$ .

Diggle (1985) considered the estimation of the intensity function,  $\lambda(x)$ , assuming that the underlying point process is a stationary Cox process (also known as a doubly stochastic Poisson process). This process is defined by :

- 1)  $\{\Lambda(x), x \in \mathbb{R}\}$  is a stationary, non-negative valued random process
- 2) conditional on the realization  $\lambda(x)$  of  $\Lambda(x)$ , the point process is a nonhomogeneous Poisson process with rate function  $\lambda(x)$ .

See Cox and Isham (1980) for a discussion of doubly stochastic Poisson processes. Diggle (1985) recommended the standard kernel intensity estimator,  $\hat{\lambda}_h(x)$ , and used an empirical Bayes method to calculate its  $MSE(h) = E[\hat{\lambda}_h(x) - \lambda(x)]^2$ . For  $x$  sufficiently far from the boundaries 0 and  $T$ , the MSE does not depend on  $x$  since the process  $\Lambda(x)$  is stationary. In order to estimate the bandwidth, Diggle (1985) proposed

substituting estimates for the unknown terms in the MSE, and then minimizing this function over  $h$ . In an earlier paper, Clevenson and Zidek (1977) developed similar methods to study  $\hat{\lambda}_h(x)$  with a uniform kernel in the stationary Cox process setting.

When  $\hat{\lambda}_h(x)$  is used with a uniform kernel, Diggle and Marron (1988) proved that Diggle's (1985) minimum MSE method and the least-squares cross-validation bandwidth selection method choose the same bandwidth. These two methods are motivated from very different viewpoints, and the fact that they both select the same bandwidth confirms that this bandwidth is a reasonable choice. More importantly, this result also indicates how asymptotic methods that were developed for density estimation can be used in the intensity setting. Finally, Diggle (1985) and Diggle and Marron (1988) recommend procedures that provide boundary adjustments for the kernel intensity estimator.

#### *1.c. Models for Intensity Estimation.*

Two mathematical models for the intensity function are considered in this thesis: they are a simple multiplicative intensity model and a stationary Cox process model. Intensity estimation under the assumptions of these models are discussed in this section.

First, consider a very simple version of the multiplicative intensity model. Let  $X_1, X_2, \dots, X_N$  be the observations from a nonhomogeneous Poisson process on the interval  $[0, T]$  with intensity  $\lambda_c(x)$ . Assume that

$$\lambda_c(x) = c \alpha(x)$$

where  $c$  is a constant, and  $\alpha(x)$  is a nonnegative deterministic function

such that  $\int_0^T \alpha(x) dx = 1$ . Moreover, the kernel estimate of  $\lambda_c$  is

$$\hat{\lambda}_h(x) = \frac{N(c)}{\sum_{i=1}^N K_h(x-X_i)} = \frac{N(c)}{\sum_{i=1}^N h^{-1} K[(x-X_i)/h]} \text{ for } x \in [0, T].$$

$K(\cdot)$  is the kernel function which we assume is a symmetric probability density function, and  $h$  is the bandwidth. Under this model,  $N$ , the number of observations that occur in  $[0, T]$ , is a Poisson random variable with mean  $\int_0^T \lambda_c(x) dx = c$ . Note that  $\alpha_c(x) = \lambda_c(x) [\int_0^T \lambda_c(u) du]^{-1}$ ; thus, conditional on  $N$ , the observations  $X_1, X_2, \dots, X_N$  have the same distribution as the order statistics of  $N$  independent random variables with probability density  $\alpha_c(x) I_{[0, T]}(x)$ . As a result, previously developed kernel density methods can be used to study  $\alpha(x) I_{[0, T]}(x)$ . Throughout the dissertation, we will assume a circular design where  $\lambda_c(0) = \lambda_c(T)$ ,  $\lambda_c'(0) = \lambda_c'(T)$  and  $\lambda_c''(0) = \lambda_c''(T)$ . This allows us to ignore the boundary effects at 0 and  $T$ .

Asymptotic analysis is a powerful tool for understanding the behavior of an estimator. In the density setting, asymptotic analysis with  $n \rightarrow \infty$  allows the researcher to evaluate the kernel estimator. In this intensity model, letting  $c \rightarrow \infty$  has the same 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  $c \rightarrow \infty$ . In this thesis, we will be considering convergence of sequences of random variables; therefore, we will construct an appropriate sequence  $\{c_s\}_{s=1}^{\infty}$  of values for  $c$  indexed by  $s$ . Throughout, we will assume that  $c_s \rightarrow \infty$ ,  $h \rightarrow 0$  and  $c_s h \rightarrow \infty$ .

Next, we look at the stationary Cox process model. Let  $X_1, X_2, \dots, X_N$  be all of the observations from a stationary Cox process or a

doubly stochastic Poisson process with intensity function  $\lambda_\mu(x)$  on the interval  $[0, T]$ . Assume that  $\lambda_\mu(x)$  is a realization of the stationary nonnegative random process  $\{\Lambda(x), x \in \mathbb{R}\}$ . Moreover, for each  $\mu$ , assume that for all  $x, y \in [0, T]$ :

$$1) E[\Lambda(x)] = \mu.$$

$$2) E[\Lambda(x)\Lambda(y)] = v(|x-y|) \quad \text{where } v(x) = \mu^2 v_0(x)$$

for  $v_0(x)$  a fixed function.

This model is identical to the Cox model used in Diggle (1985).

In this intensity model, asymptotic analysis is done by letting  $\mu \rightarrow \infty$ . Consequently, assumption 2) above ensures that the relative shape of the curve  $\Lambda(x)$  in the limiting process does not change. Again, we construct a sequence  $\{\mu_s\}_{s=1}^\infty$ , so that we can discuss convergence properties for sequences of random variables.

The kernel estimate of  $\lambda_\mu(x)$  is

$$\hat{\lambda}_h(x) = \frac{N(\mu)}{\sum_{i=1}^{N(\mu)} K_h(x-X_i)} \quad \text{for } x \in [0, T].$$

$\hat{\lambda}_h(x)$  in the Cox process model is the same as  $\hat{\lambda}_h(x)$  in the

multiplicative intensity model for estimating  $\lambda(x)$  from a data set.

The difference between the two models is only seen when the estimators are evaluated mathematically. Note that under the Cox process model,

$E[N] = E[\int_0^T \Lambda(x) dx] = \mu T$  where  $N$  is the number of observations that occur in

$[0, T]$ . Moreover, if we condition on  $\{\Lambda(x), x \in [0, T]\}$ , then  $N$  is a

Poisson random variable with mean  $(\int_0^T \Lambda)$ ; however,  $N$  is not in general a

Poisson random variable. Finally, define  $\alpha(x) = \Lambda(x) [\int_0^T \Lambda(u) du]^{-1}$ ; then,

conditional on  $N$ , the  $N$  occurrence times have the same distribution as

the order statistics from  $N$  independent identically distributed random variables with density  $\alpha(x)I_{[0,T]}(x)$ .

*1.d. Results for the Least-Squares Cross-Validation Bandwidth.*

An important consideration in kernel intensity estimation is the selection of the bandwidth or smoothing parameter. Our aim is to find a data based bandwidth selection method which has favorable asymptotic properties. For both of the above models, we are interested in finding a bandwidth that minimizes the distance between the kernel estimate  $\hat{\lambda}_h$  and the true intensity function  $\lambda$ . Therefore, we aim to minimize the integrated square error (ISE) of  $\hat{\lambda}_h$  where

$$ISE_{\lambda}(h) = \int_0^T \hat{\lambda}_h^2 - 2 \int_0^T \hat{\lambda}_h \lambda + \int_0^T \lambda^2 .$$

The first term,  $\int_0^T \hat{\lambda}_h^2$ , is a function of the data, and the third term,  $\int_0^T \lambda^2$ , is independent of  $h$ . In addition, let  $\hat{\lambda}_{hi}(x)$  be the leave-one-out estimator,

$$\hat{\lambda}_{hi}(x) = \frac{N}{\sum_{\substack{j=1 \\ j \neq i}}^N K_h(x-X_j)} \quad \text{for } x \in [0, T].$$

Then,  $\sum_{i=1}^N \hat{\lambda}_{hi}(X_i)$  is a method of moments estimator for  $\int \hat{\lambda}_h \lambda$ . As a result, we can adapt Rudemo's (1982) cross-validation score function to the intensity estimation setting by defining

$$CV_{\lambda}(h) = \int_0^T \hat{\lambda}_h^2 - 2 \sum_{i=1}^N \hat{\lambda}_{hi}(X_i) .$$

The cross-validation score function for  $\hat{\lambda}$ ,  $CV_{\lambda}(h)$ , provides a good estimate of the first two terms of  $ISE_{\lambda}(h)$ , and thus, the bandwidth that minimizes  $CV_{\lambda}(h)$  should be close to the bandwidth that minimizes  $ISE_{\lambda}(h)$ .

Again let  $\hat{h}_{cv}$  be the bandwidth that minimizes  $CV_\lambda(h)$  and  $\hat{h}_o$  the bandwidth that minimizes  $ISE_\lambda(h)$ . The goal in this thesis is to minimize the ISE; hence, there are two ways that a bandwidth  $h^*$  can be optimal. First of all,  $ISE_\lambda(h^*)$ , the ISE of the kernel estimator with bandwidth  $h^*$ , can be asymptotically equivalent to the minimum ISE. Secondly, the bandwidth  $h^*$  can be asymptotically equivalent to the bandwidth that minimizes the ISE.

It is known that the cross-validation bandwidth is asymptotically optimal for density estimation, and we show that this is also the case in the context of intensity estimation. In Chapter 2, given the simple multiplicative intensity model and several reasonable assumptions we show that  $\hat{h}_{cv}$  is asymptotically optimal in the sense that

$$\frac{ISE_\lambda(\hat{h}_{cv})}{ISE_\lambda(\hat{h}_o)} \rightarrow 1 \text{ a.s. as } c_s \rightarrow \infty, \quad (1.9)$$

$$\text{and } \hat{h}_{cv} / \hat{h}_o \xrightarrow{P} 1 \text{ as } c_s \rightarrow \infty. \quad (1.10)$$

We prove (1.9) with martingale techniques similar to those used in Hardle, Marron and Wand (1990) to study the asymptotics of density derivatives. In order to prove (1.10), we use an argument that is based on the kernel density estimation proof found in Hall (1983).

Unfortunately, it turns out that the cross-validation bandwidth is not asymptotically optimal under the Cox process model. This follows from the fact that the ISE and the MISE are not asymptotically equivalent in this setting. In Chapter 3, we discuss this more thoroughly, and we give restrictive conditions under which the cross-validation bandwidth is asymptotically optimal.

Once we determine that the least-squares cross-validation bandwidth is asymptotically optimal, the next step is to find the rate of convergence. Using ideas found in Hall and Marron (1987a), we show that under the simple multiplicative intensity model and mild assumptions,

$$\begin{aligned} c_s^{1/10} (\hat{h}_o - h_o)/h_o &\xrightarrow{D} N(0, \sigma_o^2), \\ c_s^{1/10} (\hat{h}_{cv} - \hat{h}_o)/h_o &\xrightarrow{D} N(0, \sigma_{cv}^2), \\ c_s^{-1} [ISE_\lambda(h_o) - ISE_\lambda(\hat{h}_o)] &\xrightarrow{D} \sigma_o^2 a_4 x_1^2, \\ c_s^{-1} [ISE_\lambda(\hat{h}_{cv}) - ISE_\lambda(\hat{h}_o)] &\xrightarrow{D} \sigma_{cv}^2 a_4 x_1^2 \end{aligned}$$

where  $a_4$ ,  $\sigma_o^2$ ,  $\sigma_{cv}^2$  are constants depending on  $\lambda(\cdot)$  and  $K(\cdot)$  and  $h_o$  is the bandwidth that minimizes the MISE of  $\hat{\lambda}_h(x)$ . These results imply that when  $\hat{h}_o$  is used as the basis for comparison, the cross-validation bandwidth performs as well as the bandwidth that minimizes the MISE in terms of the convergence rate. Thus, the convergence rate,  $c_s^{-1/10}$ , is slow, but we cannot reasonably expect it to be any faster. We also show that the rate of convergence under the multiplicative intensity model is comparable to the density convergence rate. In addition, the variances  $\sigma_o^2$  and  $\sigma_{cv}^2$  are similar for the two estimation settings. Finally, we examine the variances for couple of different intensity functions and see that the cross-validation bandwidth performs relatively well for the intensity function with more curvature.

In the last chapter, we use the kernel intensity estimator with the cross-validation bandwidth to analyze a coffee sales data set. In this example, we are interested in estimating the purchase rate for various brands of coffee and in relating the purchase rate to promotions. The kernel estimator with the cross-validation bandwidth

assists the researcher in seeing structure in the data sets; however, there are some problems since the cross-validation score function has multiple minima. The coffee example shows that the least-squares cross-validation bandwidth not only has good theoretical properties but also is effective for data analysis.

## CHAPTER 2

### ASYMPTOTIC OPTIMALITY OF THE LEAST-SQUARES CROSS-VALIDATION BANDWIDTH UNDER THE SIMPLE MULTIPLICATIVE INTENSITY MODEL

An important aspect of kernel intensity estimation is the selection of the bandwidth or smoothing parameter. We are interested in finding a data based bandwidth which has desirable asymptotic properties. It has been proven that in the related setting of kernel density estimation, the least-squares cross-validation bandwidth is asymptotically optimal. There are two ways in which a bandwidth can be asymptotically optimal. First, the performance of the bandwidth as measured by some error criterium can be asymptotically equivalent to the "best" performance; secondly, the bandwidth can be asymptotically equal to the "best" bandwidth.

In this chapter, a simple multiplicative intensity model is used to put a mathematical structure on the intensity function of the point process. Then, we show that the performance, measured by the integrated squared error (ISE), of the cross-validation bandwidth is asymptotically optimal under the simple multiplicative intensity model. We also prove that under this model the least-squares cross-validation bandwidth is asymptotically equivalent to the bandwidth that minimizes the ISE. In the next chapter, we explore the same asymptotic questions for the stationary Cox process model.

## 2.a. The Model and the Kernel Estimator.

Let  $X_1, X_2, \dots, X_N$  be all of the observations from a nonhomogeneous Poisson process over  $[0, T]$  with intensity  $\lambda_c(x)$ . We use a simple multiplicative intensity model to put a mathematical structure on the intensity function. The general multiplicative intensity model, introduced by Aalen (1978), is frequently used to model counting processes. Assume that

$$\lambda_c(x) = c\alpha(x) \quad \text{for } x \in \mathbb{R}$$

where  $c$  is a positive constant and  $\alpha(x)$  is a nonnegative deterministic function such that  $\int_0^T \alpha(x) dx = 1$  (i.e.  $c = \int_0^T \lambda_c(x) dx$ ). This model is appropriately called a multiplicative intensity model since frequently  $c$  can be interpreted as the number of subjects in an experiment and  $\alpha(x)$  as the "individual" intensity for each subject. Under this model,  $N$  is a Poisson random variable that has expected value equal to  $c$ . Conditional on  $N$ , the occurrence times,  $X_1, X_2, \dots, X_N$ , have the same distribution as the order statistics corresponding to  $N$  independent random variables with probability density  $\alpha(x)I_{[0, T]}(x)$ . In order to avoid boundary effects at the endpoints 0 and  $T$ , we assume a circular design such that  $\lambda_c(0) = \lambda_c(T)$ ,  $\lambda_c'(0) = \lambda_c'(T)$  and  $\lambda_c''(0) = \lambda_c''(T)$ . A circular design implies that for  $0 < r, s < T$ ,  $\lambda_c(-r) = \lambda_c(T-r)$  and  $\lambda_c(T+s) = \lambda_c(s)$ .

Finally, under the simple multiplicative intensity model, we have seen that letting  $c \rightarrow \infty$  has the desirable effect of adding observations everywhere on the interval  $[0, T]$  without changing the relative shape of the target function  $\lambda_c(x)$  in the limiting process. Since we will consider the 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  $c_s/s \rightarrow \tau$  for some constant  $\tau > 0$  as  $s \rightarrow \infty$ . Note that letting  $s \rightarrow \infty$  also implies that  $c_s \rightarrow \infty$ .

The natural kernel estimate of  $\lambda_c(x)$  is

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

$K(\cdot)$  is the kernel function, and  $h$  is the bandwidth or smoothing parameter. The kernel estimator differs from the typical kernel density estimator in two ways. First,  $\hat{\lambda}_h$  does not include a normalization factor,  $n^{-1}$ , since  $\int_r^t \lambda(x)$  is the expected number rather than the expected proportion of observations between  $r$  and  $t$ . Second,  $N$  is a random variable in the intensity estimation setting.

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  $\hat{\lambda}_h^s(x)$ . We assume that as  $c_s \rightarrow \infty$ ,  $h \rightarrow 0$  such that  $hc_s \rightarrow \infty$ . It follows from Ramlau-Hansen (1983) that  $\hat{\lambda}_h^s(x)$  is uniformly consistent and asymptotically normal as  $c_s \rightarrow \infty$ ,  $h \rightarrow 0$  and  $hc_s \rightarrow \infty$ .

Our goal is to find a data based bandwidth that minimizes the integrated square error (ISE) of  $\hat{\lambda}_h$  where

$$ISE_{\lambda}(h) = \int_0^T \hat{\lambda}_h^2 - 2 \int_0^T \hat{\lambda}_h \lambda - \int_0^T \lambda^2.$$

In the intensity estimation setting, the cross-validation score function is defined as,

$$CV_{\lambda}(h) = \int_0^T \hat{\lambda}_h^2 - 2 \sum_{i=1}^N \hat{\lambda}_{hi}(X_i)$$

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

$$\hat{\lambda}_{hi}(x) = \sum_{j \neq i} h^{-1} K((x-X_j)/h).$$

Since  $\sum_{i=1}^N \hat{\lambda}_{hi}(X_i)$  is a method of moments estimator of  $\int_0^T \hat{\lambda}_h \lambda$ , and

$\int_0^T \lambda^2$  is independent of  $h$ ,  $CV_\lambda(h)$  is a reasonable estimate of the terms in  $ISE_\lambda(h)$  that depend on  $h$ . Therefore, the bandwidth that minimizes  $CV_\lambda(h)$  should be near the bandwidth that minimizes  $ISE_\lambda(h)$ .

The mean integrated square error (MISE) is another error criterion that is used to evaluate bandwidth selection procedures. Define  $k = (\int u^2 K(u) du) / 2$ . Assuming that  $K$  is a probability density function and that  $\lambda$  has two continuous bounded derivatives,

$$\begin{aligned} MISE_\lambda(h) &= \int_0^T E[(\hat{\lambda}_h(x) - \lambda(x))^2] dx \\ &= \int_0^T \text{Var}[\hat{\lambda}_h(x)] + \{\text{Bias}[\hat{\lambda}_h(x)]\}^2 dx \\ &= h^{-1} c_s (JK^2) + h^4 c_s^2 k^2 \int_0^T (\alpha'')^2 + o(h^{-1} c_s + h^4 c_s^2) \end{aligned}$$

as  $h \rightarrow 0$ ,  $c_s \rightarrow \infty$  and  $hc_s \rightarrow \infty$ . Let  $B(h)$  be the asymptotic mean integrated square error (AMISE) of  $\hat{\lambda}$ ; then,

$$B(h) = h^{-1} c_s (JK^2) + h^4 c_s^2 k^2 \int_0^T (\alpha'')^2 .$$

There are two interesting things to note about the AMISE of  $\hat{\lambda}$ . First of all, the two terms of  $B(h)$  are in fact the asymptotic integrated variance and the asymptotic integrated squared bias. As  $h$  gets larger, there are fewer "squiggles" in the kernel estimate and hence, the variance is lower; this is reflected in the first summand where  $h^{-1}$  decreases as  $h$  increases. On the other hand, as  $h$  gets larger, the kernel estimate is smoother and misses features of the true intensity function. This causes the bias to become larger which is seen in the second summand where  $h^4$  increases along with  $h$ . The second point to note is that,  $B(h)$  is equal to the AMISE in the density setting multiplied by  $c_s^2$ . Thus, accounting for the difference of scale between the density and intensity settings, the AMISE is the same in both estimation situations.

It turns out that the squared bias of the kernel estimator in the intensity setting is  $c_s^2$  times the squared bias in the density setting. However, the variances in the two settings are different. One can show that,

$$\begin{aligned} \text{Var}[\hat{\lambda}_h(x)] &= E[\text{Var}[\hat{\lambda}_h(x) | N]] + \text{Var}[E[\hat{\lambda}_h(x) | N]] \\ &= [K_h^2 * \lambda - c^{-1}(K_h * \lambda)^2] + c^{-1}(K_h * \lambda)^2 \\ &= K_h^2 * \lambda \end{aligned} \quad (2.1)$$

where "\*" is the convolution operator. Meanwhile, for the density

$$\text{estimate } \hat{f}_h(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i),$$

$$\text{Var}[n\hat{f}_h(x)] = K_h^2 * nf - n^{-1}(K_h * nf)^2.$$

Note that the variance of  $n\hat{f}$  is similar to the first two summands of (2.1). Since the number of observations coming from a Poisson process is random, the variance of  $\hat{\lambda}$  has an additional positive term,  $\text{Var}[E[\hat{\lambda}_h(x) | N]]$ . Thus, the variance of the kernel estimator is larger in the intensity setting. However, the difference between the variances is seen only in the second and higher order terms; hence, the asymptotic variance is similar for  $\hat{\lambda}$  and  $n\hat{f}$ .

Since the number of observations in  $[0, T]$ ,  $N$ , will play an important role in the proofs in this section, we state a few properties about this random variable.  $N$  has a Poisson distribution with mean  $(\int_0^T \lambda_{c_s}(x) dx)$ , and hence,

$$E(N) = \int_0^T \lambda_{c_s}(x) dx = c_s \int_0^T \alpha(x) dx = c_s.$$

As a result,  $E[(N)_m] = E[N(N-1)(N-2)\dots(N-m+1)] = c_s^m$ , and

$$E[N^m] = c_s^m + o(c_s^m). \quad (2.2)$$

In addition, it is well known (e.g. Hoel, Port and Stone, 1971) that

given  $N$ , a Poisson random variable with mean  $c_s$ , then

$$\left[ \frac{N - c_s}{\sqrt{c_s}} \right] \xrightarrow{d} N(0,1) \text{ as } s \rightarrow \infty \text{ (or equivalently } c_s \rightarrow \infty).$$

This implies that  $N/c_s = 1 + O_p(c_s^{-1/2})$ . In other words,

$$N/c_s \xrightarrow{p} 1 \text{ as } s \rightarrow \infty. \quad (2.3)$$

## 2.b. Results.

Let  $\hat{h}_0$  be any bandwidth that minimizes  $ISE_\lambda(h)$  and  $\hat{h}_{CV}$  any bandwidth that minimizes  $CV_\lambda(h)$  (these minima always exist since  $ISE_\lambda(h)$  and  $CV_\lambda(h)$  are continuous and bounded functions). In this section, assume that:

- a) The kernel function,  $K(\cdot)$ , is a compactly supported bounded symmetric probability density function. Without loss of generality assume that  $K(\cdot)$  is supported on  $[-1,1]$ .
- b) The true intensity function,  $\lambda$ , has two continuous bounded derivatives.
- c) For each value of  $s$ , the bandwidths under consideration come from a set  $H_s$  where for some constants  $\beta, \rho, \delta > 0$ ,
 
$$\#(H_s) = \{\text{the number of elements in } H_s\} \leq c_s^\rho$$
 and for  $h \in H_s$ ,  $\beta^{-1} c_s^{-(1+\delta)} \leq h \leq \beta c_s^{-\delta}$ .
- d) For some constant  $\tau > 0$ ,  $c_s/s \rightarrow \tau$  as  $s \rightarrow \infty$ .

Assumption b) is a common technical assumption which allows Taylor expansion methods to be used for studying the error functions of  $\hat{\lambda}_h(x)$ . Assumption c) can be weakened so that  $H_s$  is a continuous interval by using a continuity argument found in <sup>11</sup>Hardle and Marron (1985). This set of possible bandwidths nearly covers the range of consistent

bandwidths. Under these assumptions, the least-squares cross-validation bandwidth is asymptotically optimal for kernel intensity estimation. That is, the performance, measured in terms of the ISE, of the kernel estimator is asymptotically optimal when the least-squares cross-validation bandwidth is used. This result is stated in Theorem 2.1.

**THEOREM 2.1 :** *If assumptions a), b), c) and d) hold, then under the simple multiplicative intensity model,*

$$\frac{ISE_{\lambda}(\hat{h}_{CV})}{ISE_{\lambda}(\hat{h}_O)} \rightarrow 1 \text{ a.s. as } s \rightarrow \infty .$$

In order to prove Theorem 2.1, 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). The theorem is a direct result of the two following lemmas. In both of these lemmas, we assume the simple multiplicative intensity model and that assumptions a), b), c) and d) hold.

**Lemma 2.1:**  $\sup_{h \in H_s} |[ISE_{\lambda}(h) - B(h)] / B(h)| \rightarrow 0 \text{ a.s. as } s \rightarrow \infty .$

**Lemma 2.2:**  $\sup_{h, b \in H_s} |[CV_{\lambda}(h) - ISE_{\lambda}(h) - CV_{\lambda}(b) + ISE_{\lambda}(b)] / [B(h) + B(b)]| \rightarrow 0 \text{ a.s. as } s \rightarrow \infty .$

Recall that  $B(h)$  is the AMISE for  $\hat{\lambda}$ . Thus, Lemma 2.1 states that the ISE and MISE are essentially the same for large  $s$ . If we let  $h_0$  be the bandwidth that minimizes  $MISE_{\lambda}(h)$ , a straight forward consequence of Theorem 2.1 is that  $\hat{h}_{cv}$  is also asymptotically optimal with respect to MISE in the sense that

$$\frac{MISE_{\lambda}(\hat{h}_{cv})}{MISE_{\lambda}(h_0)} \rightarrow 1 \text{ a.s. as } s \rightarrow \infty.$$

Furthermore, Theorem 2.1 can be used to prove that the cross-validation bandwidth is asymptotically equivalent to the minimum ISE bandwidth. However, we need to change assumption c) to

$c^*$ ) For each value of  $s$ , the bandwidths under consideration come

from a set  $H_s$  where for some constants  $\rho, \delta, \xi, \eta > 0$ ,

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

$$\text{and for } h \in H_s, \quad \xi c_s^{-1/5} \leq h \leq \eta c_s^{-1/5}.$$

The second asymptotic optimality result for the cross-validation bandwidth is given in Theorem 2.2.

**Theorem 2.2:** If assumptions a), b),  $c^*$ ) and d) hold, then under the simple multiplicative intensity model,

$$\frac{\hat{h}_{cv}}{\hat{h}_0} \rightarrow 1 \text{ in probability as } s \rightarrow \infty.$$

The proofs of both theorems and the two lemmas are given in the next section.

2.c. Proofs.

We will begin with the proof of Theorem 2.1, which states that the ISE of  $\hat{\lambda}$  with the cross-validation bandwidth is asymptotically equivalent to the minimum ISE.

*proof of Theorem 2.1:*

Lemma 2.1 says that the ISE and the asymptotic MISE of  $\hat{\lambda}_h(x)$  are asymptotically equivalent as  $s \rightarrow \infty$ ; thus we get that:

$$\text{ISE}_\lambda(h) = B(h) + o(B(h)).$$

Combining Lemma 2.2 and Lemma 2.1 implies that as  $s \rightarrow \infty$ ,

$$\sup_{h, b \in H_s} |[\text{CV}_\lambda(h) - \text{ISE}_\lambda(h) - \text{CV}_\lambda(b) + \text{ISE}_\lambda(b)] / [\text{ISE}_\lambda(h) + \text{ISE}_\lambda(b)]| \rightarrow 0 \text{ a.s.} \quad (2.4)$$

Since  $\hat{h}_0$  minimizes  $\text{ISE}_\lambda(h)$  and  $\hat{h}_{cv}$  minimizes  $\text{CV}_\lambda(h)$ ,

$$\text{ISE}_\lambda(\hat{h}_0) \leq \text{ISE}_\lambda(\hat{h}_{cv}) \text{ and } \text{CV}_\lambda(\hat{h}_{cv}) \leq \text{CV}_\lambda(\hat{h}_0). \quad (2.5)$$

As a result of (2.5), as  $s \rightarrow \infty$ ,

$$\begin{aligned} | \text{ISE}_\lambda(\hat{h}_0) / \text{ISE}_\lambda(\hat{h}_{cv}) - 1 | &= | [\text{ISE}_\lambda(\hat{h}_0) - \text{ISE}_\lambda(\hat{h}_{cv})] / \text{ISE}_\lambda(\hat{h}_{cv}) | \\ &\leq | [\text{ISE}_\lambda(\hat{h}_{cv}) - \text{ISE}_\lambda(\hat{h}_0) + \text{CV}_\lambda(\hat{h}_0) - \text{CV}_\lambda(\hat{h}_{cv})] / \text{ISE}_\lambda(\hat{h}_{cv}) | \\ &\leq 2 | [\text{CV}_\lambda(\hat{h}_0) - \text{ISE}_\lambda(\hat{h}_0) - \text{CV}_\lambda(\hat{h}_{cv}) + \text{ISE}_\lambda(\hat{h}_{cv})] / [\text{ISE}_\lambda(\hat{h}_0) + \text{ISE}_\lambda(\hat{h}_{cv})] | \end{aligned} \quad (2.6)$$

From (2.4) and (2.6), we can conclude that,

$$| \text{ISE}_\lambda(\hat{h}_0) / \text{ISE}_\lambda(\hat{h}_{cv}) - 1 | \rightarrow 0 \text{ a.s. as } s \rightarrow \infty,$$

and hence,

$$| \text{ISE}_\lambda(\hat{h}_{cv}) / \text{ISE}_\lambda(\hat{h}_0) | \rightarrow 1 \text{ a.s. as } s \rightarrow \infty.$$

This completes the proof of Theorem 2.1. \*\*

It is now necessary to prove Lemma 2.1 and Lemma 2.2. The proofs of the two lemmas are given below in full detail.

proof of Lemma 2.2:

Let  $g: (1, 2, \dots, N) \rightarrow (1, 2, \dots, N)$  be a random permutation of the numbers  $1, 2, \dots, N$ . 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 Poisson process with intensity  $\lambda(x)$ , then conditional on  $N$ , the  $Y_i$ 's are i.i.d random variables with density  $\alpha(x)I_{[0, T]}(x)$ . Therefore, the methods developed by Hardle, Marron and Wand (1990) for kernel density estimates can be applied to  $\alpha(x)I_{[0, T]}(x)$ .

We begin by introducing some notation. Define:

$$\begin{aligned}\hat{\alpha}_h(x) &\equiv c_s^{-1} \sum_{i=1}^N K_h(x-X_i) = c_s^{-1} \hat{\lambda}_h(x) \\ \hat{\alpha}_{hi}(x) &\equiv c_s^{-1} \sum_{j \neq i} h^{-1} K((x-X_j)/h) = c_s^{-1} \hat{\lambda}_{hi}(x) \\ CV(h) &\equiv \int_0^T \hat{\alpha}_h^2 - 2 c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{hi}(X_i) = c_s^{-2} CV_\lambda(h)\end{aligned}$$

These are the typical definitions for kernel density estimates of  $\alpha(x)$  except that  $N$  is a random variable and  $c_s^{-1}$  is used instead of  $n^{-1}$ . As a result, the ISE for  $\hat{\alpha}$  is,

$$\Delta(h) \equiv \int_0^T \hat{\alpha}_h^2 - 2 \int_0^T \hat{\alpha}_h \alpha + \int_0^T \alpha^2 = c_s^{-2} ISE_\lambda(h).$$

Moreover, the MISE of  $\hat{\alpha}(x)$  is

$$M(h) \equiv h^{-1} c_s^{-1} (\int K^2) + h^4 k^2 \int_0^T (\alpha''(x))^2 + o(h^{-1} c_s^{-1} + h^4),$$

and therefore, the asymptotic MISE of  $\hat{\alpha}$  is

$$A(h) \equiv h^{-1} c_s^{-1} (\int K^2) + h^4 k^2 \int_0^T (\alpha''(x))^2 = c_s^{-2} B(h).$$

Finally, define:

$$U_{ij}(h) \equiv K_h(Y_i - Y_j) - \int K_h(y - Y_j) \alpha(y) dy - \alpha(Y_i) + \int_0^T \alpha^2(y) dy$$

$$V_i(h) \equiv E(U_{ij} | Y_i)$$

$$= \int K_h(Y_i - y) \alpha(y) dy - \int \int K_h(y - z) \alpha(y) \alpha(z) dy dz - \alpha(Y_i) + \int_0^T \alpha^2(y) dy$$

$$\begin{aligned}
W_{ij}(h) &\equiv U_{ij} - V_i \\
&= K_h(Y_i - Y_j) - \int K_h(y - Y_j) \alpha(y) dy - \int K_h(Y_i - y) \alpha(y) dy \\
&\quad + \iint K_h(y - z) \alpha(y) \alpha(z) dy dz
\end{aligned}$$

$$\begin{aligned}
R_i(h) &\equiv [(N-1)c_s^{-1} - 1] [2 \int K_h(Y_i - y) \alpha(y) dy - \iint K_h(y - z) \alpha(y) \alpha(z) dy dz \\
&\quad - 2\alpha(Y_i) + \int_0^T \alpha^2(y) dy].
\end{aligned}$$

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

If we sum over all of the data points, we can replace  $Y_i$  (the unordered observations) by  $X_i$  (the ordered observations). Therefore, substituting in the definitions of  $V_i$ ,  $W_{ij}$  and  $R_i$  and recalling that we are working with a circular design,

$$\begin{aligned}
&c_s^{-1} \sum_{i=1}^N V_i + c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} + c_s^{-1} \sum_{i=1}^N R_i \\
&= c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} K_h(X_i - X_j) - c_s^{-1} \sum_{i=1}^N [\int K_h(X_i - x) \alpha(x) dx] \\
&\quad - [2(N-1)c_s^{-1} - 1] c_s^{-1} \sum_{i=1}^N \alpha(X_i) + [N(N-1)c_s^{-2}] \int_0^T \alpha^2(x) dx.
\end{aligned}$$

That is,

$$c_s^{-1} \sum_{i=1}^N V_i + c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} + c_s^{-1} \sum_{i=1}^N R_i = c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{hi}(X_i) - \int_0^T \hat{\alpha}_h \alpha - G \quad (2.7)$$

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

Toward the end of this proof, we will show that Lemma 2.2 follows directly from:

$$\sup_{h \in H_s} \left| [c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{hi}(X_i) - \int_0^T \hat{\alpha}_h \alpha - G] / A(h) \right| \rightarrow 0 \text{ a.s. as } s \rightarrow \infty. \quad (2.8)$$

By (2.7), it follows that statement (2.8) holds when (2.9), (2.10) and (2.11) hold:

$$\sup_{h \in H_s} \left| c_s^{-1} \sum_{i=1}^N V_i / A(h) \right| \rightarrow 0 \text{ a.s.} \quad (2.9)$$

$$\sup_{h \in H_s} | c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} / A(h) | \rightarrow 0 \text{ a.s.} \quad (2.10)$$

$$\sup_{h \in H_s} | c_s^{-1} \sum_{i=1}^N R_i / A(h) | \rightarrow 0 \text{ a.s.} \quad (2.11)$$

Thus, in order to prove Lemma 2.2 it is sufficient to prove (2.9), (2.10) and (2.11). Observe that the  $V_i$  and  $W_{ij}$  terms are similar to the terms that arise in the density estimation setting except that the number of observations is random. The  $R_i$  term is an additional term that must be included because we are considering data from a Poisson process.

We begin by proving (2.9). Let  $k < N$ . Then, one can easily show that

$$\begin{aligned} 1) & \sigma\{Y_1, \dots, Y_k\} \subset \sigma\{Y_1, \dots, Y_{k+1}\} \\ 2) & \sum_{i=1}^k V_i \text{ is measurable w.r.t. } \sigma\{Y_1, \dots, Y_k\} \\ 3) & E\left[ \sum_{i=1}^{k+1} V_i \mid \sigma\{Y_1, \dots, Y_k\}, N \right] = \sum_{i=1}^k V_i + E[V_{k+1} \mid \sigma\{Y_1, \dots, Y_k\}, N] \\ & = \sum_{i=1}^k V_i + E[V_{k+1} \mid N] \\ & = \sum_{i=1}^k V_i. \end{aligned}$$

Thus, conditional on  $N$ ,  $\left\{ \sum_{i=1}^k V_i \right\}_{k=1}^N$  is a martingale with respect to the  $\sigma$ -fields generated by  $\{Y_1, Y_2, \dots, Y_k\}$ .

Applying a martingale inequality given by Burkholder (1973, p. 40)

to  $\left\{ \sum_{i=1}^k V_i \right\}_{k=1}^N$  implies that

$$\begin{aligned} E\left[ \left( \sup_{k=1, \dots, N} \sum_{i=1}^k V_i \right)^{2m} \mid N \right] & \leq a E\left[ \left( \sum_{i=1}^N E[V_i^2 \mid N] \right)^m \mid N \right] \\ & + a \sum_{k=1}^{\infty} E[|V_k|^{2m} \mid N]. \end{aligned} \quad (2.12)$$

First, we find the order of the two summands on the right side of the above inequality. Then, these expressions will be substituted back into the inequality. Let "a" be a generic constant throughout this discussion and let "a<sub>m</sub>" be a generic constant that depends on m. Since the squared bias of  $\hat{\alpha}_h$  has order  $h^4$ , it follows that

$$\begin{aligned}
E\left[\left(\sum_{i=1}^N E(V_i^2 | N)^m \mid N\right) \mid N\right] &= E\left[\left(\sum_{i=1}^N E(h^{-1} \int K((Y_i - y)/h) \alpha(y) dy - \alpha(Y_i))\right.\right. \\
&\quad \left.\left. - h^{-1} \int \int K((y-z)/h) \alpha(y) \alpha(z) dy dz\right.\right. \\
&\quad \left.\left. + \int_0^T \alpha^2(y) dy\right)^2 \mid N\right]^m \mid N] \\
E\left[\left(\sum_{i=1}^N E(V_i^2 | N)^m \mid N\right) \mid N\right] &\leq E[(Nah^4)^m \mid N] \\
E\left[\left(\sum_{i=1}^N E(V_i^2 | N)^m \mid N\right) \mid N\right] &\leq a_m N^m h^{4m}. \tag{2.13}
\end{aligned}$$

Moreover,

$$\sum_{k=1}^{\infty} E[|V_k|^{2m} \mid N] \leq E\left[\sum_{k=1}^N |V_k|^{2m} \mid N\right] \leq a_m N. \tag{2.14}$$

Substituting (2.13) and (2.14) into Burkholder's inequality gives,

$$\begin{aligned}
E\left[c_s^{-1} \sum_{i=1}^N V_i\right]^{2m} &\leq c_s^{-2m} E\left[\sup_{k \geq 1} \sum_{i=1}^k V_i\right]^{2m} \\
&= c_s^{-2m} E\left[E\left[\left(\sup_{k=1, \dots, N} \sum_{i=1}^k V_i\right)^{2m} \mid N\right]\right] \\
&\leq c_s^{-2m} E\left[a E\left[\left(\sum_{i=1}^N E(V_i^2 | N)\right)^m \mid N\right] + a \sum_{k=1}^{\infty} E[|V_k|^{2m} \mid N]\right] \text{ by (2.12)} \\
&\leq c_s^{-2m} E\left[a_m N^m h^{4m} + a_m N\right] \text{ by (2.13) and (2.14)} \\
&= a_m c_s^{-2m} [h^{4m} E(N^m) + E(N)] \\
&= a_m c_s^{-2m} [h^{4m} (c_s^m + o(c_s^m)) + c_s] \text{ by (2.2)} \\
&\leq a_m (c_s^{-m} h^{4m} + c_s^{-2m+1}). \tag{2.15}
\end{aligned}$$

Recall that  $A(h)$  is the AMISE of  $\hat{\alpha}_h$ , and that  $A(h) \sim c_s^{-1} h^{-1} + h^4$ .

Then, by (2.15) for  $h \in H_s$ ,

$$\begin{aligned}
E\left[ c_s^{-1} \sum_{i=1}^N V_i / A(h) \right]^{2m} &\leq \frac{a_m (c_s^{-m} h^{4m} + c_s^{-2m+1})}{(c_s^{-1} h^{-1} + h^4)^{2m}} \\
&\leq a_m \left[ \frac{c_s^{-m} h^{4m}}{c_s^{-m} h^{3m}} + \frac{c_s^{-2m+1}}{c_s^{-2m} h^{-2m}} \right] \\
&\leq a_m [h^m + c_s h^{2m}] \\
E\left[ c_s^{-1} \sum_{i=1}^N V_i / A(h) \right]^{2m} &\leq a_m c_s^{-\gamma m} \tag{2.16}
\end{aligned}$$

for some  $\gamma > 0$  and  $m$  sufficiently large. Since (2.16) holds for an arbitrary  $h \in H_s$ , we can conclude that

$$\begin{aligned}
\sup_{h \in H_s} \{ E\left[ c_s^{-1} \sum_{i=1}^N V_i / A(h) \right]^{2m} \} &\leq a_m c_s^{-\gamma m}, \\
\sup_{h \in H_s} \{ E\left[ c_s^{-1} \sum_{i=1}^N V_i / A(h) \right]^{2m} / (c_s^{-\gamma/4})^{2m} \} &\leq a_m c_s^{-(\gamma/2)m}. \tag{2.17}
\end{aligned}$$

Using Chebychev's Theorem on (2.17) we get

$$\begin{aligned}
\sup_{h \in H_s} P\left[ \left| c_s^{-1} \sum_{i=1}^N V_i \right| / A(h) > c_s^{-\gamma/4} \right] &\leq a_m c_s^{-(\gamma/2)m}, \\
\sup_{h \in H_s} P\left[ \left| c_s^{-1} \sum_{i=1}^N V_i \right| > c_s^{-\gamma/4} A(h) \right] &\leq a_m c_s^{-(\gamma/2)m}. \tag{2.18}
\end{aligned}$$

Recall that by assumption c),  $\#(H_s) \leq c_s^\rho$ . In addition, choose  $m$  such that  $[m > 2(\rho+2)/\gamma]$ . Let  $K_1$  and  $K_2$  be finite constants. Then, given a sequence  $\{c_s\}_{s=1}^\infty$  that satisfies assumption d),

$$\begin{aligned}
&\sum_{s=1}^\infty \#(H_s) \sup_{h \in H_s} P\left[ \left| c_s^{-1} \sum_{i=1}^N V_i \right| > \epsilon A(h) \right] \\
&\leq \sum_{s=1}^\infty \#(H_s) \sup_{h \in H_s} P\left[ \left| c_s^{-1} \sum_{i=1}^N V_i \right| > \min(\epsilon, c_s^{-\gamma/4}) A(h) \right] \\
&\leq \sum_{s=1}^r \#(H_s) \sup_{h \in H_s} P\left[ \left| c_s^{-1} \sum_{i=1}^N V_i \right| > \epsilon A(h) \right] \\
&\quad + \sum_{s=r+1}^\infty \#(H_s) \sup_{h \in H_s} P\left[ \left| c_s^{-1} \sum_{i=1}^N V_i \right| > c_s^{-\gamma/4} A(h) \right]
\end{aligned}$$

$$\begin{aligned}
& \sum_{s=1}^{\infty} \#(H_s) \sup_{h \in H_s} P[|c_s^{-1} \sum_{i=1}^N V_i| > \epsilon A(h)] \\
& \leq K_1 + \sum_{s=r+1}^{\infty} \#(H_s) a_m c_s^{-(\gamma/2)m} \quad \text{by (2.18)} \\
& \leq K_1 + \sum_{s=r+1}^{\infty} c_s^{\rho} a_m c_s^{-(\rho+2)} \quad \text{since } m > 2(\rho+2)/\gamma \\
& = K_1 + a_m \sum_{s=r+1}^{\infty} c_s^{-2} \\
& \leq K_1 + a_m K_2. \quad \text{by assumption d)}
\end{aligned}$$

Hence,

$$\sum_{s=1}^{\infty} \#(H_s) \sup_{h \in H_s} P[|c_s^{-1} \sum_{i=1}^N V_i| > \epsilon A(h)] < \infty. \quad (2.19)$$

(2.19) leads to

$$\begin{aligned}
\sum_{s=1}^{\infty} P[\sup_{h \in H_s} \{|c_s^{-1} \sum_{i=1}^N V_i|/A(h)\} > \epsilon] & \leq \sum_{s=1}^{\infty} \sum_{h \in H_s} P[|c_s^{-1} \sum_{i=1}^N V_i|/A(h) > \epsilon] \\
& \leq \sum_{s=1}^{\infty} \#(H_s) \sup_{h \in H_s} P[|c_s^{-1} \sum_{i=1}^N V_i|/A(h) > \epsilon] \\
& < \infty. \quad (2.20)
\end{aligned}$$

Using the Borel-Cantelli Lemma with (2.20), we can conclude that

$$P[\sup_{h \in H_s} \{|c_s^{-1} \sum_{i=1}^N V_i|/A(h)\} > \epsilon \text{ for infinitely many } s] = 0.$$

Consequently, for the sequence  $\{c_s\}$ ,

$$\sup_{h \in H_s} |c_s^{-1} \sum_{i=1}^N V_i / A(h)| \rightarrow 0 \text{ a.s. as } s \rightarrow \infty. \quad (2.21)$$

Thus, we've proven statement (2.9).

Following a similar procedure, we will now prove statement (2.10).

For  $k < N$ ,

$$1) \sum_{i=1}^k \sum_{j=1}^{i-1} W_{ij} \text{ is measurable w.r.t. } \sigma\{Y_1, \dots, Y_k\}$$

$$\begin{aligned}
2) E[ \sum_{i=1}^{k+1} \sum_{j=1}^{i-1} W_{ij} \mid \sigma\{Y_1, \dots, Y_k\}, N ] \\
&= \sum_{i=1}^k \sum_{j=1}^{i-1} W_{ij} + E[ \sum_{j=1}^k W_{k+1,j} \mid \sigma\{Y_1, \dots, Y_k\}, N ] \\
&= \sum_{i=1}^k \sum_{j=1}^{i-1} W_{ij}.
\end{aligned}$$

Hence, conditional on  $N$ ,  $\{ \sum_{i=1}^k \sum_{j=1}^{i-1} W_{ij} \}_{k=1}^N$  is a martingale with respect to the  $\sigma$ -fields generated by  $\{Y_1, Y_2, \dots, Y_k\}$ .

Applying Burkholder's inequality to the martingale  $\{ \sum_{i=1}^k \sum_{j=1}^{i-1} W_{ij} \}_{k=1}^N$

gives

$$\begin{aligned}
E[ ( \sup_{k=1, \dots, N} \sum_{i=1}^k \sum_{j=1}^{i-1} W_{ij} )^{2m} \mid N ] \\
\leq a E[ \{ \sum_{i=1}^N E[ ( \sum_{j=1}^{i-1} W_{ij} )^2 \mid Y_1, Y_2, \dots, Y_{i-1}, N ] \}^m \mid N ] \\
+ a \sum_{k=1}^{\infty} E[ ( \sum_{j=1}^{k-1} W_{ij} )^{2m} \mid N ]. \tag{2.22}
\end{aligned}$$

Now, consider the two terms on the right side of the inequality.

Observe that for  $i \neq j \neq k$ ,

$$E[W_{ij}^2 \mid N] = ah^{-1}, \text{ and}$$

$$E(W_{ij} W_{ik} \mid N) = E[E(W_{ij} W_{ik} \mid Y_i, N) \mid N] = E[E(W_{ij} \mid Y_i, N) E(W_{ik} \mid Y_i, N) \mid N] = 0.$$

Thus, for the first term of Burkholder's inequality, we can show that,

$$\begin{aligned}
E[ \{ \sum_{i=1}^N E[ ( \sum_{j=1}^{i-1} W_{ij} )^2 \mid Y_1, \dots, Y_{i-1}, N ] \}^m \mid N ] &\leq N^{2m} (ah^{-1})^m \\
&\leq a_m N^{2m} h^{-m}. \tag{2.23}
\end{aligned}$$

For the second term of the inequality, we use Lemma 2.3 which is stated and proven below to show that.

$$\sum_{k=1}^{\infty} E[ ( \sum_{j=1}^{k-1} W_{kj} )^{2m} \mid N ] = E[ \sum_{k=1}^N ( \sum_{j=1}^{k-1} W_{kj} )^{2m} \mid N ] \leq a_m N^{m+1} h^{-2m+1}. \tag{2.24}$$

**Lemma 2.3 :** For any positive integer  $m$  and  $W_{kj}$  defined as above,

$$E\left[\sum_{k=1}^N \left(\sum_{j=1}^{k-1} W_{kj}\right)^{2m} \mid N\right] \leq a_m N^{m+1} h^{-2m+1}$$

*proof of Lemma 2.3:*

Expanding the sums gives

$$\begin{aligned} E\left[\sum_{k=1}^N \left(\sum_{j=1}^{k-1} W_{kj}\right)^{2m} \mid N\right] &\leq \sum_{k=1}^N E\left[\left(\sum_{j=1}^{k-1} W_{kj}\right)^{2m} \mid N\right] \\ &= \sum_{k=1}^N \left[ \sum_{j_1=1}^N \dots \sum_{j_{2m}=1}^N E(W_{kj_1} W_{kj_2} \dots W_{kj_{2m}} \mid N) \right]. \end{aligned}$$

For each summand  $E(W_{kj_1} W_{kj_2} \dots W_{kj_{2m}} \mid N)$ , the set of indices  $(j_1, j_2, \dots, j_{2m})$  take values from the set  $\{1, 2, \dots, N\}^{2m}$ . Define the function  $f$  such that

$$f(j_1, \dots, j_{2m}) \equiv \{\text{the number of distinct values in } (j_1, \dots, j_{2m})\}.$$

Thus, the summation could be written in terms of  $f(\cdot)$ . Note also that,

$$E(W_{kj_1} W_{kj_2} \dots W_{kj_{2m}} \mid N) = 0 \text{ if there exist some } \ell \text{ such that } j_\ell \neq j_1 \text{ for}$$

all  $i \neq \ell$ . Then,

$$\begin{aligned} E\left[\sum_{k=1}^N \left(\sum_{j=1}^{k-1} W_{kj}\right)^{2m} \mid N\right] &\leq N \left[ \sum_{r=1}^{2m} \sum_{\{(j_1, \dots, j_{2m}) : f(\cdot) = r\}} E(W_{kj_1} \dots W_{kj_{2m}} \mid N) \right] \\ &= N \left[ \sum_{r=1}^m \sum_{\{(j_1, \dots, j_{2m}) : f(\cdot) = r\}} E(W_{kj_1} \dots W_{kj_{2m}} \mid N) + N \sum_{r=m+1}^{2m} 0 \right] \\ &\leq N \left[ \sum_{r=1}^m \sum_{\{(j_1, \dots, j_{2m}) : f(\cdot) = r\}} ah^{-(2m-1)} \right]. \end{aligned}$$

For  $r \leq m$ , let  $\{i_1, \dots, i_r\}$  be a set of  $r$  distinct numbers such that  $i_\ell \in \{1, 2, \dots, N\}$  for each  $\ell$ . Thus, the number of sets  $\{i_1, \dots, i_r\}$  is less than or equal to  $N^r$ . Moreover, given a set  $\{i_1, \dots, i_r\}$ , the number of sets  $(j_1, \dots, j_{2m})$  such that  $j_\ell \in \{i_1, \dots, i_r\}$  for all  $\ell = 1, 2, \dots, 2m$  is less than or equal to  $r^{2m}$ . Thus, the number of sets

of  $(j_1, \dots, j_{2m})$  such that  $[f(\cdot) = r]$  is less than or equal to  $(N^r r^{2m})$ .

Hence, one gets the final result:

$$\begin{aligned} E\left[ \sum_{k=1}^N \left( \sum_{j=1}^{k-1} W_{kj} \right)^{2m} \mid N \right] &\leq N \left[ \sum_{r=1}^m N^r r^{2m} a h^{-(2m-1)} \right] \\ &\leq N m^{2m+1} N^m a h^{-(2m-1)} \\ &= a_m N^{m+1} h^{-2m+1}. \end{aligned}$$

Thus, we have proven Lemma 2.3. \*\*

Now, we return to the proof of Lemma 2.2. We substitute the results (2.23) and (2.24) into Burkholder's inequality (2.22),

$$\begin{aligned} E\left[ c_s^{-2} \sum_{i=1}^N \sum_{j=1}^{i-1} W_{ij} \right]^{2m} &\leq c_s^{-4m} E\left[ E\left[ \left( \sup_{k=1, \dots, N} \sum_{i=1}^k \sum_{j=1}^{i-1} W_{ij} \right)^{2m} \mid N \right] \right] \\ &\leq c_s^{-4m} E\left[ a_m N^{2m} h^{-m} + a_m N^{m+1} h^{-2m+1} \right] \\ &\leq a_m (c_s^{-2m} h^{-m} + c_s^{-3m+1} h^{-2m+1}). \end{aligned}$$

Therefore, it follows that for some  $\gamma > 0$  and  $m$  sufficiently large, and for  $h \in H_s$ ,

$$\begin{aligned} E\left[ c_s^{-2} \sum_{i=1}^N \sum_{j=1}^{i-1} W_{ij} / A(h) \right]^{2m} &\leq \frac{a_m (c_s^{-2m} h^{-m} + c_s^{-3m+1} h^{-2m+1})}{(c_s^{-1} h^{-1} + h^4)^{2m}} \\ &\leq a_m [h^m + c_s^{-m+1} h] \\ E\left[ c_s^{-2} \sum_{i=1}^N \sum_{j=1}^{i-1} W_{ij} / A(h) \right]^{2m} &\leq a_m c_s^{-\gamma m}. \end{aligned} \tag{2.25}$$

Similar to the argument used for  $\sum_{i=1}^N V_i$ , we can conclude from

(2.25) that:

$$\sup_{h \in H_s} \left| c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} / A(h) \right| \rightarrow 0 \text{ a.s.}$$

This proves statement (2.10).

Finally, consider (2.11), the term involving  $R_i$ . Using Taylor expansions on the second factor of  $\sum_{i=1}^N R_i$ , one can show that as  $s \rightarrow \infty$ ,

$$\begin{aligned}
F_2 &\equiv c_s^{-1} \sum_{i=1}^N [2 \int K_h(Y_i - y) \alpha(y) dy - \iint K_h(y-z) \alpha(y) \alpha(z) dy dz - 2\alpha(Y_i) + \int_0^T \alpha^2] \\
F_2 &= 2c_s^{-1} \sum_{i=1}^N [\int K_h(Y_i - y) \alpha(y) dy - \alpha(Y_i)] \\
&\quad - Nc_s^{-1} [\iint_0^T K_h(y-z) \alpha(y) \alpha(z) dy dz - \int_0^T \alpha^2] \\
F_2 &= h^2 k \{ 2c_s^{-1} \sum_{i=1}^N [\alpha''(Y_i)] - Nc_s^{-1} \int_0^T \alpha'' \alpha \} + o_p(h^2). \tag{2.26}
\end{aligned}$$

Note that since  $h \in H_s$ ,  $s \rightarrow \infty$  implies that  $c_s \rightarrow \infty$  as well as  $h \rightarrow 0$ .

Moreover, by assumption b)  $\alpha$  and  $\alpha''$  are bounded; thus, we get

$$E[(c_s^{-1} \sum_{i=1}^N [\alpha''(Y_i)])^2] \leq M_1 < \infty, \quad (\int_0^T \alpha \alpha'')^2 \leq M_2 < \infty, \tag{2.27}$$

$$E[(c_s^{-1} \sum_{i=1}^N [\alpha''(Y_i)])^2] = (\int_0^T \alpha \alpha'')^2 + c_s^{-1} \int_0^T \alpha^2 \alpha'' \rightarrow (\int_0^T \alpha \alpha'')^2 \text{ as } s \rightarrow \infty. \tag{2.28}$$

(2.27) and (2.28) imply that  $(c_s^{-1} \sum_{i=1}^N [\alpha''(Y_i)])$  converges to  $\int_0^T \alpha \alpha''$  in  $L^2$ .

Therefore, as  $s \rightarrow \infty$ ,

$$(c_s^{-1} \sum_{i=1}^N [\alpha''(Y_i)]) \xrightarrow{P} \int_0^T \alpha \alpha'' \tag{2.29}$$

Since  $Nc_s^{-1} \xrightarrow{P} 1$ , we can also conclude that as  $s \rightarrow \infty$ ,

$$Nc_s^{-1} \int_0^T \alpha'' \alpha \xrightarrow{P} \int_0^T \alpha \alpha'' \tag{2.30}$$

Combining (2.29) and (2.30) with (2.26) leads to the result that

$$h^{-2} F_2 \xrightarrow{P} k \{ \int_0^T \alpha'' \alpha \} \text{ as } s \rightarrow \infty,$$

and hence for  $0 < \epsilon < \delta/2$ ,

$$c_s^{-\epsilon} h^{-2} F_2 \xrightarrow{P} 0 \text{ as } s \rightarrow \infty. \tag{2.31}$$

Furthermore, since  $N$  is Poisson( $c_s$ ), it follows that as  $s \rightarrow \infty$ ,

$$c_s^{1/2} [(N-1)/c_s - 1] \xrightarrow{d} N(0,1). \tag{2.32}$$

Using the definition of  $R_i$  along with (2.31) and (2.32), we can use

Slutsky's theorem to conclude that

$$c_s^{1/2-\epsilon} h^{-2} [c_s^{-1} \sum_{i=1}^N R_i] \xrightarrow{p} 0 \text{ as } s \rightarrow \infty. \quad (2.33)$$

It can be shown that for all positive integers  $m$ ,

$$[c_s^{1/2-\epsilon} h^{-2} (c_s^{-1} \sum_{i=1}^N R_i)] \in L^{2m} \text{ and } \{[c_s^{1/2-\epsilon} h^{-2} (c_s^{-1} \sum_{i=1}^N R_i)]^{2m}: s=1,2,\dots\}$$

is a uniformly integrable family. Hence, it follows from (2.33) that

$$c_s^{m-2\epsilon m} h^{-4m} E[c_s^{-1} \sum_{i=1}^N R_i]^{2m} \rightarrow 0 \text{ as } s \rightarrow \infty.$$

Thus, there exists some  $c_0 > 0$  such that for  $c_s > c_0$ ,

$$E[c_s^{-1} \sum_{i=1}^N R_i]^{2m} \leq a_m c_s^{-m+2\epsilon m} h^{4m}.$$

Since  $h \in H_s$ , then for  $c_s > c_0$ ,

$$\begin{aligned} E[c_s^{-1} \sum_{i=1}^N R_i / A(h)]^{2m} &\leq \frac{a_m c_s^{-m+2\epsilon m} h^{4m}}{(c_s^{-1} h^{-1} + h^4)^{2m}} \\ &\leq a_m c_s^{2\epsilon m} h^m \\ E[c_s^{-1} \sum_{i=1}^N R_i / A(h)]^{2m} &\leq a_m c_s^{-\gamma m} \end{aligned} \quad (2.34)$$

for  $m$  sufficiently large and some  $\gamma > 0$ . (2.11) follows from (2.34) as seen in the proof of statement (2.9).

At this point, we collect the above results to show that (2.8) holds. By (2.7), (2.9), (2.10) and (2.11), it follows that

$$\begin{aligned} \limsup_{s \rightarrow \infty} \sup_{h \in H_s} & |(c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{hi}(X_i) - \int_0^T \hat{\alpha}_h \alpha - G) / A(h)| \\ &= \limsup_{s \rightarrow \infty} \sup_{h \in H_s} \{ |(c_s^{-1} \sum_{i=1}^N V_i + c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} + c_s^{-1} \sum_{i=1}^N R_i) / A(h)| \} \end{aligned}$$

$$\begin{aligned}
& \lim_{s \rightarrow \infty} \sup_{h \in H_s} \left| \left( c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{hi}(X_i) - \int_0^T \hat{\alpha}_h \alpha - G \right) / A(h) \right| \\
& \leq \lim_{s \rightarrow \infty} \sup_{h \in H_s} \left| c_s^{-1} \sum_{i=1}^N V_i / A(h) \right| + \lim_{s \rightarrow \infty} \sup_{h \in H_s} \left| c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} / A(h) \right| \\
& \quad \lim_{s \rightarrow \infty} \sup_{h \in H_s} \left| c_s^{-1} \sum_{i=1}^N R_i / A(h) \right| \\
& = 0 \quad \text{a.s.}
\end{aligned}$$

This limit is the same as (2.8). We now prove that (2.8) implies the final result of Lemma 2.2.

$$\begin{aligned}
& \sup_{h, b \in H_s} \left| [CV_\lambda(h) - ISE_\lambda(h) - CV_\lambda(b) + ISE_\lambda(b)] / [B(h) + B(b)] \right| \\
& = \sup_{h, b \in H_s} \left| 2 \left[ -\sum_{i=1}^N \hat{\lambda}_{hi} + \int_0^T \hat{\lambda}_h \lambda + \sum_{i=1}^N \hat{\lambda}_{bi} - \int_0^T \hat{\lambda}_b \lambda \right] / [B(h) + B(b)] \right| \\
& = \sup_{h, b \in H_s} \left| 2 \left[ -c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{hi} + \int_0^T \hat{\alpha}_h \alpha + c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{bi} - \int_0^T \hat{\alpha}_b \alpha \right] / [A(h) + A(b)] \right| \\
& \leq 2 \left[ \sup_{h \in H_s} \left| \left( c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{hi}(X_i) - \int_0^T \hat{\alpha}_h \alpha - G \right) / A(h) \right| \right. \\
& \quad \left. + \sup_{b \in H_s} \left| \left( c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{bi}(X_i) - \int_0^T \hat{\alpha}_b \alpha - G \right) / A(b) \right| \right] \\
& \rightarrow 0 \quad \text{a.s.}
\end{aligned}$$

This completes the proof of Lemma 2.2. \*\*

*proof of Lemma 2.1:*

Let  $Y_1, Y_2, \dots, Y_N$  be the "unordered" observations as in Lemma 2.2.

From the definitions of  $\Delta(h)$  and  $M(h)$  and the fact that we are using a circular design, we get that

$$\Delta(h) - M(h) = \int_0^T [\hat{\alpha}_h(y) - \alpha(y)]^2 dy - \int_0^T E[\hat{\alpha}_h(y) - \alpha(y)]^2 dy$$

$$\begin{aligned}
\Delta(h) - M(h) &= \int_0^T [\hat{\alpha}_h(y) - E\hat{\alpha}_h(y)]^2 dy + 2 \int_0^T [\hat{\alpha}_h(y) - E\hat{\alpha}_h(y)] [E\hat{\alpha}_h(y) - \alpha(y)] dy \\
&\quad - \int_0^T E[\hat{\alpha}_h(y) - E\hat{\alpha}_h(y)]^2 dy \\
&= c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} \int K_h(y - Y_i) K_h(y - Y_j) dy + c_s^{-2} \sum_{i=1}^N \int [K_h(y - Y_i)]^2 dy \\
&\quad - 2c_s^{-1} \sum_{i=1}^N \iint K_h(y - Y_i) K_h(y - x) \alpha(x) dx dy \\
&\quad + \iint \iint K_h(y - x) K_h(y - z) \alpha(x) \alpha(z) dx dz dy \\
&\quad + 2 [c_s^{-1} \sum_{i=1}^N \iint K_h(y - Y_i) K_h(y - x) \alpha(x) dx dy - c_s^{-1} \sum_{i=1}^N \int K_h(y - Y_i) \alpha(y) dy \\
&\quad - \iint \iint K_h(y - x) K_h(y - z) \alpha(x) \alpha(z) dx dz dy + \iint K_h(y - x) \alpha(x) \alpha(y) dx dy] \\
&\quad - c_s^{-1} \iint [K_h(x - y)]^2 \alpha(x) dx dy.
\end{aligned}$$

Define:

$$V_i = \iint K_h(y - Y_i) K_h(y - x) \alpha(x) dx dy - \int K_h(y - Y_i) \alpha(y) dy$$

$$- \iint \iint K_h(y - x) K_h(y - z) \alpha(x) \alpha(z) dx dz dy + \iint K_h(y - x) \alpha(x) \alpha(y) dx dy$$

$$T_i = c_s^{-1} \int K_h^2(Y_i - y) dy - c_s^{-1} \iint K_h^2(x - y) \alpha(x) dx dy$$

$$W_{ij} = \int K_h(y - Y_i) K_h(y - Y_j) dy - \iint K_h(y - Y_i) K_h(y - x) \alpha(x) dx dy$$

$$- \iint K_h(y - x) K_h(y - Y_j) \alpha(x) dx dy + \iint \iint K_h(y - x) K_h(y - z) \alpha(x) \alpha(z) dx dz dy$$

$$R = -[N(N-1)/c_s^2 - 2N/c_s + 1] \iint \iint K_h(y - x) K_h(y - z) \alpha(x) \alpha(z) dx dz dy$$

$$+ 2[N/c_s - 1] [c_s^{-1} \sum_{i=1}^N \iint K_h(y - Y_i) K_h(y - x) \alpha(x) dx dy - \iint K_h(y - x) \alpha(x) \alpha(y) dx dy]$$

$$+ [N/c_s - 1] c_s^{-1} \iint [K_h(x - y)]^2 \alpha(x) dx dy$$

$$- 2c_s^{-2} \sum_{i=1}^N \iint K_h(y - Y_i) K_h(y - x) \alpha(x) dx dy.$$

Therefore, it follows that

$$\Delta(h) - M(h) = c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} + 2c_s^{-1} \sum_{i=1}^N V_i + c_s^{-1} \sum_{i=1}^N T_i + R. \quad (2.35)$$

Consider each of the summands in the expression above. First of

all, note that

$$T_i = c_s^{-2} \int_0^T \text{Var}[\hat{\lambda}_h(y)] dy = o(c_s^{-1} h^{-1}).$$

With assumption b), this leads to

$$\sup_{h \in H_s} |c_s^{-1} \sum_{i=1}^N T_i / A(h)| \rightarrow 0 \text{ as } s \rightarrow \infty. \quad (2.36)$$

Using Burkholder's inequality, as in Lemma 2.2, it can be shown that,

$$\sup_{h \in H_s} |c_s^{-1} \sum_{i=1}^N V_i / A(h)| \rightarrow 0 \text{ a.s. as } s \rightarrow \infty. \quad (2.37)$$

$$\sup_{h \in H_s} |c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} / A(h)| \rightarrow 0 \text{ a.s. as } s \rightarrow \infty. \quad (2.38)$$

Finally, consider the term  $R/A(h)$ . We can write  $R = P_1 + P_2 + P_3 + P_4$

where:

$$P_1 = [N(N-1)/c_s^2 - 2N/c_s + 1] \iiint K_h(y-x) K_h(y-z) \alpha(x) \alpha(z) dx dz dy$$

$$P_2 = [N/c_s - 1] c_s^{-1} \iint [K_h(x-y)]^2 \alpha(x) dx dy$$

$$P_3 = [N/c_s - 1] [c_s^{-1} \sum_{i=1}^N \iint K_h(y-Y_i) K_h(y-x) \alpha(x) dx dy - \iint K_h(y-x) \alpha(x) \alpha(y) dx dy]$$

$$P_4 = c_s^{-2} \sum_{i=1}^N \iint K_h(y-Y_i) K_h(y-x) \alpha(x) dx dy.$$

Since  $K$  and  $\alpha$  are bounded and  $c_s^{1/2} [N/c_s - 1] \xrightarrow{d} N(0,1)$ , one can easily show that

$$P_1 = o_p(c_s^{-1}). \quad (2.39)$$

$$P_2 = o_p(c_s^{-3/2}) = o_p(c_s^{-1}). \quad (2.40)$$

$$P_4 = o_p(c_s^{-1}). \quad (2.41)$$

First look at  $P_1$  and  $P_4$ . (2.39) and (2.41) imply that for  $0 < \epsilon < \delta/2$ ,

$$c_s^{1-\epsilon} (P_1 + P_4) \xrightarrow{p} 0 \text{ as } s \rightarrow \infty. \quad (2.42)$$

Since for all positive integers  $m$ ,  $[c^{1-\epsilon}(P_1 + P_4)] \in L^{2m}$  and  $\{[c^{1-\epsilon}(P_1 + P_4)]^{2m} : s=1,2,\dots\}$  is a uniformly integrable family, it follows from (2.42) that

$$c_s^{(2-2\epsilon)m} E[P_1 + P_4]^{2m} \rightarrow 0 \quad \text{as } s \rightarrow \infty.$$

Thus, there exists some  $c_1 > 0$  such that for  $c_s > c_1$ ,

$$E[P_1 + P_4]^{2m} \leq a_m c_s^{-(2-2\epsilon)m}. \quad (2.43)$$

Next consider  $P_3$ . Taylor expansions can be used to show that as  $s \rightarrow \infty$

$$c_s^{-1} \sum_{i=1}^N \iint K_h(y - Y_i) K_h(y - x) \alpha(x) dx dy - \iint \alpha(y) K_h(y - x) \alpha(x) dx dy = o_p(h^2).$$

Hence, it follows from Slutsky's theorem that,

$$P_3 = o_p(c_s^{-1/2+\epsilon} h^2). \quad (2.44)$$

As seen above, (2.40), (2.44) can be used to show that there exists

some  $c_2 > 0$  such that for  $c_s > c_2$ ,

$$E[P_2 + P_3]^{2m} \leq a_m (c_s^{-2m} + c_s^{-(1-2\epsilon)m} h^{4m}). \quad (2.45)$$

Thus, using (2.43) and (2.45), there exists some  $c_0 > 0$  such that

for  $c_s > c_0$  and  $h \in H_s$ ,

$$\begin{aligned} E[R/A(h)]^{2m} &\leq \frac{a_m [c_s^{-(2-2\epsilon)m} + c_s^{-2m} + c_s^{-(1-2\epsilon)m} h^{4m}]}{(c_s^{-1} h^{-1} + h^4)^{2m}} \\ &\leq a_m [c_s^{2\epsilon m} h^{2m} + h^{2m} + c_s^{2\epsilon m} h^m] \\ &\leq a_m c_s^{-\gamma m} \end{aligned}$$

for  $m$  sufficiently large and some  $\gamma > 0$ . Hence, using similar methods as seen in Lemma 2.2,

$$\sup_{h \in H_s} |R / A(h)| \rightarrow 0 \quad \text{a.s. as } s \rightarrow \infty. \quad (2.46)$$

Combining the results (2.36), (2.37), (2.38) and (2.46) in (2.35),

it follows that

$$\begin{aligned} &\limsup_{s \rightarrow \infty} \sup_{h \in H_s} |[\Lambda(h) - M(h)] / A(h)| \\ &= \limsup_{s \rightarrow \infty} \sup_{h \in H_s} |[2c_s^{-1} \sum_{i=1}^N V_i + c_s^{-1} \sum_{i=1}^N T_i + c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} + R] / A(h)| \end{aligned}$$

$$\begin{aligned}
& \limsup_{s \rightarrow \infty} \sup_{h \in H_s} |[\Delta(h) - M(h)] / A(h)| \\
& \leq 2 \limsup_{s \rightarrow \infty} \sup_{h \in H_s} |c_s^{-1} \sum_{i=1}^N V_i / A(h)| + \limsup_{s \rightarrow \infty} \sup_{h \in H_s} |c_s^{-1} \sum_{i=1}^N T_i / A(h)| \\
& \quad + \limsup_{s \rightarrow \infty} \sup_{h \in H_s} |c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} / A(h)| + \limsup_{s \rightarrow \infty} \sup_{h \in H_s} |R / A(h)| \\
& = 0 \quad \text{a.s.} \tag{2.47}
\end{aligned}$$

One can also show that for  $A(h)$ , the asymptotic MISE of  $\hat{\alpha}_h$ , we get

$$\sup_{h \in H_s} |[M(h) - A(h)] / A(h)| \rightarrow 0 \quad \text{a.s. as } s \rightarrow \infty. \tag{2.48}$$

As a result of (2.47) and (2.48),

$$\begin{aligned}
\sup_{h \in H_s} |[ISE_{\lambda}(h) - B(h)] / B(h)| &= \sup_{h \in H_s} |[\Delta(h) - A(h)] / A(h)| \\
&\leq \sup_{h \in H_s} |[\Delta(h) - M(h)] / A(h)| + \sup_{h \in H_s} |[M(h) - A(h)] / A(h)| \\
&\rightarrow 0 \quad \text{a.s.}
\end{aligned}$$

Thus, we have proven Lemma 2.1. \*\*

Finally, we prove the second theorem; that is, we show that the cross-validation bandwidth is asymptotically equivalent to the minimum ISE bandwidth.

*proof of Theorem 2.2:*

As seen earlier in this chapter, the asymptotic MISE of the intensity function  $\lambda$  is defined by

$$B(h) = c_s h^{-1} (\int K^2) + c_s^2 h^4 k^2 \int_0^T (\alpha'')^2.$$

as  $c_s \rightarrow \infty$ ,  $h \rightarrow 0$  and  $c_s h \rightarrow \infty$ . Let  $h^{\circ}$  be the value that minimizes  $B(h)$ ; then,  $h^{\circ} = a_0 c_s^{-1/5}$  for some constant  $a_0$  which depends on  $\alpha$  and  $K$ . Moreover,

$$B(h^{\circ}) = a_1 c_s^{-4/5} \tag{2.49}$$

for some constant  $a_1$ . Therefore, in order to prove Theorem 2.2, it is sufficient to prove the following statements:

$$\hat{h}_o/h^o \rightarrow 1 \text{ in probability as } s \rightarrow \infty. \quad (2.50)$$

$$\hat{h}_{cv}/h^o \rightarrow 1 \text{ in probability as } s \rightarrow \infty. \quad (2.51)$$

By Lemma 2.1,  $ISE_\lambda(h)$  and  $B(h)$  are asymptotically equivalent, and hence, one can show that

$$ISE_\lambda(\hat{h}_o)/B(h^o) \rightarrow 1 \text{ a.s. as } s \rightarrow \infty. \quad (2.52)$$

As a direct result of Theorem 2.1 and (2.52), we get that

$$ISE_\lambda(\hat{h}_{cv})/B(h^o) \rightarrow 1 \text{ a.s. as } s \rightarrow \infty. \quad (2.53)$$

Given (2.52) and (2.53), the argument proving (2.50) and (2.51) is similar to the proof found in Hall (1983) (p. 1160). We will prove (2.51) here; the proof of (2.50) is analogous.

Note that  $\hat{h}_{cv}$  is a function of  $s$ . By assumption  $c^*$ , it follows that for all  $s$ ,  $c_s^{1/5} \hat{h}_{cv} \in [\xi, \eta]$ , and thus,

$$(\hat{h}_{cv}/h^o) \in [\xi/a_o, \eta/a_o]. \quad (2.54)$$

Given  $c_s$ , let  $F_s$  be the distribution function for  $(\hat{h}_{cv}/h^o)$ . Then by (2.54),  $\{F_s : s=1,2,\dots\}$  is said to be "tight" sequence of functions. Using Helly's extraction principle, there is subsequence  $\{s_k\}_{k=1}^\infty$  and a distribution function  $F$  such that  $F_{s_k}(x) \rightarrow F(x)$  as  $k \rightarrow \infty$  for all  $x \in \mathbb{R}$ . That is,  $F_{s_k}$  converges weakly to  $F$ . (2.54) further implies that the limiting distribution is proper and supported on  $[\xi/a_o, \eta/a_o]$ . Let  $\ell > 0$  be a point of support of the limiting distribution.

Now suppose that (2.51) is not true. Then, it is possible to choose a subsequence  $\{s_k\}$  where  $\ell \neq 1$ . Let  $0 < \ell < 1$ ; the case of  $\ell > 1$  follows in a similar fashion, but we will not present that argument

here. Define  $\zeta = 1/2 \min\{\ell, 1-\ell\}$ . Now choose  $d > 0$  sufficiently small so that for all  $z \in [\ell - \zeta, \ell + \zeta]$  and some  $\rho > 1$ ,

$$\frac{c_s (h^\circ z)^{-1} \left\{ \int_d^{T-d} \alpha(x) dx \right\} \left\{ \int K^2(y) dy \right\} + c_s^2 (h^\circ z)^4 k^2 \int_d^{T-d} \{\alpha''(x)\}^2 dx}{c_s (h^\circ)^{-1} \int K^2(y) dy + c_s^2 (h^\circ)^4 k^2 \int_0^T \{\alpha''(x)\}^2 dx} > \rho. \quad (2.55)$$

It is always possible to find such a  $d$  because  $h^\circ$  minimizes the denominator. Moreover, since  $h^\circ = a_{0s}^{-1/5}$ , the left side of (2.55) is independent of  $s$ . From the definition of  $B(h)$  and Lemma 2.1, one can show that

$$\begin{aligned} \text{ISE}_\lambda(\hat{h}_{cv}) &= c_s \hat{h}_{cv}^{-1} (\int K^2) + c_s^2 \hat{h}_{cv}^4 k^2 \int_0^T (\alpha'')^2 + o_p(c_s^{-4/5}) \\ &\geq c_s \hat{h}_{cv}^{-1} (\int_d^{T-d} \alpha) (\int K^2) + c_s^2 \hat{h}_{cv}^4 k^2 \int_d^{T-d} (\alpha'')^2 + o_p(c_s^{-4/5}). \end{aligned} \quad (2.56)$$

When  $(\hat{h}_{cv}/h^\circ) \in [\ell - \zeta, \ell + \zeta]$ , (2.55) and (2.56) imply that

$$\begin{aligned} \text{ISE}_\lambda(\hat{h}_{cv}) &> \rho [c_s (h^\circ)^{-1} (\int K^2) + c_s^2 (h^\circ)^4 k^2 \int_0^T (\alpha'')^2] + o_p(c_s^{-4/5}) \\ &= \rho B(h^\circ) + o_p(c_s^{-4/5}). \end{aligned}$$

Therefore,

$$\text{ISE}_\lambda(\hat{h}_{cv})/B(h^\circ) > \rho + o_p(1) \quad \text{for } (\hat{h}_{cv}/h^\circ) \in [\ell - \zeta, \ell + \zeta]. \quad (2.57)$$

Furthermore, since  $\ell$  is a point of support of the subsequence limit of  $(\hat{h}_{cv}/h^\circ)$ , it follows that

$$\limsup_{s \rightarrow \infty} P\{(\hat{h}_{cv}/h^\circ) \in [\ell - \zeta, \ell + \zeta]\} > 0. \quad (2.58)$$

The combination of (2.57) and (2.58) contradicts statement (2.53).

Hence, (2.51) must be true, and consequently the theorem has been proven. \*\*

CHAPTER 3  
THE LEAST-SQUARES CROSS-VALIDATION BANDWIDTH UNDER THE  
STATIONARY COX PROCESS MODEL

For kernel intensity estimation, we would like to find a data based bandwidth has asymptotically appealing properties. In Chapter 2, we proved that the least-squares cross-validation bandwidth is asymptotically optimal under the simple multiplicative intensity model. In this chapter, a stationary Cox process model is considered for putting a mathematical structure on the intensity function. This situation is more complicated since the true intensity function is a random function. We will show that in general the the least-squares cross-validation bandwidth is not asymptotically optimal (measured by the ISE) for a stationary Cox process model.

3.a. *The Model.*

A second model that is frequently used to put a mathematical structure on counting processes is the stationary Cox process model. Throughout this section, let  $X_1, X_2, \dots, X_N$  form a partial realization of a stationary Cox process also known as a doubly stochastic Poisson process with intensity function  $\lambda_\mu(x)$  on the interval  $[0, T]$ . A stationary Cox process is defined by :

- 1)  $\{A(x), x \in \mathbb{R}\}$  is a stationary nonnegative valued random process,

2) conditional on the realization  $\lambda_\mu(x)$  of  $\Lambda(x)$ , the point process is a nonhomogeneous Poisson process with rate function  $\lambda_\mu(x)$ .

Furthermore, for each  $\mu$ , assume that for all  $x, y \in [0, T]$ ,

$$3) E[\Lambda(x)] = \mu,$$

$$4) E[\Lambda(x)\Lambda(y)] = \mu(|x-y|) \quad \text{where } \mu(x) = \mu^2 v_0(x)$$

for  $v_0(x)$  a fixed function.

The above definition and assumptions are identical to those found in Diggle and Marron (1987). In addition, to avoid boundary effects, we employ a circular design where  $\lambda_\mu(0) = \lambda_\mu(T)$ ,  $\lambda'_\mu(0) = \lambda'_\mu(T)$  and  $\lambda''_\mu(0) = \lambda''_\mu(T)$ . The Cox process model differs from the simple multiplicative intensity model since the true intensity function  $\lambda_\mu(x)$  of a stationary Cox process is a *random* function with mean  $\mu$ .

Under the Cox process model, letting  $\mu \rightarrow \infty$  again has the effect of adding observations everywhere on the interval  $[0, T]$  and not changing the relative shape of the target function  $\lambda_\mu(x)$  in the limiting process. Therefore, we require the values of  $\mu$  to come from the sequence  $\{\mu_s\}_{s=1}^\infty$  such that  $\mu_s/s \rightarrow \kappa$  for some constant  $\kappa > 0$ .

The kernel estimate of  $\lambda_{\mu_s}(x)$  is

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

The kernel estimate of  $\lambda$  under the Cox process model is identical to  $\hat{\lambda}_h(x)$  in the simple multiplicative intensity model; the difference between the two models is only seen when the estimators are analyzed mathematically.

The ISE of  $\hat{\lambda}_h(x)$  and the cross-validation score function are defined as

$$\begin{aligned} \text{ISE}_\lambda(h) &= \int_0^T \hat{\lambda}_h^2 - 2 \int_0^T \hat{\lambda}_h \Lambda - \int_0^T \Lambda^2. \\ \text{CV}_\lambda(h) &= \int_0^T \hat{\lambda}_h^2 - 2 \sum_{i=1}^N \hat{\lambda}_{hi}(X_i) \quad \text{where} \quad \hat{\lambda}_{hi}(x) = \sum_{j \neq i} K_h(x - X_j). \end{aligned}$$

In the Cox process setting, it is natural to consider the expectation of the ISE of  $\hat{\lambda}_h$  conditional on  $\{\Lambda(x), x \in [0, T]\}$ . When  $k = (\int u^2 K)/2$ ,

$$\begin{aligned} E[\text{ISE}_\lambda(h) \mid \Lambda] &= E\left[\int_0^T (\hat{\lambda}(x) - \Lambda(x))^2 \mid \Lambda\right] \\ &= \iint K_h^2(x-y) \Lambda(y) dy dx - \int [\int K_h(x-y) \Lambda(y) dy - \Lambda(x)]^2 dx \\ &= h^{-1} (\int_0^T \Lambda) (\int K^2) + h^4 k^2 \int_0^T (\Lambda'')^2 + o(h^{-1} \int_0^T \Lambda + h^4 \int_0^T (\Lambda'')^2) \end{aligned}$$

as  $\mu_s \rightarrow \infty$ ,  $h \rightarrow 0$  and  $\mu_s h \rightarrow \infty$ . Let  $B(h)$  be the asymptotic conditional expectation of the ISE of  $\hat{\lambda}_h(x)$ , then we define

$$B(h) = h^{-1} (\int_0^T \Lambda) (\int K^2) + h^4 k^2 \int_0^T (\Lambda'')^2.$$

Finally, one can show that  $E[(\Lambda''(x))^2] = \mu_s^2 v_0^{(4)}(0)$ , and hence, the MISE and the asymptotic MISE (AMISE) of  $\hat{\lambda}_h(x)$  are

$$\begin{aligned} \text{MISE}_\lambda(h) &= E\left[\int_0^T (\hat{\lambda}(x) - \Lambda(x))^2\right] \\ &= \mu_s T h^{-1} (\int K^2) + \mu_s^2 T h^4 E[(\Lambda''/\mu)^2] k^2 + o(\mu^2 h^4) \\ &= \mu_s T h^{-1} (\int K^2) + \mu_s^2 T h^4 v_0^{(4)}(0) k^2 + o(\mu^2 h^4) \\ \text{AMISE}_\lambda(h) &= \mu_s T h^{-1} (\int K^2) + \mu_s^2 T h^4 v_0^{(4)}(0) k^2. \end{aligned}$$

### 3.b. Differences Between the Stationary Cox Process Model and the Simple Multiplicative Intensity Model.

In Chapter 2, we saw that the number of observations in  $[0, T]$  under the multiplicative intensity model is a Poisson random variable with mean  $c_s$ . As a result,  $N$  is asymptotically normal with standard deviation  $c_s^{1/2}$ , and  $N/c_s$  converges to 1 in probability. Under the stationary Cox process model,  $N$  is not a Poisson random variable, and the relationship between  $N$  and  $E[N]$  is different. Note that,

$$E[N] = E[E[N|\Lambda]] = E[\int_0^T \Lambda(x) dx] = \mu_s T$$

In addition,

$$E[N^2] = E[(\int_0^T \Lambda(x) dx)^2 + \int_0^T \Lambda(x) dx] = \mu_s^2 \int_0^T \int_0^T v_0(|x-y|) dx dy + \mu_s T$$

If we define  $G(T) = \int_0^T \int_0^T v_0(|x-y|) dx dy$ , then,

$$\text{Var}(N) = \mu_s^2 G(T) + \mu_s T - (\mu_s T)^2 = (G(T) - T^2) \mu_s^2 + \mu_s T,$$

and as  $s \rightarrow \infty$ ,

$$\text{Var}[N/(\mu_s T)] = [(G(T) - T^2) \mu_s^2 + \mu_s T] / (\mu_s T)^2 \rightarrow [G(T) - T^2] / T^2.$$

Hence, under the stationary Cox process model, as  $s \rightarrow \infty$ ,

- 1)  $N/E[N] = N/(\mu_s T)$  does not converge in probability,
- 2)  $(\mu_s T)^{1/2} (N/(\mu_s T) - 1)$  does not converge in distribution to  $N(0,1)$ .

Through out this chapter, let  $c = \int_0^T \Lambda(x) dx$ . Restrict the values of  $c$  to the sequence  $\{c_s\}_{s=1}^{\infty}$  such that  $c_s/s \rightarrow \tau$  for some constant  $\tau > 0$ . Then conditional on  $\{c_s\}$ ,

- 1)  $N$  is a  $\text{Poisson}(c_s)$ ,
- 2)  $c_s^{1/2} (N/c_s - 1) \xrightarrow{d} N(0,1)$  as  $s \rightarrow \infty$ ,
- 3)  $N/E[N|c_s] = 1 + o_p(c_s^{-1/2})$ .

As a result, we might want to consider the effect of conditioning on  $(\int_0^T \Lambda)$  or the effect of conditioning on  $\{\Lambda(x), x \in \mathbb{R}\}$  in the Cox process setting.

The ISE and the MISE are two error criteria used for evaluating  $\hat{\lambda}_h$ . In the density estimation setting and under the multiplicative intensity model, the ISE and the MISE are asymptotically equivalent. We will show that this is not the case under the stationary Cox process model.

For a simple example, consider a homogeneous doubly stochastic Poisson process. That is, let  $\Lambda(x) = \Lambda_0$  for all  $x \in [0, T]$ , and assume that  $\Lambda_0$  has an exponential distribution with mean  $\mu_s$ .

In this situation, for  $x, y \in [0, T]$ ,

$$E[\Lambda(x)\Lambda(y)] = E[\Lambda_0^2] = 2\mu_s^2 \quad (\text{i.e. } v_0(|x-y|) = 2)$$

$$\text{and } \Lambda''(x) = 0.$$

Using the definitions given in section 3.a,

$$\text{AMISE}_\lambda(h) = \mu T h^{-1} (Jk^2)$$

$$E[\text{ISE}_\lambda(h) | \Lambda] = (J_0^T \Lambda_0) h^{-1} (Jk^2) + o(h^{-1} J_0^T \Lambda_0)$$

$$\text{Var}[E[\text{ISE}_\lambda(h) | \Lambda]] = \mu T h^{-1} (Jk^2) + o(h^{-1} \mu)$$

Thus,

$$\text{Var} \left| \frac{E[\text{ISE}_\lambda | \Lambda]}{\text{AMISE}} \right| \rightarrow 1 \quad \text{as } s \rightarrow \infty.$$

This implies that for the homogeneous doubly stochastic Poisson process, the variance of  $|(E[\text{ISE}_\lambda(h) | \Lambda] - \text{MISE}(\hat{\lambda}_h))/\text{AMISE}_\lambda(h)|$  converges to one, and hence,  $|(E[\text{ISE}_\lambda(h) | \Lambda] - \text{MISE}(\hat{\lambda}_h))/\text{AMISE}_\lambda(h)|$  does not converge to zero in probability.

Furthermore, consider the difference between  $E[\text{ISE}_\lambda(h) | \Lambda]$  and  $\text{MISE}(h)$  under a more general stationary Cox process model. First, note that,

$$E[(\int_0^T \Lambda) / \mu_s T] = 1$$

$$E[(\int_0^T (\Lambda'')^2) / (\mu_s^2 T)] = \mu_s^{-2} T^{-1} \int_0^T E[(\Lambda''(x))^2] = v_0^{(4)}(0)$$

Second, define the random variables  $Z_1$  and  $Z_2$  to be

$$Z_1 \equiv [(\int_0^T \Lambda) / \mu_s T - 1] \quad \text{and} \quad Z_2 \equiv [(\int_0^T (\Lambda'')^2) / (\mu_s^2 T) - v_0^{(4)}(0)].$$

If we choose  $h$  such that  $h \sim \mu_s^{-1/5}$ , then, as  $s \rightarrow \infty$ ,

$$\begin{aligned} E[\text{ISE}_\lambda(h) | \Lambda] - \text{MISE}_\lambda(h) &= \mu_s T h^{-1} (Jk^2) Z_1 + \mu_s^2 T h^4 k^2 Z_2 + o(\mu_s h^{-1} + \mu_s^2 h^4) \\ &= \mu_s^{6/5} T (Jk^2) Z_1 + \mu_s^{6/5} T k^2 Z_2 + o(\mu_s^{6/5}) \end{aligned}$$

and hence,

$$\frac{E[ISE_{\lambda}(h) | \Lambda] - MISE_{\lambda}(h)}{AMISE_{\lambda}(h)} = aZ_1 + bZ_2 + o(1)$$

where  $a = (\int K^2) / [(\int K^2) + k^2 v_0^{(4)}(0)]$  and  $b = k^2 / [(\int K^2) + k^2 v_0^{(4)}(0)]$ .

Moreover, one can show that

$$\begin{aligned} \text{Var}[aZ_1 + bZ_2] &= a^2 \{ T^{-2} \int_0^T \int_0^T v_0(|x-y|) dx dy - 1 \} \\ &\quad + b^2 \{ \mu_s^{-4} T^{-2} E[\int_0^T \int_0^T (\Lambda''(x) \Lambda''(y))^2 dx dy] - [v_0^{(4)}(0)]^2 \} \\ &\quad - 2ab \{ \mu_s^{-3} T^{-2} E[\int_0^T \int_0^T \Lambda(x) (\Lambda''(y))^2 dx dy] - v_0^{(4)}(0) \} \end{aligned}$$

As a result,  $\text{Var}[aZ_1 + bZ_2]$  may be a strictly positive constant. When this is the case, the variance of  $\{|E[ISE_{\lambda}(h) | \Lambda] - MISE_{\lambda}(h)\} / AMISE_{\lambda}(h) |$  converges to a strictly positive constant as  $s \rightarrow \infty$ , and hence,

$\{|E[ISE_{\lambda}(h) | \Lambda] - MISE_{\lambda}(h)\} / AMISE_{\lambda}(h) |$  does not converge to zero in probability.

Consequently,  $E[ISE_{\lambda}(h) | \Lambda]$  and  $MISE_{\lambda}(h)$  are not asymptotically equivalent in general. On the other hand,  $ISE_{\lambda}(h)$  and  $E[ISE_{\lambda}(h) | \Lambda]$  are asymptotically equivalent under the stationary Cox process model. Since we are interested in minimizing the ISE of  $\hat{\lambda}_h$ , we will study the asymptotic behavior of the cross-validation bandwidth conditional on the process  $\Lambda$ . Unfortunately, conditioning on  $\Lambda$  makes the stationary Cox process much less interesting. Essentially, given  $\{\Lambda(x), x \in [0, T]\}$ , the stationary Cox process setting is equivalent to the simple multiplicative intensity setting. Thus, conditional on  $\{\Lambda(x), x \in [0, T]\}$ , we can get the same asymptotic properties for the stationary Cox process model as we proved for the simple multiplicative intensity model.

3.c. Results.

As in Chapter 2, let  $\hat{h}_0$  minimize  $ISE_\lambda(h)$  and  $\hat{h}_{CV}$  minimize  $CV_\lambda(h)$ .

Also, assume that

a)  $K$  is a compactly supported bounded symmetric probability density function.

b)  $\Lambda$  has two continuous bounded derivatives.

c) For each value of  $s$ , the bandwidths under consideration come from a set  $H_s$  where for some constants  $\beta, \rho, \delta > 0$ ,

$$\#(H_s) \leq c_s^\rho, \text{ and for } h \in H_s, \beta^{-1} c_s^{(-1+\delta)} \leq h \leq \beta c_s^{-\delta}.$$

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

Under these assumptions, the least-squares cross-validation bandwidth is asymptotically optimal for the conditional Cox process setting.

This is stated in Theorem 3.1.

**THEOREM 3.1:** Let assumptions a), b), c) and d) and the stationary Cox process model hold. Then, conditional on the random variable  $c_s = \int_0^T \Lambda$  and the random function  $\{a(x) = \Lambda(x) / (\int_0^T \Lambda), x \in [0, T]\}$ ,

$$\frac{ISE_\lambda(\hat{h}_{CV})}{ISE_\lambda(\hat{h}_0)} \rightarrow 1 \text{ a.s. as } s \rightarrow \infty.$$

The proof of Theorem 3.1 is very similar to the proof of Theorem 2.1 in the simple multiplicative case. Again the theorem is a direct result of the two lemmas below. In these lemmas, assume that the assumptions a), b), c), and d) hold.

**Lemma 3.1:** Under the stationary Cox process model, and conditional on  $c_s = \int_0^T \Lambda$  and  $\{\alpha(x) = \Lambda(x) / (\int_0^T \Lambda), x \in [0, T]\}$ ,

$$\sup_{h \in H_s} |[ISE_\lambda(h) - B(h)] / B(h)| \rightarrow 0 \text{ a.s. as } s \rightarrow \infty.$$

**Lemma 3.2:** Under the stationary Cox process model, and conditional on  $c_s = \int_0^T \Lambda$  and  $\{\alpha(x) = \Lambda(x) / (\int_0^T \Lambda), x \in [0, T]\}$ ,

$$\sup_{h, b \in H_s} |[(CV_\lambda(h) - ISE_\lambda(h) - CV_\lambda(b) + ISE_\lambda(b))] / [B(h) + B(b)]| \rightarrow 0 \text{ a.s.}$$

as  $s \rightarrow \infty$ .

First of all, note that under the Cox process model,  $B(h)$  is the asymptotic expression of  $E[ISE_\lambda(h) | \Lambda]$ . Thus, the proof of the Lemma 3.1 above is identical to the proof of Lemma 2.1 in Chapter 2 except that  $E[ISE_\lambda(h) | \Lambda]$  is used instead of  $MISE_\lambda(h)$ . In this Chapter, we will only present an outline of the proof of Lemma 3.2.

As in Chapter 2, we can use Theorem 3.1 to give results regarding the convergence of  $(h_{cv} / \hat{h}_0)$  conditional on the process  $\Lambda$ .

3.d. Proofs.

proof of Lemma 3.2:

We begin with some notation. For  $c_s = \int_0^T \Lambda$ , define,

$$\alpha(x) \equiv \Lambda(x) / c_s.$$

Then, we can estimate  $\alpha(x)$  by

$$\hat{\alpha}_h(x) = c_s^{-1} \sum_{i=1}^N K_h(x - X_i).$$

Again, define

$$\Delta(h) \equiv \int_0^T \hat{\alpha}_h^2 - 2 \int_0^T \hat{\alpha}_h \alpha - \int_0^T \alpha^2.$$

$$CV(h) \equiv \int_0^T \hat{\alpha}_h^2 - 2c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{hi}(X_i) \quad \text{where } \hat{\alpha}_{hi}(x) = c_s^{-1} \sum_{j \neq i} K_h(x-X_j) .$$

Therefore,

$$E[\Lambda(h) \mid \Lambda] = h^{-1} c_s^{-1} (\int K^2) + h^4 k^2 \int_0^T (\alpha'')^2 + o(h^{-1} c_s^{-1} + h^4)$$

Finally, let  $A(h)$  be the asymptotic conditional expectation of  $\Lambda(h)$ ,

$$A(h) \equiv c_s^{-1} h^{-1} (\int K^2) + h^4 k^2 \int_0^T (\alpha'')^2 = c_s^{-2} B(h) .$$

Let  $Y_1, Y_2, \dots, Y_N$  be the "unordered" observations as in Lemma 2.2.

Conditional on  $N$  and on  $\{\Lambda(x), x \in [0, T]\}$ , the  $Y_i$ 's are i.i.d. random variables with probability density function  $\alpha(x) I_{[0, T]}(x)$ .

Define:

$$U_{ij}(h) = h^{-1} K_h(Y_i - Y_j) - \int K_h(y - Y_j) \alpha(y) dy - \alpha(Y_i) + \int_0^T \alpha^2(y) dy$$

$$V_i(h) = E(U_{ij} \mid Y_i, N)$$

$$W_{ij}(h) = U_{ij} - V_i$$

$$R_i(h) = [(N-1)c_s^{-1} - 1] [2 \int K_h(Y_i - y) \alpha(y) dy - \int \int K_h(y-z) \alpha(y) \alpha(z) dy dz - 2\alpha(Y_i) + \int_0^T \alpha^2(y) dy]$$

Conditional on  $N$  and on the process  $\Lambda$ ,  $\left\{ \sum_{i=1}^k V_i \right\}_{k=1}^N$  and

$\left\{ \sum_{i=1}^k \sum_{j=1}^{i-1} W_{ij} \right\}_{k=1}^N$  are both martingales with respect to the  $\sigma$ -fields

generated by  $\{Y_1, Y_2, \dots, Y_k\}$ . Now we proceed exactly as in Lemma 2.2

for the simple multiplicative intensity case except that we use conditional expectations and probabilities given  $\{\Lambda(x), x \in [0, T]\}$ .

By Burkholder's inequality, it follows that

$$\begin{aligned} E\left[\left(c_s^{-1} \sum_{i=1}^N V_i\right)^{2m} \mid \Lambda\right] &\leq c_s^{-2m} E\left[E\left[\left(\sup_{k=1, \dots, N} \sum_{i=1}^k V_i\right)^{2m} \mid N, \Lambda\right] \mid \Lambda\right] \\ &\leq c_s^{-2m} E\left[a E\left[\left(\sum_{i=1}^N E[V_i^2 \mid N, \Lambda]\right)^m \mid N, \Lambda\right] + a \sum_{k=1}^{\infty} E\left[|V_k|^{2m} \mid N, \Lambda\right] \mid \Lambda\right]. \end{aligned}$$

In addition,

$$\begin{aligned} E[(c_s^{-1} \sum_{i=1}^N V_i)^{2m} | \Lambda] &\leq c_s^{-2m} E[aN^m h^{4m} + aN | \Lambda] \\ &\leq c_s^{-2m} [h^{4m} (a_m c_s^m + o(c_s^m)) + ac_s] \\ &\leq a_m [c_s^{-m} h^{4m} + c_s^{-2m+1}] \end{aligned}$$

Similar to Lemma 2.2, since  $A(h) \sim c_s^{-1} h^{-1} + h^4$ , there exists some  $\tau > 0$  so that for  $m$  sufficiently large and  $h \in H_s$ ,

$$E[(c_s^{-1} \sum_{i=1}^N V_i / A(h))^{2m} | \Lambda] \leq \frac{a_m [c_s^{-m} h^{4m} + c_s^{-2m+1}]}{[c_s^{-1} h^{-1} + h^4]^{2m}} \leq a_m c_s^{-\tau m}.$$

Equivalently, one can write,

$$E[(c_s^{-1} \sum_{i=1}^N V_i / A(h))^{2m} | \alpha, c_s] \leq a_m c_s^{-\tau m}.$$

Therefore, using the same procedure as seen in Lemma 2.2, conditional on  $\alpha$  and  $c_s$ ,

$$\sup_{h \in H_s} |c_s^{-1} \sum_{i=1}^N V_i / A(h)| \rightarrow 0 \text{ a.s. as } s \rightarrow \infty \quad (3.1)$$

Using conditional expectations and probabilities and arguments similar to Lemma 2.2 in the simple multiplicative case it can be shown that given  $\alpha$  and  $c_s$ ,

$$\sup_{h \in H_s} |c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} / A(h)| \rightarrow 0 \text{ a.s. as } s \rightarrow \infty, \quad (3.2)$$

$$\sup_{h \in H_s} |c_s^{-1} \sum_{i=1}^N R_i / A(h)| \rightarrow 0 \text{ a.s. as } s \rightarrow \infty. \quad (3.3)$$

Now, let  $G = [2(N-1)c_s^{-1} - 1] c_s^{-1} \sum_{i=1}^N \alpha(X_i) - [N(N-1)c_s^{-2}] \int_0^T \alpha^2(x) dx$ .

Then, conditional on  $\alpha$  and  $c_s$ .

$$\begin{aligned}
& \sup_{h, b \in H_s} | [CV_\lambda(h) - ISE_\lambda(h) - CV_\lambda(b) + ISE_\lambda(b)] / [B(h) + B(b)] | \\
&= \sup_{h, b \in H_s} | [-2 \sum_{i=1}^N \hat{\lambda}_{hi} + 2 \int_0^T \hat{\lambda}_h^\top \lambda + 2 \sum_{i=1}^N \hat{\lambda}_{bi} - 2 \int_0^T \hat{\lambda}_b^\top \lambda] / [B(h) + B(b)] | \\
&\leq 4 \sup_{h \in H_s} | [c_s^{-1} \sum_{i=1}^N \hat{\alpha}_{hi} - \int_0^T \hat{\alpha}_h^\top \alpha - G] / A(h) | \\
&\leq 4 [ \sup_{h \in H_s} | c_s^{-1} \sum_{i=1}^N V_i / A(h) | + \sup_{h \in H_s} | c_s^{-2} \sum_{i=1}^N \sum_{j \neq i} W_{ij} / A(h) | \\
&\quad + \sup_{h \in H_s} | c_s^{-1} \sum_{i=1}^N R_i / A(h) | ] \\
&\rightarrow 0 \quad \text{a.s. as } s \rightarrow \infty \quad \text{by (3.1), (3.2) and (3.3)}
\end{aligned}$$

This completes the proof of Lemma 3.2. \*\*

## CHAPTER 4

### THE ASYMPTOTIC DISTRIBUTION OF THE LEAST-SQUARES CROSS-VALIDATION BANDWIDTH UNDER THE SIMPLE MULTIPLICATIVE INTENSITY MODEL

In Chapter 2, we showed that the least-squares cross-validation bandwidth is asymptotically optimal for kernel intensity estimation under the simple multiplicative intensity model. The next problem to consider is the rate of convergence of this bandwidth to an optimal bandwidth. If the rate of convergence for an asymptotically optimal bandwidth is extraordinarily slow, then that bandwidth will often perform poorly for moderately sized data sets. Therefore, the convergence rate of the cross-validation bandwidth,  $\hat{h}_{cv}$ , is critical. There are two bandwidths that could be considered "optimal"; they are  $\hat{h}_0$ , the bandwidth that minimizes the integrated square error of  $\hat{\lambda}$ , and  $h_0$ , the bandwidth that minimizes the mean integrated square error of  $\hat{\lambda}$ . If we aim to approximate the optimal bandwidth  $\hat{h}_0$ , then, it turns out that the cross-validation bandwidth does as well as  $h_0$  to the first and second order. We examine the relationship between these bandwidths by finding the asymptotic distributions of  $(\hat{h}_{cv} - \hat{h}_0)$  and  $(h_0 - \hat{h}_0)$ . Moreover, in order to evaluate the performance of the cross-validation bandwidth, we find the asymptotic distribution of the distance between the ISE with the cross-validation bandwidth and the minimum ISE.

4.a. Assumptions and Notation.

Assume that  $X_1, X_2, \dots, X_N$  are observations in the interval  $[0, T]$  from a nonhomogeneous Poisson process with a simple multiplicative intensity function

$$\lambda_{c_s}(x) = c_s \alpha(x) \quad \text{for } x \in \mathbb{R}$$

where  $c_s$  is a positive constant,  $\alpha(x)$  is a nonnegative deterministic function such that  $\int_0^T \alpha(x) dx = 1$ . Under this model,  $N$  is a Poisson random variable with mean  $c_s$ . Moreover, conditional on  $N$ ,  $\alpha(x) I_{[0, T]}(x)$  is the density function of the "unordered" observations. We assume a circular design with  $\lambda_{c_s}(0) = \lambda_{c_s}(T)$ ,  $\lambda'_{c_s}(0) = \lambda'_{c_s}(T)$  and  $\lambda''_{c_s}(0) = \lambda''_{c_s}(T)$ .

The kernel estimate of the intensity function is defined as in Chapter 2.

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

where  $K_h(x) = h^{-1} K(x/h)$ . Again,  $\hat{\alpha}_h(x) = \hat{\lambda}_h(x)/c_s$  will be the kernel estimator of  $\alpha(x)$ .

We will assume the following technical conditions throughout this Chapter.

- 1)  $K$  is a compactly supported, symmetric probability density function on  $\mathbb{R}$  with Holder continuous derivative  $K'$  such that,  $(\int z^2 K(z) dz)/2 \equiv k \neq 0$ . Without loss of generality, assume that  $K$  is supported on  $[-1, 1]$ .
- 2)  $\lambda$  is bounded and twice differentiable.  $\lambda'$  and  $\lambda''$  are bounded and integrable.  $\lambda''$  is uniformly continuous.
- 3) For some constant  $\tau > 0$ ,  $c_s/s \rightarrow \tau$  as  $s \rightarrow \infty$ .

Regarding the first assumption, recall that a function  $g$  is Holder

continuous if  $|g(x)-g(y)| \leq a|x-y|^\epsilon$  for some  $a, \epsilon > 0$  and all  $x, y$ .

Assumption 2) implies that the same properties hold for  $\alpha$  as well as for  $\lambda$ . Since we are going to deal with sequences of random variables we need a countable sequence of indices  $\{c_s\}_{s=1}^\infty$ . The third assumption gives a sequence of values  $\{c_s\}$  such that  $s \rightarrow \infty$  implies that  $c_s \rightarrow \infty$ .

We define the following functions of the ISE and MISE of  $\hat{\lambda}_h$ :

$$\Delta(h) \equiv c_s^{-2} \text{ISE}_\lambda(h) = c_s^{-2} \int_0^T (\hat{\lambda}_h(x) - \lambda(x))^2 dx$$

$$M(h) \equiv c_s^{-2} \text{MISE}_\lambda(h) = c_s^{-2} \int_0^T E[(\hat{\lambda}_h(x) - \lambda(x))^2] dx$$

$$D(h) \equiv \Delta(h) - M(h)$$

Essentially,  $\Delta(h)$  and  $M(h)$  are the ISE and the MISE for  $\hat{\alpha}(x)$ . Let  $\hat{\lambda}_{hi}(x) \equiv \sum_{j \neq i} h^{-1} K((x-X_j)/h)$  be the leave-one-out estimator, then define several functions of the cross-validation score function of  $\hat{\lambda}_h$ :

$$CV(h) \equiv c_s^{-2} CV_\lambda(h) = c_s^{-2} \left[ \int_0^T \hat{\lambda}_h^2(x) dx - 2 \sum_{i=1}^N \hat{\lambda}_{hi}(X_i) \right]$$

$$\delta(h) \equiv 2c_s^{-2} \left[ \int_0^T \hat{\lambda}_h(x) \lambda(x) dx - \sum_{i=1}^N \hat{\lambda}_{hi}(X_i) \right]$$

$$\text{Thus, } CV(h) = \Delta(h) + \delta(h) - \int_0^T \alpha^2 dx.$$

Finally, let  $h_0$ ,  $\hat{h}_0$  and  $\hat{h}_{CV}$  be the bandwidths that minimize  $\text{MISE}_\lambda(h)$ ,  $\text{ISE}_\lambda(h)$  and  $CV_\lambda(h)$  respectively. Hence, these bandwidths also minimize  $M(h)$ ,  $\Delta(h)$  and  $CV(h)$  respectively. Define

$$a_1 \equiv \int K^2 \quad \text{and} \quad a_2 \equiv k^2 \int_0^T (\alpha'')^2 dx.$$

Then, given a circular design, we can conclude that as  $c_s \rightarrow \infty$ ,  $h \rightarrow 0$ , and  $c_s h \rightarrow \infty$ ,

$$M(h) = c_s^{-1} h^{-1} a_1 + h^4 a_2 + o(c_s^{-1} h^{-1} + h^4).$$

$$M''(h) = 2c_s^{-1} h^{-3} a_1 + 12h^2 a_2 + o(c_s^{-1} h^{-3} + h^2).$$

Thus,  $M(h)$  is minimized by  $h_0$  such that  $h_0 \sim a_0 c_s^{-1/5}$  where  $a_0 \equiv [a_1/(4a_2)]^{1/5}$ . Furthermore,  $M''(h_0) \sim a_3 c_s^{-2/5}$  where

$a_3 \equiv 2a_1a_0^{-3} + 12a_2a_0^2$ . It will also be useful to denote  $a_4 \equiv (a_0^2a_3)/2$ .

Finally, define:

$$L(z) \equiv -zK'(z)$$

$$\sigma_0^2 = 8/(a_0^5a_3^2) \int_0^T \alpha^2 \int \{ \int K(y+z)[K(z)-L(z)]dz \}^2 dy + 16k^2/(a_3^2) [ \int_0^T (\alpha'')^2 \alpha + \int_0^T \alpha'' \alpha^2 ] \quad (4.1)$$

$$\sigma_{cv}^2 = 8/(a_0^5a_3^2) \int_0^T \alpha^2 (\int L^2) + 16k^2/(a_3^2) [ \int_0^T (\alpha'')^2 \alpha + \int_0^T \alpha'' \alpha^2 ] \quad (4.2)$$

#### 4.b. Results and Discussion.

Given all of the definitions and assumptions in the previous section, we are now prepared to present the asymptotic distribution functions for  $(\hat{h}_{cv} - \hat{h}_0)$  and  $(\hat{h}_0 - h_0)$ . It is reasonable to standardize these differences by dividing them by  $h_0$  which is of order  $c_s^{-1/5}$ . Hence,  $[(\hat{h}_0 - h_0)/h_0]$  is the relative distance between the minimum MISE bandwidth and the minimum ISE bandwidth, while  $[(\hat{h}_{cv} - \hat{h}_0)/h_0]$  is the appropriately scaled distance between the cross-validation bandwidth and the minimum ISE bandwidth. Using these variables, we get the asymptotic distributions found in Theorem 4.1.

**Theorem 4.1:** Given assumptions 1), 2) and 3), under the simple multiplicative intensity model

$$c_s^{1/10} \left[ \frac{\hat{h}_0 - h_0}{h_0} \right] \xrightarrow{D} N(0, \sigma_0^2) \text{ as } s \rightarrow \infty,$$

and

$$c_s^{1/10} \left[ \frac{\hat{h}_{cv} - \hat{h}_0}{h_0} \right] \xrightarrow{D} N(0, \sigma_{cv}^2) \text{ as } s \rightarrow \infty.$$

There are a few important things to notice regarding this theorem. First of all, the convergence rate for  $[(\hat{h}_{cv} - \hat{h}_0)/h_0]$  and  $[(\hat{h}_0 - h_0)/h_0]$  is  $c_s^{-1/10}$ . Thus, it turns out that the cross-validation bandwidth converges very slowly to the bandwidth that minimizes the ISE; however, in terms of the asymptotic rate,  $\hat{h}_{cv}$  does as well as the minimum MISE bandwidth.

In the kernel density estimation setting, Hall and Marron (1987) showed that  $[(\hat{h}_{cv} - \hat{h}_0)/h_0]$  and  $[(\hat{h}_0 - h_0)/h_0]$  converge in distribution with rate  $n^{-1/10}$ . Since  $c_s$  is the expected value of the number of observations, the convergence rate for the multiplicative intensity model is parallel to the density estimation result. Furthermore, Hall and Marron (1987) proved that  $[(\hat{h}_{cv} - \hat{h}_0)/h_0]$  and  $[(\hat{h}_0 - h_0)/h_0]$  are each asymptotically normal. Considering this fact, our result for the intensity estimation setting makes sense intuitively. Conditional on  $N$ ,  $[(\hat{h}_{cv} - \hat{h}_0)/h_0]$  and  $[(\hat{h}_0 - h_0)/h_0]$  are asymptotically normal as in density estimation. Moreover, the number of observations,  $N$ , is a Poisson random variable, and a Poisson random variable is known to be asymptotically normally distributed. Essentially, the unconditional asymptotic cumulative distributions of  $[(\hat{h}_{cv} - \hat{h}_0)/h_0]$  and  $[(\hat{h}_0 - h_0)/h_0]$  are the limits of weighted normal cumulative distributions where the weight function is the normal density. Taking the limit of this weighted normal cumulative distribution, we find that the asymptotic distributions of  $[(\hat{h}_{cv} - \hat{h}_0)/h_0]$  and  $[(\hat{h}_0 - h_0)/h_0]$  are also normal for intensity estimation.

The variances for the asymptotic distributions in the intensity setting are similar to the variances in the density setting. However,

$\sigma_o^2$  are not  $\sigma_{cv}^2$  equivalent. We compare these variances in more detail after we present the second theorem.

The integrated square error measures the distance between the kernel estimator and the true intensity function  $\lambda$ , and hence,  $ISE_\lambda(h^*)$  evaluates the performance of the kernel estimator with the bandwidth  $h^*$ . Since  $h_o$  minimizes  $ISE_\lambda$ ,  $h_o$  performs "best" according this error criterion. Therefore, when  $|ISE_\lambda(h^*) - ISE_\lambda(h_o)|$  is comparatively small, the kernel estimator with bandwidth  $h^*$  is doing a relatively good job of estimating  $\lambda$ .

At this point, we will compare the degree to which  $ISE_\lambda(h_o)$  and  $ISE_\lambda(h_{cv})$  differ from  $ISE(h_o)$ . Notice that the ISE of  $\hat{\lambda}_h$  is of order  $c_s^2$ . Thus, it makes sense to consider the scaled version of the integrated square error  $\Delta(h) = c_s^{-2} ISE(h)$ . In order to compare the ISE's of these bandwidths, an extra technical condition is necessary. Thus, assume that a fourth condition holds:

- 4)  $K$  has a second derivative on  $\mathbb{R}$ , and  $K''$  is Holder continuous.

With the above condition as well as all of the conditions and definitions from the previous section, we can calculate the asymptotic distributions found in Theorem 4.2.

**Theorem 4.2:** Given assumptions 1), 2), 3) and 4), under the simple multiplicative intensity model

$$c_s \{ \Delta(h_o) - \Delta(\hat{h}_o) \} \xrightarrow{D} \sigma_o^2 a_4 \chi_1^2 \quad \text{as } s \rightarrow \infty,$$

and

$$c_s \{ \Delta(\hat{h}_{cv}) - \Delta(\hat{h}_o) \} \xrightarrow{D} \sigma_{cv}^2 a_4 \chi_1^2 \quad \text{as } s \rightarrow \infty.$$

Our aim is to find a bandwidth that minimizes  $ISE(h)$  or similarly  $\Lambda(h)$ . In general, we believe that the minimum MISE bandwidth,  $h_0$ , would perform quite well; however,  $h_0$  is unknown. Theorem 2.2 implies that the data based cross-validation bandwidth performs almost as well as  $h_0$  since the distance from the minimum scaled integrated square error (i.e.  $\Lambda(\hat{h}_0)$ ) is of order  $c_s^{-1}$  for both  $\Lambda(\hat{h}_{cv})$  and  $\Lambda(h_0)$ .

Again, the convergence rate  $c_s^{-1}$  seen in Theorem 4.2 is analogous to the  $n^{-1}$  rate of convergence found by Hall and Marron (1987) for density estimation. In addition, both estimation settings yield asymptotic Chi-square distributions.

At this point, it is interesting to compare the asymptotic variances in Theorems 4.1 and 4.2 for a nonhomogeneous Poisson process to the respective variances derived from a density function. Let  $A_i(K)$   $i=1,2,3$  be constants in terms of the kernel function  $K(\cdot)$ . Then, under the multiplicative intensity model,

$$\begin{aligned}\sigma_0^2 &= A_1(K) \left[ \int_0^T (\alpha'')^2 \right]^{-1/5} \left( \int_0^T \alpha^2 \right) \\ &\quad + A_2(K) \left[ \int_0^T (\alpha'')^2 \right]^{-6/5} \left[ \int_0^T (\alpha'')^2 \alpha + \left( \int_0^T \alpha'' \alpha \right)^2 \right], \\ \sigma_{cv}^2 &= A_3(K) \left[ \int_0^T (\alpha'')^2 \right]^{-1/5} \left( \int_0^T \alpha^2 \right) \\ &\quad + A_2(K) \left[ \int_0^T (\alpha'')^2 \right]^{-6/5} \left[ \int_0^T (\alpha'')^2 \alpha + \left( \int_0^T \alpha'' \alpha \right)^2 \right].\end{aligned}$$

For kernel estimation of a density function  $f$  estimation, let the asymptotic variances be defined by  $AVar[n^{1/10}(\hat{h}_0 - h_0)/h_0] \equiv \sigma_{od}^2$  and  $AVar[n^{1/10}(\hat{h}_{cv} - \hat{h}_0)/h_0] \equiv \sigma_{cd}^2$ . Then, it follows from Hall and Marron (1987) that,

$$\begin{aligned}\sigma_{od}^2 &= A_1(K) \left[ \int (f'')^2 \right]^{-1/5} \left( \int f^2 \right) \\ &\quad + A_2(K) \left[ \int (f'')^2 \right]^{-6/5} \left[ \int (f'')^2 f - \left( \int f'' f \right)^2 \right],\end{aligned}$$

$$\sigma_{cd}^2 = A_3(K) [\int (f'')^2]^{-1/5} (\int f^2) + A_2(K) [\int (f'')^2]^{-6/5} [\int (f'')^2 f - (\int f'' f)^2].$$

Hence, the asymptotic variances in the density setting are similar to the asymptotic variances in the intensity setting. Recall that when we contrasted the variances of the kernel estimators in the two settings (Chapter 2, section 2.a), we found that the variance of the kernel intensity estimator was larger than the variance of the kernel density estimator, but that the highest order terms were the same (i.e. the asymptotic variances were equal). Similarly, the variance of  $[(\hat{h}_{cv} - \hat{h}_0)/h_0]$  and the variance of  $[(\hat{h}_0 - h_0)/h_0]$  are larger in the intensity estimation setting since the number of observations from a Poisson process is random; however, this difference shows up only in smaller order terms. Hence, the asymptotic distribution of the cross-validation bandwidth is equivalent for the two estimation settings.

The more important issue is the comparison of  $\sigma_0^2$  to  $\sigma_{cv}^2$  under the simple multiplicative intensity model. First of all, we will show that  $\sigma_{cv}^2 \geq \sigma_0^2$ . By the Cauchy-Schwartz inequality and the fact that  $\int K=1$ , it follows that

$$\begin{aligned} \{\int K(y+z)[K(z)-L(z)]dz\}^2 &\leq \int K(y+z)dz \int K(y+z)[K(z)-L(z)]^2 dz \\ &= \int K(y+z)[K(z)-L(z)]^2 dz . \end{aligned}$$

Integrating over  $y$  leads to the conclusion that

$$\int \{\int K(y+z)[K(z)-L(z)]dz\}^2 dy \leq \int [K(z)-L(z)]^2 dz = \int L^2(z) dz.$$

Thus, using the definitions of  $\sigma_0^2$  and  $\sigma_{cv}^2$  given in (4.1) and (4.2), we get the result:

$$\begin{aligned} \sigma_{cv}^2 - \sigma_o^2 &= 8(a_o^5 a_3^2)^{-1} (\int_o^T \alpha^2) [(\int L^2) - \int \{ \int K(y+z) [K(z) - L(z)] dz \}^2 dy] \\ \sigma_{cv}^2 - \sigma_o^2 &\geq 0. \end{aligned} \quad (4.3)$$

As  $s \rightarrow \infty$ , the asymptotic variances of  $[(\hat{h}_o - h_o)/h_o]$  and  $[(\hat{h}_{cv} - \hat{h}_o)/h_o]$  are  $(c_s^{-1/5} \sigma_o^2)$  and  $(c_s^{-1/5} \sigma_{cv}^2)$  respectively, and the asymptotic variances of  $\{\Delta(h_o) - \Delta(\hat{h}_o)\}$  and  $\{\Delta(\hat{h}_{cv}) - \Delta(\hat{h}_o)\}$  are  $(2c_s^{-2} a_4^2 \sigma_o^4)$  and  $(2c_s^{-2} a_4^2 \sigma_{cv}^4)$  respectively. Therefore, since our goal is to minimize ISE, (4.3)

implies that  $h_o$  performs slightly better than  $\hat{h}_{cv}$ .

Secondly, we illustrate the relative sizes of  $\sigma_o^2$  and  $\sigma_{cv}^2$  by giving a couple of examples. Consider the intensity functions  $\lambda_i(x) = c\alpha_i(x)$   $i=1,2$  where

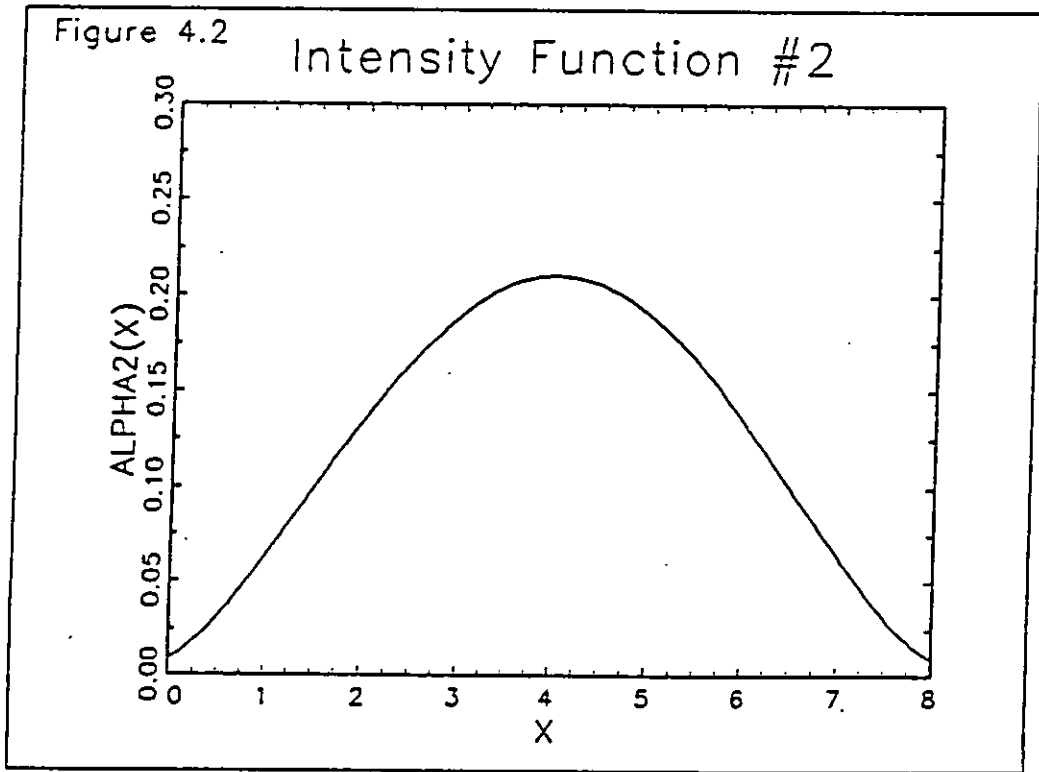
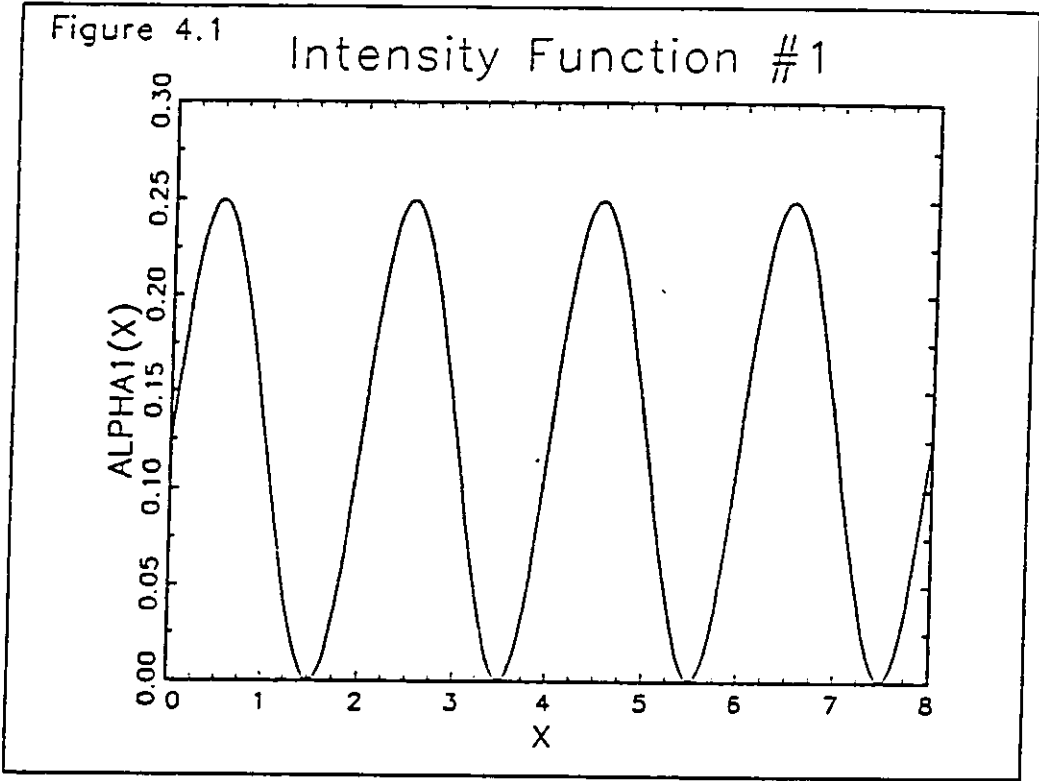
$$\begin{aligned} \alpha_1(x) &= (1/8) [\sin(\pi x) + 1] I_{[0,8]}(x), \\ \alpha_2(x) &= (.0005255) [20 - (x-4)^2]^2 I_{[0,8]}(x). \end{aligned}$$

See figures 4.1 and 4.2.

We calculate  $\sigma_o^2$  and  $\sigma_{cv}^2$  from their definitions for the two intensity functions  $\lambda_1$  and  $\lambda_2$ . The results appear in the following table.

	$\sigma_o^2$	$\sigma_{cv}^2$
$\lambda_1$	.1468	.1787
$\lambda_2$	.5283	.6318

Both variances,  $\sigma_{cv}^2$  and  $\sigma_o^2$ , are bigger for  $\lambda_2$  compared to  $\lambda_1$ . Of the two intensity functions,  $\lambda_1$  has more "curvature" as it rises and falls four times over the interval  $[0,8]$ . Therefore, from the two examples it follows that  $\sigma_{cv}^2$  can be lower when the intensity function



has more curvature. The curvature of a function is often measured by the second derivative of that function. Notice that  $\alpha''$  (which is proportional to  $\lambda''$ ) appears in the the formulas of  $\sigma_{cv}^2$  and  $\sigma_o^2$ . In general, both of the variance functions are the sum of two ratios of measures of curvature (i.e.  $(\int_0^T \alpha^2) / [\int_0^T (\alpha'')^2]^{1/5}$  and  $[\int_0^T (\alpha'')^2 \alpha + (\int_0^T \alpha'' \alpha)^2] / [\int_0^T (\alpha'')^2]^{6/5}$  ).

The fact that  $\sigma_{cv}^2$  is comparatively small when the underlying intensity function has more curvature is very encouraging. As stated earlier, the convergence rate  $c_s^{1/10}$  is disappointingly slow; on the other hand, this slow convergence rate implies that the constants in this situation are more important than usual. Therefore, it is useful to know that the cross-validation bandwidth might perform reasonably well when there exists a fair amount of curvature in the true intensity function.

#### 4.c. Proofs of the Theorems.

In this section, we will present the proofs of the two theorems given in section 4.b. The proof Theorem 4.1 depends on a number of lemmas which are stated and proven in section 4.d. Theorem 4.2 is a direct consequence of Theorem 4.1 and Lemma 4.6.

*proof of Theorem 4.1:*

We will use the lemmas that are found in section 4.d to prove this theorem. We begin by proving the asymptotic distribution of  $(\hat{h}_o - h_o)$ . First using the fact that  $\hat{h}_o$  minimizes the function  $\Lambda(h)$  and second using the Mean-Value Theorem, we get

$$0 = \Delta'(\hat{h}_0) = M'(\hat{h}_0) + D'(\hat{h}_0) = (\hat{h}_0 - h_0) M''(h^*) + D'(\hat{h}_0) \quad (4.3)$$

where  $h^*$  lies between  $h_0$  and  $\hat{h}_0$ . It follows from Lemma 4.3. that for some  $\epsilon > 0$ ,

$$\hat{h}_0 = h_0 + o_p(c_s^{-1/5-\epsilon}) \quad \text{as } s \rightarrow \infty. \quad (4.4)$$

Thus, by letting  $h_1 = h_0$ , Lemma 4.2 implies that

$$D'(\hat{h}_0) = D'(h_0) + o_p(c_s^{-7/10}). \quad (4.5)$$

From Lemma 4.4, we know that  $c_s^{7/10} D'(h_0)$  is asymptotically normal with zero mean and standard deviation  $a_0 a_3 \sigma_0$ . Therefore, it follows that

$$c_s^{7/10} D'(\hat{h}_0) \xrightarrow{D} N(0, (a_0 a_3)^2 \sigma_0^2) \quad \text{as } s \rightarrow \infty. \quad (4.6)$$

Since (4.4) implies that  $h^*/h_0 \xrightarrow{p} 1$  and since  $M''(h_0) \sim a_3 c_s^{-2/5}$ , one can show that

$$M''(h^*) = a_3 c_s^{-2/5} + o_p(c_s^{-2/5}). \quad (4.7)$$

With (4.6), we know that  $D'(\hat{h}_0) = o_p(c_s^{-7/10})$ . Combining this with (4.3) and (4.7), we get

$$-(\hat{h}_0 - h_0) = D'(\hat{h}_0) / M''(h^*) = a_3^{-1} c_s^{2/5} D'(\hat{h}_0) + o_p(c_s^{-3/10}). \quad (4.8)$$

Using the fact that  $h_0 \sim a_0 c_s^{-1/5}$ , gives the result,

$$-c_s^{1/10} (\hat{h}_0 - h_0)/h_0 = c_s^{7/10} (a_0 a_3)^{-1} D'(\hat{h}_0) + o_p(1). \quad (4.9)$$

Consequently, using Slutsky's Theorem with (4.6) and (4.9) leads to the final limit,

$$c_s^{1/10} (\hat{h}_0 - h_0)/h_0 \xrightarrow{D} N(0, \sigma_0^2) \quad \text{as } s \rightarrow \infty.$$

Next, we'll prove the asymptotic distribution of  $(\hat{h}_{cv} - \hat{h}_0)$ .

Again, since  $\hat{h}_{cv}$  minimizes  $CV(h)$ ,

$$\begin{aligned} 0 &= CV'(\hat{h}_{cv}) = M'(\hat{h}_{cv}) + D'(\hat{h}_{cv}) + \delta'(\hat{h}_{cv}) \\ &= (\hat{h}_{cv} - h_0) M''(h^*) + D'(\hat{h}_{cv}) + \delta'(\hat{h}_{cv}) \end{aligned} \quad (4.10)$$

where  $h^*$  lies in between  $\hat{h}_{cv}$  and  $h_0$ . By Lemma 4.3,  $h^*/h_0 \xrightarrow{p} 1$ , and

hence,

$$M''(h^*) = a_3 c_s^{-2/5} + o_p(c_s^{-2/5}) \text{ as } s \rightarrow \infty. \quad (4.11)$$

As above, utilizing Lemma 4.2 and Lemma 4.3 and then utilizing Lemma 4.4 and Lemma 4.5 leads to the result that as  $s \rightarrow \infty$ ,

$$\begin{aligned} D'(\hat{h}_{cv}) + \delta'(\hat{h}_{cv}) &= D'(h_o) + \delta'(h_o) + o_p(c_s^{-7/10}) \\ &= o_p(c_s^{-7/10}). \end{aligned} \quad (4.12)$$

Combining (4.11) and (4.12) with (4.10) gives

$$0 = (\hat{h}_{cv} - h_o) [a_3 c_s^{-2/5} + o_p(c_s^{-2/5})] + o_p(c_s^{-7/10}).$$

This implies that  $(\hat{h}_{cv} - h_o) = o_p(c_s^{-3/10})$ , and hence, it follows from (4.11) that

$$(\hat{h}_{cv} - h_o) M''(h^*) = (\hat{h}_{cv} - h_o) a_3 c_s^{-2/5} + o_p(c_s^{-7/10}). \quad (4.13)$$

Therefore, substituting (4.13) into (4.10) results in,

$$0 = (\hat{h}_{cv} - h_o) a_3 c_s^{-2/5} + D'(h_o) + \delta'(h_o) + o_p(c_s^{-7/10}). \quad (4.14)$$

From (4.5) and (4.8) in the first half of the proof, we can conclude that

$$0 = (\hat{h}_o - h_o) a_3 c_s^{-2/5} + D'(h_o) + o_p(c_s^{-7/10}). \quad (4.15)$$

Subtracting (4.15) from (4.14) gives,

$$0 = (\hat{h}_{cv} - \hat{h}_o) a_3 c_s^{-2/5} + \delta'(h_o) + o_p(c_s^{-7/10}),$$

and consequently,

$$-c_s^{1/10} (\hat{h}_{cv} - \hat{h}_o) / h_o = c_s^{7/10} (a_o a_3)^{-1} \delta'(h_o) + o_p(1). \quad (4.16)$$

By Lemma 4.5, we know that

$$c_s^{7/10} \delta'(h_o) \xrightarrow{D} N(0, (a_o a_3)^2 \sigma_{cv}^2) \text{ as } s \rightarrow \infty. \quad (4.17)$$

Finally, using Slutsky's Theorem with (4.16) and (4.17) gives,

$$c_s^{1/10} (\hat{h}_{cv} - \hat{h}_o) / h_o \xrightarrow{D} N(0, \sigma_{cv}^2) \text{ as } s \rightarrow \infty.$$

This completes the proof of Theorem 4.1. \*\*

proof of Theorem 4.2:

First, consider  $[\Delta(h_o) - \Delta(\hat{h}_o)]$ . Observe that by repeatedly applying the Mean-Value Theorem, we get,

$$\begin{aligned} \Delta(h_o) - \Delta(\hat{h}_o) &= (h_o - \hat{h}_o) \Delta'((h_o - \hat{h}_o)/2) = (h_o - \hat{h}_o) [\Delta'((h_o - \hat{h}_o)/2) - \Delta'(\hat{h}_o)] \\ &= (1/2)(h_o - \hat{h}_o)^2 \Delta''(h^*) \end{aligned} \quad (4.18)$$

where  $h^*$  lies between  $h_o$  and  $\hat{h}_o$ . Since  $h^*/h_o \xrightarrow{p} 0$ , it is immediate from Lemma 4.6 that as  $s \rightarrow \infty$ ,

$$|\Delta''(h^*) - M''(h^*)| = |D''(h^*)| = o_p(c_s^{-2/5}),$$

and hence,

$$\Delta''(h^*) = M''(h^*) + o_p(c_s^{-2/5}) = a_3 c_s^{-2/5} + o_p(c_s^{-2/5}). \quad (4.19)$$

By Theorem 4.1, it directly follows that  $(h_o - \hat{h}_o) = o_p(c_s^{-3/10})$ . Thus, substituting (4.19) into (4.18) yields,

$$\Delta(h_o) - \Delta(\hat{h}_o) = (1/2)(h_o - \hat{h}_o)^2 a_3 c_s^{-2/5} + o_p(c_s^{-1}) \quad (4.20)$$

Finally, by Theorem 4.1, we know that as  $s \rightarrow \infty$ ,

$$c_s^{3/10} (h_o - \hat{h}_o) \xrightarrow{D} N(0, a_o^2 \sigma_o^2). \quad (4.21)$$

Therefore, using Slutsky's Theorem, (4.20) and (4.21) imply that as  $s \rightarrow \infty$ ,

$$c_s \{\Delta(h_o) - \Delta(\hat{h}_o)\} \xrightarrow{D} (a_o^2 a_3 / 2) \sigma_o^2 \chi_1^2 = a_4 \sigma_o^2 \chi_1^2.$$

The proof for  $[\Delta(h_o) - \Delta(\hat{h}_o)]$  is exactly the same as the argument presented above replacing  $h_o$  and  $\sigma_o^2$  by  $\hat{h}_{cv}$  and  $\sigma_{cv}^2$  respectively. Thus, the proof of Theorem 4.2 is finished. \*\*

#### 4.d. The Lemmas and Their Proofs.

Assume throughout this section that the definitions and the three technical assumptions given in section 4.a hold. We will begin by

presenting some material that is useful for the proofs of the six lemmas.

The number of observations,  $N$ , is a Poisson random variable with mean  $c_s$ , and therefore, it can be shown that

$$E[N^m; N>0] = E[N^m] = c_s^m + o(c_s^m) \quad (4.22)$$

$$(N-c_s)/\sqrt{c_s} \xrightarrow{d} N(0,1) \quad \text{as } c_s \rightarrow \infty \quad (4.23)$$

$$(N/c_s) = 1 + O_p(c_s^{-1/2}) \quad \text{as } c_s \rightarrow \infty \quad (4.24)$$

Recall that  $L(z) \equiv -zK'(z)$ . Using integration by parts, one can show that

$$\int L(z)dz = \int K(z)dz = 1, \quad \text{and} \quad (4.25)$$

$$\int z^2 L(z) = 3 \int z^2 K(z)dz = 3k \quad (4.26)$$

Moreover, since  $K(z)$  is symmetric about zero,

$$\int zL(z)dz = \int z^2(-K'(z))dz = 0 \quad (4.27)$$

We will denote  $\hat{\gamma}_h(x) \equiv c_s^{-1} h^{-1} \sum_{i=1}^N L((x-X_i)/h)$ .

It is also useful for us to calculate several derivatives and moments of  $\hat{\alpha}_h(x)$ . Suppose that  $\{Y_1, Y_2, \dots, Y_N\}$  are the unordered observations (i.e. the  $Y_i$ 's are the observations in some random order). In addition, assume that  $Y$  and  $Z$  are independent random variables with density function  $\alpha(x)I_{[0,T]}(x)$ . Then, for  $x \in [0, T]$ ,

$$\begin{aligned} d[\hat{\alpha}_h(x)]/dh &= c_s^{-1} \sum_{i=1}^N [-h^{-1}K'((x-Y)/h) ((x-Y)/h^2) - h^{-2}K((x-Y)/h)] \\ &= c_s^{-1} \sum_{i=1}^N h^{-1} [L_h(x-X_i) - K_h(x-X_i)] \\ d[\hat{\alpha}_h(x)]/dh &= h^{-1} [\hat{\gamma}_h(x) - \hat{\alpha}_h(x)]. \end{aligned} \quad (4.28)$$

Moreover,

$$E[\hat{\alpha}_h(x)] = c_s^{-1} \int K_h(x-y)\lambda(y)dy$$

$$\begin{aligned} E[\hat{\alpha}_h(x)] &= \int K_h(x-y)\alpha(y)dy = E[K_h(x-Y)] \\ &= \alpha(x) + h^2\alpha''(x)[\int u^2K(u)/2] + o(h^2). \end{aligned} \quad (4.29)$$

and,

$$\begin{aligned} E[\hat{\alpha}_h(x)^2] &= E[c_s^{-2} \sum_{i=1}^N K_h^2(x-X_i) + c_s^{-2} \sum_{i \neq j} K_h(x-X_i) K_h(x-X_j)] \\ &= c_s^{-2} \int K_h^2(x-y)\lambda(y)dy + c_s^{-2} \iint K_h(x-y)K_h(x-z)\lambda(y)\lambda(z)dydz \\ E[\hat{\alpha}_h(x)^2] &= c_s^{-1}E[K_h^2(x-Y)] + E^2[K_h(x-Y)]. \end{aligned} \quad (4.30)$$

Given (4.29) and (4.30), one can show that

$$d(E\hat{\alpha}_h)/dh = E[d(K_h(x-Y))/dh] = h^{-1}E\hat{\gamma}_h - h^{-1}E\alpha_h. \quad (4.31)$$

and

$$\begin{aligned} d(E\hat{\alpha}_h^2)/dh &= -2c^{-1}h^{-1}E[K_h^2(x-Y)] + 2c^{-1}h^{-1}E[K_h(x-Y)L_h(x-Y)] \\ &\quad - 2h^{-1}E^2[K_h(x-Y)] + 2h^{-1}E[K_h(x-Y)] E[L_h(x-Y)] \\ d(E\hat{\alpha}_h^2)/dh &= -2h^{-1}\{c^{-1}E[K_h^2(x-Y)] - c^{-1}E[K_h(x-Y)L_h(x-Y)] \\ &\quad + [E\hat{\alpha}_h(x)]^2 - E[\alpha_h(x)]E[\gamma_h(x)]\}. \end{aligned} \quad (4.32)$$

For these lemmas, we are interested in calculating

$$D_1(h) = -(h/2)D'(h) = (-h/2)\Delta'(h) - (-h/2)M'(h).$$

We will find a decomposition of  $D_1(h)$  that will be convenient for the proofs of several lemmas. Define:

$$K_i(x) = K_h(x-Y_i) - E[K_h(x-Y)] \quad \text{and} \quad L_i(x) = L_h(x-Y_i) - E[L_h(x-Y)].$$

Then, since we are using a circular design, we claim that

$$D_1(h) \equiv -(h/2)D'(h) = S_1(h) + S_2(h) + S_3(h) + S_4(h) + S_5(h) + S_6(h) \quad (4.33)$$

where  $S_1 = S_{11} - S_{12}$ ,  $S_2 = S_{21} + S_{22}$ ,  $S_3 = S_{31} - S_{32}$ ,  $S_4 = S_{41} - S_{42}$ ,

$S_5 = S_{51} + S_{52}$ , and  $S_6 = S_{61} - S_{62}$ , and

$$S_{11}(h) = 2c_s^{-2} \sum_{1 \leq i < j \leq N} \int K_i(x) K_j(x) dx$$

$$S_{12}(h) = c_s^{-2} \sum_{1 \leq i < j \leq N} \int \{K_i(x)L_j(x) + L_i(x)K_j(x)\} dx$$

$$\begin{aligned}
S_{21}(h) &= c_s^{-1} \sum_{i=1}^N \int K_i(x) \{2E\hat{\alpha}_h(x) - E\hat{\gamma}_h(x) - \alpha(x)\} dx \\
S_{22}(h) &= c_s^{-1} \sum_{i=1}^N \int L_i(x) \{\alpha(x) - E\hat{\alpha}_h(x)\} dx \\
S_{31}(h) &= c_s^{-2} \sum_{i=1}^N \int K_h^2(x-Y_i) - E[K_h^2(x-Y)] dx \\
S_{32}(h) &= c_s^{-2} \sum_{i=1}^N \int K_h(x-Y_i) L_h(x-Y_i) - E[K_h(x-Y) L_h(x-Y)] dx \\
S_{41}(h) &= \{(N-1)c_s^{-1}-1\} c_s^{-1} \sum_{i=1}^N \int K_h(x-Y_i) E[K_h(x-Y)] \\
&\quad - (1/2) \{K_h(x-Y_i) E[L_h(x-Y)] + L_h(x-Y_i) E[K_h(x-Y)]\} dx \\
S_{42}(h) &= \{N(N-1)c_s^{-2}-1\} \int \{E^2[K_h(x-Y)] - E[K_h(x-Y)] E[L_h(x-Y)]\} dx \\
S_{51}(h) &= (Nc_s^{-1}-1) \int E[K_h(x-Y)] \{2E\hat{\alpha}_h(x) - E\hat{\gamma}_h(x) - \alpha(x)\} dx \\
S_{52}(h) &= (Nc_s^{-1}-1) \int E[L_h(x-Y)] \{\alpha(x) - E\hat{\alpha}_h(x)\} dx \\
S_{61}(h) &= (Nc_s^{-1}-1) c_s^{-1} \int E[K_h^2(x-Y)] dx \\
S_{62}(h) &= (Nc_s^{-1}-1) c_s^{-1} \int E[K_h(x-Y) L_h(x-Y)] dx.
\end{aligned}$$

Now, we must prove the claim (4.33). First, expand  $(-h/2)\Delta'(h)$ .

$$\begin{aligned}
(-h/2)\Delta'(h) &= -h/2 \, d[\int_0^T \hat{\alpha}_h^2 - 2\int_0^T \hat{\alpha}_h \alpha - \int_0^T \alpha^2] / dh \\
&= -h/2 \, [\int_0^T (\hat{\alpha}_h^2)' - 2\int_0^T \hat{\alpha}_h' \alpha] \\
&= -h \, [\int_0^T \hat{\alpha}_h' (\hat{\alpha}_h - \alpha) dx].
\end{aligned}$$

Using (4.28), we get

$$\begin{aligned}
(-h/2)\Delta'(h) &= \int_0^T (\hat{\alpha}_h - \alpha) (\hat{\alpha}_h - \hat{\gamma}_h) dx \\
(-h/2)\Delta'(h) &= \int_0^T (\hat{\alpha}_h - E\hat{\alpha}_h)^2 - \int_0^T (\hat{\alpha}_h - E\hat{\alpha}_h) (\hat{\gamma}_h - E\hat{\gamma}_h) \\
&\quad + \int_0^T (\hat{\alpha}_h - E\hat{\alpha}_h) (2E\hat{\alpha}_h - E\hat{\gamma}_h - \alpha) + \int_0^T (\hat{\gamma}_h - E\hat{\gamma}_h) (\alpha - E\hat{\alpha}_h) \\
&\quad + \int_0^T (E\hat{\alpha}_h - \alpha) (E\hat{\alpha}_h - E\hat{\gamma}_h). \tag{4.34}
\end{aligned}$$

By substituting the definitions of  $\hat{\alpha}_h$  and  $\hat{\gamma}_h$  into  $[\int_0^T (\hat{\alpha}_h - E\hat{\alpha}_h)^2 - \int_0^T (\hat{\alpha}_h - E\hat{\alpha}_h) (\hat{\gamma}_h - E\hat{\gamma}_h)]$  and then expanding these terms as a sum of integrals of squares plus a sum of integrals of products, one

can show that:

$$\begin{aligned} & \int_0^T (\hat{\alpha}_h - E\hat{\alpha}_h)^2 - \int_0^T (\hat{\alpha}_h - E\hat{\alpha}_h)(\hat{\gamma}_h - E\hat{\gamma}_h) \\ &= S_{11} - S_{12} + S_{41} - S_{42} + c_s^{-2} \sum_{i=1}^N \int K_h^2(x-Y_i) dx \\ & \quad - c_s^{-2} \sum_{i=1}^N \int K_h(x-Y_i) L_h(x-Y_i) dx. \end{aligned}$$

Likewise,

$$\int_0^T (\hat{\alpha}_h - E\hat{\alpha}_h)(2E\hat{\alpha}_h - E\hat{\gamma}_h - \alpha) + \int_0^T (\hat{\gamma}_h - E\hat{\gamma}_h)(\alpha - E\hat{\alpha}_h) = S_{21} + S_{22} + S_{51} + S_{52}.$$

Thus, by (4.34),

$$\begin{aligned} (-h/2)\Delta'(h) &= S_1 + S_2 + S_4 + S_5 + c_s^{-2} \sum_{i=1}^N \int K_h^2(x-Y_i) dx \\ & \quad - c_s^{-2} \sum_{i=1}^N \int K_h(x-Y_i) L_h(x-Y_i) dx + \int_0^T (E\hat{\alpha}_h - \alpha)(E\hat{\alpha}_h - E\hat{\gamma}_h). \end{aligned} \quad (4.35)$$

In addition, we need to calculate  $(-h/2)M'(h)$ . First of all,

$$(-h/2)M'(h) = (-h/2) d[\int_0^T E(\hat{\alpha}_h - \alpha)^2] / dh = (-h/2) \int_0^T E(\hat{\alpha}_h^2)' + h \int_0^T E(\hat{\alpha}_h)' \alpha.$$

Secondly, using (4.31) and (4.32), one can show that

$$\begin{aligned} (-h/2)M'(h) &= c_s^{-1} \int E[K_h^2(x-Y)] dx - c_s^{-1} \int E[K_h(x-Y) L_h(x-Y)] dx + \int_0^T (E\hat{\alpha}_h)^2 \\ & \quad - \int_0^T E\hat{\alpha}_h E\hat{\gamma}_h - \int_0^T E\hat{\alpha}_h \alpha + \int_0^T E\hat{\gamma}_h \alpha \\ &= c_s^{-1} \int E[K_h^2(x-Y)] dx - c_s^{-1} \int E[K_h(x-Y) L_h(x-Y)] dx \\ & \quad + \int_0^T (E\hat{\alpha}_h - \alpha)(E\hat{\alpha}_h - E\hat{\gamma}_h). \end{aligned} \quad (4.36)$$

Combining (4.35) and (4.36), we can conclude that (4.33) holds:

$$\begin{aligned} D_1(h) &= (-h/2)D'(h) \\ &= (-h/2)\Delta'(h) - (-h/2)M'(h) \\ &= S_1 + S_2 + S_4 + S_5 + c_s^{-2} \sum_{i=1}^N \int K_h^2(x-Y_i) dx - c_s^{-2} \sum_{i=1}^N \int K_h(x-Y_i) L_h(x-Y_i) dx \\ & \quad + \int_0^T (E\hat{\alpha}_h - \alpha)(E\hat{\alpha}_h - E\hat{\gamma}_h) - \{c_s^{-1} \int E[K_h^2(x-Y)] dx \\ & \quad - c_s^{-1} \int E[K_h(x-Y) L_h(x-Y)] dx + \int_0^T (E\hat{\alpha}_h - \alpha)(E\hat{\alpha}_h - E\hat{\gamma}_h)\} \\ &= S_1 + S_2 + S_3 + S_4 + S_5 + S_6. \end{aligned} \quad (4.37)$$

Furthermore, we will be interested in  $\delta'(h)$ . Define

$$\delta_1(h) \equiv (h/2)\delta'(h).$$

Then,

$$\begin{aligned} \delta_1(h) &= h \, d[\int_0^T \hat{\alpha}_h(x) \alpha_h(x) dx - c_s^{-1} \sum_{i=1}^N \alpha_{hi}(Y_i)]/dh \\ &= h \int_0^T \hat{\alpha}'_h(x) \alpha_h(x) dx - c_s^{-1} \sum_{i=1}^N \alpha'_{hi}(Y_i) \\ &= c_s^{-1} \sum_{i=1}^N \int L_h(x-Y_i) \alpha(x) dx - c_s^{-1} \sum_{i=1}^N \int K_h(x-Y_i) \alpha(x) dx \\ &\quad + c_s^{-2} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N K_h(Y_i-Y_j) - c_s^{-2} \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N L_h(Y_i-Y_j). \end{aligned}$$

Let  $Y$  and  $Z$  be random variables with distribution  $\alpha(x)I_{[0,T]}(x)$ .

Then,  $\delta_1(h)$  can be decomposed such that

$$\delta_1(h) = (h/2)\delta'(h) = T_1(h) + T_2(h) + T_3(h)$$

where  $T_1 = T_{11} - T_{12}$ ,  $T_2 = T_{21} - T_{22}$  and  $T_3 = T_{31} - T_{32}$ , and

$$\begin{aligned} T_{11}(h) &= 2c_s^{-2} \sum_{1 \leq i < j \leq N} \{K_h(Y_i-Y_j) - \int K_h(Y_i-y) \alpha(y) dy - \int K_h(y-Y_j) \alpha(y) dy \\ &\quad + \iint K_h(y-z) \alpha(y) \alpha(z) dy dz\} \end{aligned}$$

$$\begin{aligned} T_{12}(h) &= 2c_s^{-2} \sum_{1 \leq i < j \leq N} \{L_h(Y_i-Y_j) - \int L_h(Y_i-y) \alpha(y) dy - \int L_h(y-Y_j) \alpha(y) dy \\ &\quad + \iint L_h(y-z) \alpha(y) \alpha(z) dy dz\} \end{aligned}$$

$$T_{21}(h) = c_s^{-1} \sum_{i=1}^N \{ \int K_h(Y_i-y) \alpha(y) dy - \iint K_h(y-z) \alpha(y) \alpha(z) dy dz - \alpha(Y_i) + \int_0^T \alpha^2 \}$$

$$T_{22}(h) = c_s^{-1} \sum_{i=1}^N \{ \int L_h(Y_i-y) \alpha(y) dy - \iint L_h(y-z) \alpha(y) \alpha(z) dy dz - \alpha(Y_i) + \int_0^T \alpha^2 \}$$

$$T_{31}(h) = [(N-1)c_s^{-1} - 1] c_s^{-1} \sum_{i=1}^N \{ \int K_h(Y_i-y) \alpha(y) dy - \int L_h(Y_i-y) \alpha(y) dy \}$$

$$\begin{aligned} T_{32}(h) &= [N(N-1)c_s^{-2} - Nc_s^{-1}] [\iint K_h(y-z) \alpha(y) \alpha(z) dy dz \\ &\quad - \iint L_h(y-z) \alpha(y) \alpha(z) dy dz]. \end{aligned}$$

It is convenient to denote

$$D(h) = D^*(h) + R^*(h) \quad \text{and} \quad \delta(h) = \delta^*(h) + \rho^*(h)$$

such that

$$(-h/2) D^{*\prime}(h) \equiv (S_1 + S_2 + S_3)$$

$$(-h/2) R^{*\prime}(h) \equiv (S_4 + S_5 + S_6)$$

$$(-h/2) \delta^{*\prime}(h) \equiv (T_1 + T_2)$$

$$(-h/2) \rho^{*\prime}(h) \equiv T_3.$$

Essentially,  $D^*$  and  $\delta^*$  include the terms that are found in the density estimation case, and  $R^*$  and  $\rho^*$  include the extra terms that are introduced by considering a nonhomogenous Poisson process. Thus, the intensity setting differs from the density setting in two ways. First, the number of observations in  $D^*$  and  $\delta^*$  is now a random variable, and second,  $R^*$  and  $\rho^*$  are new terms that have to be considered. At this stage, we will present and prove the Lemmas that are necessary to prove Theorem 4.1. and Theorem 4.2.

**Lemma 4.1:** For each  $0 < a < b < \infty$  and all positive integers  $\ell$ ,

$$\sup_{s; a \leq t \leq b} E[|c_s^{7/10} D^{\mathbf{x}, (c_s^{-1/5} t)}|^{2\ell} ; N > 0] \leq A_1(a, b, \ell), \quad (1)$$

$$\sup_{s; a \leq t \leq b} E[|c_s^{7/10} \delta^{\mathbf{x}, (c_s^{-1/5} t)}|^{2\ell} ; N > 0] \leq A_1(a, b, \ell). \quad (2)$$

Furthermore  $\exists \epsilon > 0$  such that

$$E[|c_s^{7/10} \{D^{\mathbf{x}, (c_s^{-1/5} r)} - D^{\mathbf{x}, (c_s^{-1/5} t)}\}|^{2\ell} ; N > 0] \leq A_2(a, b, \ell) |r-t|^{\epsilon\ell}, \quad (3)$$

$$E[|c_s^{7/10} \{\delta^{\mathbf{x}, (c_s^{-1/5} r)} - \delta^{\mathbf{x}, (c_s^{-1/5} t)}\}|^{2\ell} ; N > 0] \leq A_2(a, b, \ell) |r-t|^{\epsilon\ell} \quad (4)$$

whenever  $a \leq r \leq t \leq b$ .

proof of Lemma 4.1:

We will present the proofs of parts (2) and (3) of Lemma 4.1. We begin by looking at part (3). Since  $h \sim c_s^{-1/5}$ , in order to prove (3) it is sufficient to prove that for some  $\epsilon > 0$  and constant A:

$$E[|c_s^{9/10} \{S_{11}(c_s^{-1/5} r) - S_{11}(c_s^{-1/5} t)\}|^{2\ell} ; N > 0] \leq A|r-t|^{\epsilon\ell}, \quad (4.38)$$

$$E[|c_s^{9/10} \{S_{12}(c_s^{-1/5} r) - S_{12}(c_s^{-1/5} t)\}|^{2\ell} ; N > 0] \leq A|r-t|^{\epsilon\ell},$$

$$E[|c_s^{9/10} \{S_{21}(c_s^{-1/5} r) - S_{21}(c_s^{-1/5} t)\}|^{2\ell} ; N > 0] \leq A|r-t|^{\epsilon\ell}, \quad (4.39)$$

$$E[|c_s^{9/10} \{S_{22}(c_s^{-1/5} r) - S_{22}(c_s^{-1/5} t)\}|^{2\ell} ; N > 0] \leq A|r-t|^{\epsilon\ell},$$

$$E[|c_s^{13/10} \{S_{31}(c_s^{-1/5} r) - S_{31}(c_s^{-1/5} t)\}|^{2\ell} ; N > 0] \leq A|r-t|^{\epsilon\ell}, \quad (4.40)$$

$$E[|c_s^{13/10} \{S_{32}(c_s^{-1/5} r) - S_{32}(c_s^{-1/5} t)\}|^{2\ell} ; N > 0] \leq A|r-t|^{\epsilon\ell}.$$

In this chapter, we will prove (4.38), (4.39), and (4.40). The proofs of the other inequalities are similar.

Throughout this proof let  $a \leq r, t \leq b$  and let A be a generic constant that depends on K, a, b and  $\ell$ . First, consider (4.38). We can write  $S_{11}$  as

$$S_{11}(c_s^{-1/5} t) = c_s^{-2} \sum_{1 \leq i < j \leq N} U_{ij}(t)$$

where  $E[U_{ij}(t) | Y_i, N] = E[U_{ij}(t) | Y_j, N] = 0$ . Furthermore, let

$$W_{ij}(r, t) = U_{ij}(r) - U_{ij}(t).$$

The left side of (4.38) can be written in terms of  $W_{ij}(s, t)$ :

$$\begin{aligned} & E[|c_s^{9/10} \{S_{11}(c_s^{-1/5} r) - S_{11}(c_s^{-1/5} t)\}|^{2\ell} ; N > 0] \\ &= E[|c_s^{9/10} c_s^{-2} \sum_{1 \leq i < j \leq N} W_{ij}(r, t)|^{2\ell} ; N > 0] \\ &= c_s^{-11\ell/5} E[E[|\sum_{1 \leq i < j \leq N} W_{ij}(r, t)|^{2\ell} | N] ; N > 0]. \end{aligned} \quad (4.41)$$

Consequently, we are interested in finding an upper bound for

$E[|\sum_{1 \leq i < j \leq N} W_{ij}(r, t)|^{2\ell} | N]$ . Given the definition of  $K_i(x)$ , we can expand

$W_{ij}$  such that  $W_{ij}(r, t) = W_1(r, t, Y_i, Y_j) + W_2(r, t, Y_i) + W_3(r, t)$  where

$$W_1(r, t, Y_i, Y_j) \equiv \int K_{(c_s^{-1/5} r)}(x-Y_i) K_{(c_s^{-1/5} r)}(x-Y_j) dx \\ - \int K_{(c_s^{-1/5} t)}(x-Y_i) K_{(c_s^{-1/5} t)}(x-Y_j) dx.$$

$$W_2(r, t, Y_i) \equiv \int K_{(c_s^{-1/5} r)}(x-Y_i) E[K_{(c_s^{-1/5} r)}(x-Y)] dx \\ - \int K_{(c_s^{-1/5} t)}(x-Y_i) E[K_{(c_s^{-1/5} t)}(x-Y)] dx.$$

$$W_3(r, t) \equiv \int E^2[K_{(c_s^{-1/5} r)}(x-Y)] dx - \int E^2[K_{(c_s^{-1/5} t)}(x-Y)] dx.$$

We begin with  $W_1$ . For the first integral, use the substitution  $u=(x-Y_i)/(c_s^{-1/5} r)$  (hence,  $du=(c_s^{-1/5} r)^{-1} dx$ ); and for the second integral, use  $u=(x-Y_i)/(c_s^{-1/5} t)$ . Then,

$$|W_1(r, t, Y_i, Y_j)| = |(c_s^{-1/5} r)^{-2} \int K((x-Y_i)/(c_s^{-1/5} r)) K((x-Y_j)/(c_s^{-1/5} r)) dx \\ - (c_s^{-1/5} t)^{-2} \int K((x-Y_i)/(c_s^{-1/5} t)) K((x-Y_j)/(c_s^{-1/5} t)) dx| \\ = |(c_s^{-1/5} r)^{-1} \int_{-1}^1 K(u) K(u+(Y_i-Y_j)/(c_s^{-1/5} r)) du \\ - (c_s^{-1/5} t)^{-1} \int_{-1}^1 K(u) K(u+(Y_i-Y_j)/(c_s^{-1/5} t)) du| \\ \leq (c_s^{-1/5} r)^{-1} \int_{-1}^1 K(u) |K(u+(Y_i-Y_j)/(c_s^{-1/5} r)) - K(u+(Y_i-Y_j)/(c_s^{-1/5} t))| du \\ + |(c_s^{-1/5} r)^{-1} - (c_s^{-1/5} t)^{-1}| \int_{-1}^1 K(u) K(u+(Y_i-Y_j)/(c_s^{-1/5} t)) du.$$

Since  $K$  is supported on  $[-1, 1]$ ,  $K(u+(Y_i-Y_j)/(c_s^{-1/5} r))$  equals zero when  $|(Y_i-Y_j)/(c_s^{-1/5} b)| > 2$ . In addition,  $K$  is bounded and Holder continuous. Therefore,

$$|W_1(r, t, Y_i, Y_j)| \leq A 1_{[-2, 2]} \{(Y_i-Y_j)/(c_s^{-1/5} b)\} \\ [(c_s^{-1/5} r)^{-1} \int_{-1}^1 |(Y_i-Y_j)/(c_s^{-1/5} r) - (Y_i-Y_j)/(c_s^{-1/5} t)|^{\epsilon_1} du \\ + |(c_s^{-1/5} r)^{-1} - (c_s^{-1/5} t)^{-1}|] \\ \leq A (c_s^{-1/5} b)^{-1} 1_{[-2, 2]} \{(Y_i-Y_j)/(c_s^{-1/5} b)\} \\ [|(r-t) (Y_i-Y_j)/(c_s^{-1/5} rt)|^{\epsilon_1} + (rt)^{-1} |r-t|].$$

Since  $|(Y_i - Y_j)/c_s^{-1/5}| \leq 2b$  and  $(rt)^{-1} \leq a^{-2}$ ,

$$|W_1(r, t, Y_i, Y_j)| \leq A (c_s^{-1/5} b)^{-1} 1_{[-2, 2]} \left\{ \frac{(Y_i - Y_j)}{c_s^{-1/5} b} \right\} \\ \left[ (2b/a^2)^{\epsilon_1} |r-t|^{\epsilon_1} + a^{-2} |r-t| \right] \\ |W_1(r, t, Y_i, Y_j)| \leq A |r-t|^{\epsilon_2} (c_s^{-1/5} b)^{-1} 1_{[-2, 2]} \left\{ \frac{(Y_i - Y_j)}{c_s^{-1/5} b} \right\}. \quad (4.42)$$

Next consider  $W_2$ .

$$|W_2(r, t, Y_i)| = \left| \int K_{(c_s^{-1/5} r)}(x - Y_i) \left[ \int K_{(c_s^{-1/5} r)}(x - y) \alpha(y) dy \right] dx \right. \\ \left. - \int K_{(c_s^{-1/5} t)}(x - Y_i) \left[ \int K_{(c_s^{-1/5} t)}(x - y) \alpha(y) dy \right] dx \right| \\ \leq \int \left| \int K_{(c_s^{-1/5} r)}(x - Y_i) K_{(c_s^{-1/5} r)}(x - y) dx \right. \\ \left. - \int K_{(c_s^{-1/5} t)}(x - Y_i) K_{(c_s^{-1/5} t)}(x - y) dx \right| \alpha(y) dy \\ = \int |W_1(r, t, Y_i, y)| \alpha(y) dy.$$

By (4.42), we get

$$|W_2(r, t, Y_i)| \leq \int A |r-t|^{\epsilon_2} (c_s^{-1/5} b)^{-1} 1_{[-2, 2]} \left\{ \frac{(Y_i - y)}{c_s^{-1/5} b} \right\} \alpha(y) dy.$$

Since  $\alpha(y)$  is bounded,

$$|W_2(r, t, Y_i)| \leq A |r-t|^{\epsilon_2} \int (c_s^{-1/5} b)^{-1} 1_{[-2, 2]} \left\{ \frac{(Y_i - y)}{c_s^{-1/5} b} \right\} dy \\ |W_2(r, t, Y_i)| \leq A |r-t|^{\epsilon_2}. \quad (4.43)$$

Finally, look at  $W_3$ . Using similar methods as those used to obtain (4.43), one can show that

$$|W_3(r, t)| \leq \left| \iiint K_{(c_s^{-1/5} r)}(x - y) K_{(c_s^{-1/5} r)}(x - z) \alpha(y) \alpha(z) dy dz dx \right. \\ \left. - \iiint K_{(c_s^{-1/5} t)}(x - y) K_{(c_s^{-1/5} t)}(x - z) \alpha(y) \alpha(z) dy dz dx \right| \\ \leq \iiint \left| \int K_{(c_s^{-1/5} r)}(x - y) K_{(c_s^{-1/5} r)}(x - z) dx \right. \\ \left. - \int K_{(c_s^{-1/5} t)}(x - y) K_{(c_s^{-1/5} t)}(x - z) dx \right| \alpha(y) \alpha(z) dy dz \\ \leq \iiint A |r-t|^{\epsilon_2} (c_s^{-1/5} b)^{-1} 1_{[-2, 2]} \left\{ \frac{(y - z)}{c_s^{-1/5} b} \right\} \alpha(y) \alpha(z) dy dz$$

Hence,

$$|W_3(r,t)| \leq A|r-t|^{\epsilon_2}. \quad (4.44)$$

Substituting (4.42), (4.43) and (4.44) into the definition of

$W_{ij}(s,t)$  gives

$$\begin{aligned} |W_{ij}(s,t)| &\leq 2[A(c_s^{-1/5} b)^{-1} 1_{[-2,2]} \{(Y_i - Y_j)/c_s^{-1/5} b\} |r-t|^{\epsilon_2} \\ &\quad + 2A|r-t|^{\epsilon_2} + A|r-t|^{\epsilon_2}] \\ |W_{ij}(s,t)| &\leq A|r-t|^{\epsilon_2} [(c_s^{-1/5} b)^{-1} 1_{[-2,2]} \{(Y_i - Y_j)/c_s^{-1/5} b\} + 1]. \end{aligned} \quad (4.45)$$

Observe that, conditional on  $N$ ,  $\{\sum_{1 \leq i < j \leq k} W_{ij}(s,t)\}_{k=1}^N$  is a martingale. As a result, we can utilize a martingale inequality found in Burkholder (1973, p. 40).

$$\begin{aligned} E[|\sum_{1 \leq i < j \leq N} W_{ij}(r,t)|^{2\ell} | N] &\leq E[|\sup_{k=1, \dots, N} \sum_{1 \leq i < j \leq k} W_{ij}(r,t)|^{2\ell} | N] \\ &\leq AE[(\sum_{j=1}^N E[(\sum_{i=1}^{j-1} W_{ij}(r,t))^2 | \{Y_1, \dots, Y_{j-1}\}, N])^\ell | N] \\ &\quad + A \sum_{k=1}^{\infty} E[(\sum_{i=1}^{k-1} W_{kj}(r,t))^{2\ell} | N]. \end{aligned} \quad (4.46)$$

Consider the first term on the right side of the inequality in (4.46).

Observe that

$$\begin{aligned} &E[(\sum_{i=1}^{N-1} W_{iN}(r,t))^2 | Y_1, \dots, Y_{N-1}, N] \\ &= \sum_{i=1}^{N-1} E[W_{iN}^2 | Y_1, \dots, Y_{N-1}, N] \\ &\quad + \sum_{i=1}^{N-1} \sum_{\substack{k=1 \\ k \neq i}}^{N-1} E[W_{iN} | Y_1, \dots, Y_{N-1}, N] E[W_{kN} | Y_1, \dots, Y_{N-1}, N] \\ &= N E[W_{1N}^2 | Y_1, N] + \sum_{j=1}^N \sum_{\substack{k=1 \\ k \neq j}}^N 0. \end{aligned}$$

Using the result in (4.45) gives

$$\begin{aligned}
& E[(\sum_{i=1}^{N-1} W_{iN}(r,t))^2 | \{Y_1, \dots, Y_{N-1}\}, N] \\
& \leq N A |r-t|^{2\epsilon_2} E[(c_s^{-1/5} b)^{-1} 1_{[-2,2]} \{ (Y_1 - Y_N) / c_s^{-1/5} b \} + 1)^2 | Y_1, N] \\
& \leq N A |r-t|^{2\epsilon_2} E[(c_s^{-1/5} b)^{-2} 1_{[-2,2]} \{ (Y_1 - Y_N) / c_s^{-1/5} b \} + 1 | Y_1, N] \\
& \leq N A |r-t|^{2\epsilon_2} [ \int (c_s^{-1/5} b)^{-2} 1_{[-2,2]} \{ (Y_1 - y) / c_s^{-1/5} b \} \alpha(y) dy + 1 ] \\
& \leq N A |r-t|^{2\epsilon_2} [(c_s^{-1/5} b)^{-1} + 1] \\
& \leq A |r-t|^{2\epsilon_2} N c_s^{1/5}. \tag{4.47}
\end{aligned}$$

Then, the entire first term in the inequality of (4.46) equals,

$$\begin{aligned}
& E[(\sum_{j=1}^N E[(\sum_{i=1}^{j-1} W_{ij}(r,t))^2 | \{Y_1, \dots, Y_{j-1}\}, N])^\ell | N] \\
& \leq E[N^\ell (E[(\sum_{i=1}^{N-1} W_{iN}(r,t))^2 | \{Y_1, \dots, Y_{N-1}\}, N])^\ell | N] \\
& \leq E[N^\ell (A |r-t|^{2\epsilon_2} N c_s^{1/5})^\ell | N] \\
& \leq A |r-t|^{2\epsilon_2 \ell} N^{2\ell} c_s^{\ell/5}. \tag{4.48}
\end{aligned}$$

Next, consider the second term in Burkholder's inequality. Using a partitioning argument as in Lemma 2.3 in Section 2.c, we can show that

$$\begin{aligned}
\sum_{k=1}^{\infty} E[(\sum_{i=1}^{k-1} W_{ik}(r,t))^{2\ell} | N] &= E[\sum_{k=1}^N (\sum_{i=1}^{k-1} W_{ik}(r,t))^{2\ell} | N] \\
&\leq AN^{\ell+1} E[W_{12}^{2\ell}(r,t) | N].
\end{aligned}$$

From (4.45), this leads to the statement

$$\begin{aligned}
& \sum_{k=1}^{\infty} E[(\sum_{i=1}^{k-1} W_{ik}(r,t))^{2\ell} | N] \\
& \leq AN^{\ell+1} E[A |r-t|^{2\epsilon_2 \ell} [(c_s^{-1/5} b)^{-1} 1_{[-2,2]} \{ (Y_1 - Y_2) / c_s^{-1/5} b \} + 1]^{2\ell} | N] \\
& \leq A |r-t|^{2\epsilon_2 \ell} N^{\ell+1} [(c_s^{-1/5} b)^{-2\ell+1} \\
& \quad E[(c_s^{-1/5} b)^{-1} 1_{[-2,2]} \{ (Y_i - Y_j) / c_s^{-1/5} b \} | N] + 1] \\
& = A |r-t|^{2\epsilon_2 \ell} N^{\ell+1} [c_s^{2\ell/5 - 1/5} + 1]
\end{aligned}$$

Since  $\ell \geq 1$ ,

$$\sum_{k=1}^{\infty} E[(\sum_{i=1}^{k-1} W_{ik}(r,t))^{2\ell} | N] \leq A|r-t|^{2\epsilon_2\ell} N^{\ell+1} c_s^{2\ell/5-1/5}. \quad (4.49)$$

Thus, substituting (4.48) and (4.49) into (4.46), we get

$$E[|\sum_{1 \leq i < j \leq N} W_{ij}(r,t)|^{2\ell} | N] \leq A|r-t|^{\epsilon\ell} [N^{2\ell} c_s^{\ell/5} + N^{\ell+1} c_s^{2\ell/5-1/5}],$$

and hence, from (4.41), we can conclude that

$$\begin{aligned} & E[|c_s^{9/10} \{S_{11}(c_s^{-1/5} r) - S_{11}(c_s^{-1/5} t)\}|^{2\ell} ; N > 0] \\ & \leq c_s^{-11\ell/5} E[E[|\sum_{1 \leq i < j \leq N} W_{ij}(r,t)|^{2\ell} | N] ; N > 0] \\ & E[|c_s^{9/10} \{S_{11}(c_s^{-1/5} r) - S_{11}(c_s^{-1/5} t)\}|^{2\ell} ; N > 0] \\ & \leq c_s^{-11\ell/5} E[A|r-t|^{\epsilon\ell} [N^{2\ell} c_s^{\ell/5} + N^{\ell+1} c_s^{2\ell/5-1/5}] ; N > 0] \\ & \leq A|r-t|^{\epsilon\ell} [1 + c_s^{-4\ell/5+4/5}] \quad \text{by (4.22)} \\ & \leq A|r-t|^{\epsilon\ell}. \end{aligned}$$

Thus, we have proven (4.38)

Next, we consider (4.39). We can write,

$$S_{21}(c_s^{-1/5} t) = N^{-1} \sum_{i=1}^N V_i(t) \quad \text{where} \quad E[V_i(t) | N] = 0.$$

Let  $B_i(r,t) = V_i(r) - V_i(t)$ . Note that, utilizing Taylor expansion methods along with (4.25), (4.26) and (4.27) gives

$$\begin{aligned} & 2E[\hat{\alpha}_{(c_s^{-1/5} t)}(x)] - E[\hat{\gamma}_{(c_s^{-1/5} t)}(x)] - \alpha(x) \\ & = \int [2K(u)\alpha(x-c_s^{-1/5} tu) - L(u)\alpha(x-c_s^{-1/5} tu)] du - \alpha(x) \\ & = 2\alpha(x) + (c_s^{-1/5} t)^2 \int K(u)u^2 du \alpha''(x) + o((c_s^{-1/5} t)^2) - \alpha(x) \\ & \quad - (c_s^{-1/5} t)^2 \int L(u)u^2 du \alpha''(x)/2 + o((c_s^{-1/5} t)^2) - \alpha(x) \\ & = (c_s^{-1/5} t)^2 \alpha''(x) \int [K(u) + (L(u)/2)] u^2 du + o((c_s^{-1/5} t)^2). \end{aligned}$$

With arguments similar to those used to prove (4.38), one can prove that

$$\begin{aligned}
|B_i(r, t)| \leq & \int \{ |E[\hat{\alpha}_{(c_s^{-1/5} r)}(x)] - E[\hat{\gamma}_{(c_s^{-1/5} r)}(x)] - \alpha(x)| \\
& \{ |K_{(c_s^{-1/5} r)}(x-Y_i) - E[K_{(c_s^{-1/5} r)}(x-Y)] \\
& - K_{(c_s^{-1/5} t)}(x-Y_i) + E[K_{(c_s^{-1/5} t)}(x-Y)] | \} dx \\
& + \int \{ |K_{(c_s^{-1/5} t)}(x-Y_i) - E[K_{(c_s^{-1/5} t)}(x-Y)] | \} \\
& \{ |E[\hat{\alpha}_{(c_s^{-1/5} r)}(x)] - E[\hat{\gamma}_{(c_s^{-1/5} r)}(x)] - \alpha(x) \\
& - E[\hat{\alpha}_{(c_s^{-1/5} t)}(x)] + E[\hat{\gamma}_{(c_s^{-1/5} t)}(x)] + \alpha(x) | \} dx,
\end{aligned}$$

and hence,

$$|B_i(r, t)| \leq A c_s^{-2/5} |r-t|^{\epsilon_1}. \quad (4.50)$$

Again, Burkholder's martingale inequality can be used for the

martingale  $\{ \sum_{i=1}^k B_i(r, t) \}_{k=1}^N$  to show that

$$\begin{aligned}
E[ | \sum_{i=1}^N B_i(r, t) |^{2\ell} | N ] & \leq AE[ ( \sum_{i=1}^N E[B_i^2(r, t)] )^\ell | N ] + A \sum_{i=1}^\infty E[ |B_i(r, t)|^{2\ell} | N ] \\
& \leq AE[ N^\ell (E[B_i^2(r, t)])^\ell | N ] + AE[ \sum_{i=1}^N |B_i(r, t)|^{2\ell} | N ] \\
& \leq AE[ N^\ell (c_s^{-4/5} |r-t|^{2\epsilon_1})^\ell | N ] + AE[ N (c_s^{-2/5} |r-t|^{\epsilon_1})^{2\ell} | N ] \\
& \leq AN^\ell c_s^{-4\ell/5} |r-t|^{2\epsilon_1 \ell} + AN c_s^{-4\ell/5} |r-t|^{2\epsilon_1 \ell} \\
& \leq AN^\ell c_s^{-4\ell/5} |r-t|^{\epsilon \ell}.
\end{aligned} \quad (4.51)$$

As before, the left side of (4.39) can be written in terms of  $B_i(s, t)$ :

$$\begin{aligned}
& E[ | c_s^{9/10} \{ S_{21}(c_s^{-1/5} r) - S_{21}(c_s^{-1/5} t) \} |^{2\ell} ; N > 0 ] \\
& = E[ | c_s^{9/10} c_s^{-1} \sum_{i=1}^N B_i(r, t) |^{2\ell} ; N > 0 ] \\
& = c_s^{-\ell/5} E[ E[ | \sum_{i=1}^N B_i(r, t) |^{2\ell} | N ] ; N > 0 ] \\
& \leq c_s^{-\ell/5} E[ (AN^\ell c_s^{-4\ell/5} |r-t|^{\epsilon \ell}) ; N > 0 ]
\end{aligned}$$

$$\begin{aligned}
E[|c_s^{9/10}\{S_{21}(c_s^{-1/5}r) - S_{21}(c_s^{-1/5}t)\}|^{2\ell}; N>0] \\
&= A|r-t|^{\epsilon\ell} c_s^{-\ell} E[N^\ell; N>0] \\
&\leq A|r-t|^{\epsilon\ell} \quad \text{by (4.22)}.
\end{aligned}$$

Thus, (4.39) has been proven.

Finally, we'll prove (4.40). As before, we represent

$$S_{31}(c_s^{-1/5}t) = c_s^{-2} \sum_{i=1}^N V_i(t) \quad \text{where} \quad E[V_i(t)|N]=0.$$

Let  $B_i(r,t) = V_i(r) - V_i(t)$ . Then,  $B_i(r,t)$  can be written as

$$B_i(r,t) = B_1(r,t,Y_i) + B_2(r,t) \quad \text{where,}$$

$$B_1(r,t,Y_i) \equiv \int_{(c_s^{-1/5}r)}^{K^2} (x-Y_i) dx - \int_{(c_s^{-1/5}t)}^{K^2} (x-Y_i) dx,$$

$$B_2(r,t) \equiv \int E[K_{(c_s^{-1/5}r)}^2 (x-Y)] dx - \int E[K_{(c_s^{-1/5}t)}^2 (x-Y)] dx.$$

First consider  $B_1$ . Using the substitution  $w=(x-Y_i)/(c_s^{-1/5}r)$  for the first integral and  $w=(x-Y_i)/(c_s^{-1/5}t)$  for the second integral, we get

$$\begin{aligned}
|B_1(r,t,Y_i)| &= |(c_s^{-1/5}r)^{-2} \int_{(c_s^{-1/5}r)}^{K^2} ((x-Y_i)/(c_s^{-1/5}r)) dx \\
&\quad - (c_s^{-1/5}t)^{-2} \int_{(c_s^{-1/5}t)}^{K^2} ((x-Y_i)/(c_s^{-1/5}t)) dx| \\
&= |(c_s^{-1/5}r)^{-1} \int_{-1}^1 K^2(w) dw - (c_s^{-1/5}t)^{-1} \int_{-1}^1 K^2(w) dw| \\
&\leq A c_s^{1/5} [|1/r - 1/t| \int_{-1}^1 K^2(w) dx] \\
&\leq A c_s^{1/5} [(rt)^{-1} |r-t| \int_{-1}^1 K^2(w) dx].
\end{aligned}$$

Since  $K$  is bounded and  $(rt)^{-1} \leq a^{-2}$ ,

$$|B_1(r,t,Y_i)| \leq A c_s^{1/5} |r-t|^{\epsilon_2}. \quad (4.52)$$

Likewise,  $|B_2(r,t)| \leq A c_s^{1/5} |r-t|^{\epsilon_2}$ , and hence,

$$|B_i(r,t)| \leq A c_s^{1/5} |r-t|^{\epsilon_2}. \quad (4.53)$$

Now, (4.53) and Burkholder's inequality imply that

$$\begin{aligned}
& E[|c_s^{13/10} \{S_{31a}(c_s^{-1/5} r) - S_{31a}(c_s^{-1/5} t)\}|^{2\ell} ; N > 0] \\
&= E[|c_s^{13/10} c_s^{-2} \sum_{i=1}^N B_i(r, t)|^{2\ell} ; N > 0] \\
&= c_s^{-7\ell/5} E[E[|\sum_{i=1}^N B_i(r, t)|^{2\ell} | N] ; N > 0]. \\
&\leq c_s^{-7\ell/5} E[A|r-t|^{\epsilon\ell} (N^\ell c_s^{2\ell/5} + N c_s^{2\ell/5}) ; N > 0] \\
&\leq A|r-t|^{\epsilon\ell} c_s^{-\ell} E[N^\ell ; N > 0] \\
&\leq A|r-t|^{\epsilon\ell}.
\end{aligned}$$

Therefore, we've proven (4.40). This concludes the proof of part (3) of Lemma 4.1.

Now we prove part (2) of Lemma 4.1. Since  $h \sim c_s^{-1/5}$ , in order to show that part (2) holds, it is sufficient to prove:

$$\sup_{s; a \leq t \leq b} E[|c_s^{9/10} T_{11}(c_s^{-1/5} t)|^{2\ell} ; N > 0] \leq A, \quad (4.54)$$

$$\sup_{s; a \leq t \leq b} E[|c_s^{9/10} T_{12}(c_s^{-1/5} t)|^{2\ell} ; N > 0] \leq A.$$

$$\sup_{s; a \leq t \leq b} E[|c_s^{9/10} T_{21}(c_s^{-1/5} t)|^{2\ell} ; N > 0] \leq A, \quad (4.55)$$

$$\sup_{s; a \leq t \leq b} E[|c_s^{9/10} T_{22}(c_s^{-1/5} t)|^{2\ell} ; N > 0] \leq A.$$

We will present the proofs of (4.54) and (4.55) here. The other parts are similar. We begin with (4.54). Notice that

$$T_{11}(c_s^{-1/5} t) = 2c_s^{-2} \sum_{1 \leq i < j \leq N} W_{ij}(t)$$

where  $E[W_{ij}(t) | Y_i, N] = E[W_{ij}(t) | Y_j, N] = 0$ . Consequently, we can write,

$$\begin{aligned}
& \sup_{s; a \leq t \leq b} E[|c_s^{9/10} T_{11}(c_s^{-1/5} t)|^{2\ell} ; N > 0] \\
&= \sup_{s; a \leq t \leq b} E[|c_s^{9/10} 2c_s^{-2} \sum_{1 \leq i < j \leq N} W_{ij}(t)|^{2\ell} ; N > 0]
\end{aligned}$$

$$\begin{aligned} & \sup_{s; a \leq t \leq b} E[|c_s^{9/10} T_{11}(c_s^{-1/5} t)|^{2\ell} ; N > 0] \\ & = A c_s^{9\ell/5} E[ \sup_{s; a \leq t \leq b} |c_s^{-2} \sum_{1 \leq i < j \leq N} W_{ij}(t)|^{2\ell} ; N > 0]. \end{aligned}$$

We now can use the same argument as seen in Lemma 2.2 from Chapter 2.

That is, one can use Burkholder's inequality to show that for  $h \sim c_s^{-1/5}$  and  $\ell \geq 1$ ,

$$\sup_{s; a \leq t \leq b} E[|c_s^{9/10} T_{11}(c_s^{-1/5} t)|^{2\ell} ; N > 0] \leq A c_s^{9\ell/5} [A(c_s^{-9\ell/5} + c_s^{-13\ell/5 + 4/5})] \leq A.$$

Next, look at (4.55). We can represent

$$T_{21}(c_s^{-1/5} t) = c_s^{-1} \sum_{i=1} V_i(t) \quad \text{where } E[V_i(t) | N] = 0.$$

Thus, we can write,

$$\begin{aligned} & \sup_{s; a \leq t \leq b} E[|c_s^{9/10} T_{21}(c_s^{-1/5} t)|^{2\ell} ; N > 0] \\ & \leq \sup_{s; a \leq t \leq b} E[|c_s^{9/10} c_s^{-1} \sum_{i=1} V_i(t)|^{2\ell} ; N > 0] \\ & \leq c_s^{9\ell/5} E[ \sup_{s; a \leq t \leq b} |c_s^{-1} \sum_{i=1} V_i(t)|^{2\ell} ; N > 0]. \end{aligned}$$

Again using the same argument as in Lemma 2.2 (Chapter 2) for  $h \sim c_s^{-1/5}$ ,

$$\sup_{s; a \leq t \leq b} E[|c_s^{9/10} T_{21}(c_s^{-1/5} t)|^{2\ell} ; N > 0] \leq c_s^{9\ell/5} [A c_s^{-9\ell/5}] \leq A.$$

Hence, we have proved part (2) of the lemma. This completes the proof of Lemma 4.1. \*\*

**Lemma 4.2:** For some  $\epsilon > 0$  and any  $0 < a < b < \infty$ ,

$$\sup_{a \leq t \leq b} \{ |D'(c_s^{-1/5} t)| + |\delta'(c_s^{-1/5} t)| \} = O_p(c_s^{-3/5-\epsilon}) \text{ as } s \rightarrow \infty. \quad (1)$$

Furthermore, for any  $\epsilon_1 > 0$  and any nonrandom  $h_1$  asymptotic to a constant multiple of  $c_s^{-1/5}$ ,

$$\sup_{|t - c_s^{-1/5} h_1| \leq c_s^{-\epsilon_1}} c_s^{7/10} \{ |D'(c_s^{-1/5} t) - D'(h_1)| + |\delta'(c_s^{-1/5} t) - \delta'(h_1)| \} \xrightarrow{P} 0 \text{ as } s \rightarrow \infty. \quad (2)$$

*proof of Lemma 4.2:*

We'll present a proof for the  $D'$  term in part (2) and for the  $\delta'$  term in (1) of Lemma 4.2. Start by considering  $D'(h)$ . Note that

$$D'(h) = D^{*'}(h) + R^{*'}(h).$$

Thus, we can treat the two summands separately. The methods used on  $D^{*'}$  are similar to those in Lemma 3.2 in Hall and Marron (1987).

Since  $K$  and  $L$  are both Hölder continuous and have compact support,

there exists a  $\beta > 0$  sufficiently large such that

$$\sup_{\substack{a \leq r \leq t \leq b \\ |r-t| \leq c_s^{-\beta+1/5}}} |D^{*'}(c_s^{-1/5} r) - D^{*'}(c_s^{-1/5} t)| = O(c_s^{-1}). \quad (4.56)$$

Assume that  $a < \lim_{s \rightarrow \infty} c_s h_1 < b$ . Then, a sequence can be constructed such

that

$$c_s^{1/5} h_1 - c_s^{-\epsilon_1} = t_0 < t_1 < \dots < t_{m-1} \leq c_s^{1/5} h_1 + c_s^{-\epsilon_1} < t_m$$

where  $t_i - t_{i-1} = c_s^{-\beta}$  for each  $i$ . Let

$$\Theta = \{ (t_i, t_j) : 0 < t_i - t_j \leq c_s^{-1/5-\epsilon_1}, i \leq m \}.$$

Note that there are  $(2c_s^{-\epsilon_1}/c_s^{-\beta})$  points  $t_i$ , and for each  $t_i$  there are  $(c_s^{-1/5-\epsilon_1}/c_s^{-\beta})$  points  $t_j$  such that  $(t_i - t_j) \leq c_s^{-1/5-\epsilon_1}$ . Therefore, the

number of pairs in  $\Theta$  has order  $c_s^{-2\epsilon_1-1/5+2\beta}$ . Given any  $\eta > 0$ ,

$$\begin{aligned}
& P\left[ \sup_{(t_i, t_j) \in \Theta} c_s^{7/10} |D^*(c_s^{-1/5} t_i) - D^*(c_s^{-1/5} t_j)| > \eta \right] \\
& \leq P\left[ \sup_{(t_i, t_j) \in \Theta} c_s^{7/10} |D^*(c_s^{-1/5} t_i) - D^*(c_s^{-1/5} t_j)| > \eta; N > 0 \right] \\
& \quad + P[N=0] \\
& \leq \sum_{(t_i, t_j) \in \Theta} E[\eta^{-1} c_s^{7/10} |D^*(c_s^{-1/5} t_i) - D^*(c_s^{-1/5} t_j)|^{2\ell}; N > 0] \\
& \quad + \exp(-c_s).
\end{aligned}$$

Using Lemma 4.1, we can conclude that

$$\begin{aligned}
& P\left[ \sup_{(t_i, t_j) \in \Theta} c_s^{7/10} |D^*(c_s^{-1/5} t_i) - D^*(c_s^{-1/5} t_j)| > \eta \right] \\
& \leq \sum_{(t_i, t_j) \in \Theta} \eta^{-2\ell} A_2(c_s^{-1/5-\epsilon_1})^{\epsilon_2 \ell} + \exp(-c_s) \\
& \leq A \eta^{-2\ell} c_s^{-2\epsilon_1-1/5+2\beta} (c_s^{-1/5-\epsilon_1})^{\epsilon_2 \ell} + \exp(-c_s).
\end{aligned}$$

By allowing  $\ell$  to be sufficiently large, this term converges to zero as  $s \rightarrow \infty$ . Thus,

$$\sup_{(t_i, t_j) \in \Theta} c_s^{7/10} |D^*(c_s^{-1/5} t_i) - D^*(c_s^{-1/5} t_j)| \xrightarrow{P} 0. \quad (4.57)$$

(4.57) together with (4.56) imply that as  $s \rightarrow \infty$ ,

$$\sup_{|t-c_s^{-1/5} h_1| \leq c_s^{-\epsilon_1}} c_s^{7/10} \{|D^*(c_s^{-1/5} t_i) - D^*(c_s^{-1/5} t_j)|\} \xrightarrow{P} 0. \quad (4.58)$$

Now consider  $R^*(h) = (-h/2)^{-1} [S_4(h) + S_5(h) + S_6(h)]$ . Begin by looking at  $S_5$ . We can define

$$S_5(h) = (Nc_s^{-1} - 1) Q(h).$$

Then, as  $s \rightarrow \infty$  (and hence  $h \rightarrow 0$ ),

$$\begin{aligned}
Q(h) &= \int E[K_h(x-Y)] \{2E\hat{\alpha}_h(x) - E\hat{\gamma}_h(x) - \alpha(x)\} dx \\
&\quad + \int E[L_h(x-Y)] \{\alpha(x) - E\hat{\alpha}_h(x)\} dx \\
&= \int_0^T \{\alpha(x) + h^2 \alpha''(x) [\int u^2 K/2] + o(h^2)\} \{h^2 \alpha''(x) [\int u^2 (2K-L)/2] + o(h^2)\} dx \\
&\quad + \int_0^T \{\alpha(x) + h^2 \alpha''(x) [\int u^2 L/2] + o(h^2)\} \{-h^2 \alpha''(x) [\int u^2 K/2] + o(h^2)\} dx \\
&= h^2 (\int_0^T \alpha'' \alpha) [\int u^2 (2K-L)/2 - \int u^2 K/2] + o_p(h^2) \\
Q(h) &= h^2 (\int_0^T \alpha'' \alpha) (-2k) + o_p(h^2). \tag{4.59}
\end{aligned}$$

For any  $t$  such that  $|t - c_s^{-1/5} h_1| \leq c_s^{-\epsilon_1}$ ,

$$|Q(c_s^{-1/5} t) - Q(h_1)| = |c_s^{-2/5} t^2 - h_1^2| (\int_0^T \alpha'' \alpha) (-2k) + o_p(c_s^{-2/5}).$$

Since  $h_1 \sim c_s^{-1/5}$  and  $|t - c_s^{-1/5} h_1| \leq c_s^{-\epsilon_1}$ , we know that  $|t + c_s^{-1/5} h_1| = o_p(1)$ .

In addition,  $\alpha$  and  $\alpha''$  are bounded. Therefore,

$$|Q(c_s^{-1/5} t) - Q(h_1)| \leq A c_s^{-2/5} |t - c_s^{-1/5} h_1| + o_p(c_s^{-2/5}).$$

Moreover,  $|t - c_s^{-1/5} h_1| \leq c_s^{-\epsilon_1}$  implies that

$$c_s^{2/5} [Q(c_s^{-1/5} t) - Q(h_1)] \xrightarrow{P} 0 \text{ as } s \rightarrow \infty. \tag{4.60}$$

By (4.24), we know that

$$c_s^{1/2} \{Nc_s^{-1} - 1\} \xrightarrow{d} N(0,1) \text{ as } s \rightarrow \infty. \tag{4.61}$$

Using Slutsky's theorem on (4.60) and (4.61) gives,

$$c_s^{9/10} \{|S_5(c_s^{-1/5} t) - S_5(h_1)|\} \xrightarrow{P} 0 \text{ as } s \rightarrow \infty.$$

Using the same procedure for  $S_4$  and  $S_6$ ,

$$c_s^{7/10} \{|R^*(c_s^{-1/5} t) - R^*(h_1)|\} \xrightarrow{P} 0 \text{ as } s \rightarrow \infty.$$

Since  $t$  was chosen arbitrarily,

$$\sup_{|t - c_s^{-1/5} h_1| \leq c_s^{-\epsilon_1}} c_s^{7/10} \{|R^*(c_s^{-1/5} t) - R^*(h_1)|\} \xrightarrow{P} 0 \text{ as } s \rightarrow \infty. \tag{4.62}$$

Combining (4.58) and (4.62) gives the main result for  $D'$  in part (2) of

Lemma 4.2.

Next, consider the  $\delta'$  term in part (1) of Lemma 4.2. First, recall that  $\delta(h) = \delta^*(h) + \rho^*(h)$ . Using an argument that involves the partitioning method found in Hall and Marron (1987) and seen above for  $D^*$ , one can show that

$$\sup_{a \leq t \leq b} \{ |\delta^*(c_s^{-1/5}t)| \} = O_p(c_s^{-3/5-\epsilon}) \quad \text{as } s \rightarrow \infty. \quad (4.63)$$

Now, look at  $\rho^*(h) = (-h/2)^{-1} T_3(h)$ . First of all,

$$T_3(h) = [(N-1)c_s^{-1} - 1] B(h) \quad (4.64)$$

where as  $s \rightarrow \infty$ ,

$$\begin{aligned} B(h) &= c_s^{-1} \sum_{i=1}^N \{ \int K_h(x-Y_i) \alpha(x) dx - \int L_h(x-Y_i) \alpha(x) dx \} \\ &\quad - Nc_s^{-1} \{ \int \int K_h(x-y) \alpha(x) \alpha(y) dx dy - \int \int L_h(x-y) \alpha(x) \alpha(y) dx dy \} \\ &= h^2 c_s^{-1} \sum_{i=1}^N \alpha''(Y_i) [\int u^2(K-L)/2] - Nc_s^{-1} h^2 (\int_0^T \alpha'' \alpha) [\int u^2(K-L)/2] \\ &\quad + o_p(h^2) \\ &= h^2 (\int_0^T \alpha'' \alpha) [\int u^2(K-L)/2] - h^2 (\int_0^T \alpha'' \alpha) [\int u^2(K-L)/2] + o_p(h^2) \\ &= o_p(h^2). \end{aligned}$$

Hence, for any  $t$  such that  $0 < a \leq t \leq b < \infty$ ,

$$B(c_s^{-1/5}t) = o_p(c_s^{-2/5}) \quad \text{as } s \rightarrow \infty.$$

Since  $[(N-1)c_s^{-1} - 1]$  is  $O_p(c_s^{-1/2})$ , for some  $\epsilon > 0$ ,

$$\sup_{a \leq t \leq b} \{ |T_3(c_s^{-1/5}t)| \} = O_p(c_s^{-4/5-\epsilon}).$$

Therefore, as  $s \rightarrow \infty$ ,

$$\sup_{a \leq t \leq b} \{ |\rho^*(c_s^{-1/5}t)| \} = O_p(c_s^{-3/5-\epsilon}). \quad (4.65)$$

(4.65) together with (4.63) imply that the main result for  $\delta'$  in part (1) of the lemma holds. This completes Lemma 4.2. \*\*

**Lemma 4.3:** For some  $\epsilon > 0$ ,

$$|\hat{h}_o - h_o| + |\hat{h}_{cv} - h_o| = o_p(c_s^{-3/5-\epsilon}) \text{ as } s \rightarrow \infty.$$

*proof of Lemma 4.3:*

The proof of this lemma is parallel to the proof of Lemma 3.3 in Hall and Marron (1987). Consider first  $|\hat{h}_{cv} - h_o|$ . By definition of  $\hat{h}_{cv}$ ,  $\delta$  and  $D$ ,

$$\begin{aligned} CV'(h_o) &= CV'(h_o) - CV'(\hat{h}_{cv}) \\ &= \Delta'(h_o) - \Delta'(\hat{h}_{cv}) + \delta'(h_o) - \delta'(\hat{h}_{cv}) \\ CV'(h_o) &= M'(h_o) - M'(\hat{h}_{cv}) + D'(h_o) - D'(\hat{h}_{cv}) + \delta'(h_o) - \delta'(\hat{h}_{cv}). \end{aligned}$$

We know from the proof of Theorem 2.2 that  $\hat{h}_{cv}/h_o \xrightarrow{p} 1$ . Hence, by Lemma 4.2,

$$CV'(h_o) = M'(h_o) - M'(\hat{h}_{cv}) + o_p(c_s^{-3/5-\epsilon}) \text{ as } s \rightarrow \infty. \quad (4.66)$$

In addition, using Lemma 4.2 and the fact that  $h_o$  minimizes  $M(h)$ , we get

$$\begin{aligned} CV'(h_o) &= \Delta'(h_o) + \delta'(h_o) \\ &= M'(h_o) + D'(h_o) + \delta'(h_o) \\ &= D'(h_o) + \delta'(h_o) \\ CV'(h_o) &= o_p(c_s^{-3/5-\epsilon}) \text{ as } s \rightarrow \infty. \end{aligned} \quad (4.67)$$

Combining (4.66) and (4.67) implies that

$$o_p(c_s^{-3/5-\epsilon}) = M'(h_o) - M'(\hat{h}_{cv}) = (h_o - \hat{h}_{cv}) M''(h^*) \quad (4.68)$$

when  $h^*$  lies between  $h_o$  and  $\hat{h}_{cv}$ . Since  $h^*/h_o \xrightarrow{p} 1$  and  $M''(h_o) \sim a_3 c_s^{-2/5}$ ,

$$M''(h^*) = a_3 c_s^{-2/5} + o_p(c_s^{-2/5}).$$

Substituting this into (4.68) gives

$$(h_o - \hat{h}_{cv}) = o_p(c_s^{-1/5-\epsilon}) \text{ as } s \rightarrow \infty. \quad (4.69)$$

This is the first part of the Lemma.

Consider now  $|\hat{h}_0 - h_0|$ . In the proof of Theorem 2.2, we saw that  $\hat{h}_0/h_0 \xrightarrow{p} 1$ . Thus, Lemma 4.2 implies that as  $s \rightarrow \infty$ ,

$$\begin{aligned} \Delta'(h_0) &= \Delta'(h_0) - \Delta'(\hat{h}_0) \\ &= M'(h_0) - M'(\hat{h}_0) + D'(h_0) - D'(\hat{h}_0) \\ &= M'(h_0) - M'(\hat{h}_0) + O_p(c_s^{-3/5-\epsilon}). \end{aligned} \quad (4.70)$$

Since  $h_0$  is the minimizer of  $M(h)$ , and then by Lemma 4.2 it follows that

$$\Delta'(h_0) = D'(h_0) = O_p(c_s^{-3/5-\epsilon}). \quad (4.71)$$

Using (4.70) and (4.71), we get

$$O_p(c_s^{-3/5-\epsilon}) = M'(h_0) - M'(\hat{h}_0).$$

As above, this implies that as  $s \rightarrow \infty$

$$(h_0 - \hat{h}_0) = O_p(c_s^{-1/5-\epsilon}). \quad (4.72)$$

Combining (4.69) and (4.72) gives Lemma 4.3. \*\*

**Lemma 4.4:**  $c_s^{7/10} D'(h_0) \xrightarrow{D} N(0, (a_0 a_3)^2 \sigma_0^2)$  as  $s \rightarrow \infty$ .

*proof of Lemma 4.4:*

Recall that  $D_1(h) = -(h/2)D'(h) = S_1 + S_2 + S_3 + S_4 + S_5 + S_6$ . Thus, it is sufficient to prove that

$$c_s^{9/10} D_1(h_0) \xrightarrow{D} N(0, a_4^2 \sigma_0^2). \quad (4.73)$$

We saw in the proof of Lemma 4.1 that  $E[S_3^2(h_0)] = O(c_s^{-13/5})$ , and consequently  $S_3(h_0) = o_p(c_s^{-9/10})$ . Similarly,  $S_6(h_0) = o_p(c_s^{-9/10})$ .

Next consider  $S_4(h_0)$ .

$$S_4(h) = [(N-1)c_s^{-1} - 1] B(h)$$

where

$$\begin{aligned}
B(h) &= c_s^{-1} \sum_{i=1}^N \int \{K_h(x-Y_i)E[K_h(x-Y)] \\
&\quad - (1/2)(K_h(x-Y_i)E[L_h(x-Y)] + (1/2)L_h(x-Y_i)E[K_h(x-Y)])\} dx \\
&\quad - (Nc_s^{-1}+1) \int_0^T \{E^2[K_h(x-Y)] - E[K_h(x-Y)]E[L_h(x-Y)]\} dx + o_p(1) \\
&= h^2 c_s^{-1} \sum_{i=1}^N \alpha''(Y_i) [\int u^2(K-L)/2] \\
&\quad - (Nc_s^{-1}+1) h^2 (\int_0^T \alpha'' \alpha) [\int u^2(K-L)/2] + o_p(h^2) \\
&= h^2 (\int_0^T \alpha'' \alpha) (-2k) - 2h^2 (\int_0^T \alpha'' \alpha) (-2k) + o_p(h^2) \\
B(h) &= 2kh^2 (\int_0^T \alpha'' \alpha) + o_p(h^2).
\end{aligned}$$

Recall that  $S_5(h) = (Nc_s^{-1}-1) Q(h)$  where by (4.59),

$$Q(h) = -2kh^2 (\int_0^T \alpha'' \alpha) + o(h^2).$$

Thus, for  $h \sim c_s^{-1/5}$ ,

$$[S_4(h) + S_5(h)] = (Nc_s^{-1}-1) [0 + o(c_s^{-2/5})].$$

Since  $c_s^{1/2} [Nc_s^{-1}-1]$  is asymptotically  $N(0,1)$ , and  $h_0 \sim a_0 c_s^{-1/5}$ ,

one can show using Slutsky's theorem that  $[S_4(h_0) + S_5(h_0)] = o_p(c_s^{-9/10})$ .

Let  $S_i = S_i(h_0)$  in this proof. Then, in order to prove (4.73) and hence the Lemma, it is sufficient to prove that

$$c_s^{9/10} (S_1 + S_2) \xrightarrow{D} N(0, a_4^2 \sigma_0^2) \text{ as } s \rightarrow \infty. \quad (4.74)$$

Define:

$$\begin{aligned}
\sigma_1^2 &\equiv (2/a_0) (\int_0^T \alpha^2) \int \{ \int K(y+z)[K(z) - L(z)] dz \}^2 dy \\
\sigma_2^2 &\equiv 4a_0^4 k^2 [\int_0^T (\alpha'')^2 \alpha - (\int_0^T \alpha'' \alpha)^2]
\end{aligned}$$

and observe that  $\sigma_0^2 = a_4^{-2} (\sigma_1^2 + \sigma_2^2)$ .

In order to prove (4.74) and hence Lemma 4.4, it is sufficient to prove that conditional on  $N$ ,

$$(c_s^{9/10} S_1, c_s^{9/10} S_2) \xrightarrow{D} (Z_1, Z_2) \text{ as } c_s \rightarrow \infty \quad (4.75)$$

where  $Z_1$  and  $Z_2$  are independent normal variables with zero mean,  $\text{var}(Z_1) = N^2 c_s^{-2} \sigma_1^2$  and  $\text{var}(Z_2) = N c_s^{-1} \sigma_2^2$ . First, we will prove (4.75), and then we will show how (4.75) implies that (4.74) holds.

Conditional on  $N$ ,  $S_1$  and  $S_2$  are similar to the terms found in the density setting. Therefore, the proof of (4.75) is parallel to the proof found in Hall and Marron (1987). It can be shown that conditional on  $N$ ,  $S_1$  and  $S_2$  are uncorrelated. In addition, we can write

$$S_1 = \sum_{i < j} \sum V(Y_i, Y_j) \quad \text{and} \quad S_2 = \sum_{i=1}^N v(Y_i) \quad (4.76)$$

such that  $E[V(Y_i, Y_j) | Y_j, N] = 0$  a.s. for  $j < i$ . Hence, conditional on  $N$ , for any  $a$  and  $b$ ,

$$aS_1 + bS_2 = \sum_{i=1}^N W_i = \sum_{i=1}^N [a \sum_{j < i} V(Y_i, Y_j) + bv(Y_i)]$$

is a martingale with respect to the  $\sigma$ -fields generated by  $\{Y_1, \dots, Y_N\}$ .

The method found in the proof of Theorem 1 in Hall (1984) can be utilized to show that conditional on  $N$ ,

$$(aS_1 + bS_2) \xrightarrow{D} N(0, a^2 \text{var}[S_1 | N] + b^2 \text{var}[S_2 | N]). \quad (4.77)$$

Moreover, we show below that

$$\begin{aligned} c_s^{9/5} \text{var}[S_1 | N] &\rightarrow N^2 c_s^{-2} \sigma_1^2 \\ &= N^2 c_s^{-2} 2a_0^{-1} (\int_0^T \alpha^2) \int [\int K(y+z) \{K(z) - L(z)\} dz]^2 dy. \end{aligned} \quad (4.78)$$

$$c_s^{9/5} \text{var}[S_2 | N] \rightarrow N c_s^{-1} \sigma_2^2 = N c_s^{-1} 4a_0^4 k^2 \{ \int_0^T (\alpha'')^2 \alpha - (\int_0^T \alpha'' \alpha)^2 \}. \quad (4.89)$$

Using the Cramer-Wold device, (4.77), (4.78) and (4.79) imply that (4.75) holds.

At this stage, we will prove (4.78) and (4.79). From Hall and Marron (1987), we can conclude that given observations  $Y_i, Y_j$  from a density  $f(x)$ ,

$$\text{Var}[V(Y_i, Y_j)] = 2h_0^{-1} (\int f^2) \{ \int [\int K(y+z) \{K(z) - L(z)\} dz]^2 dy \} + o(h_0^{-1}) \quad (4.80)$$

$$\text{Var}[v(Y_i)] = 4h_0^4 k^2 \{ \int (f'')^2 f - (\int f'' f)^2 \} + o(h_0^4) \quad (4.81)$$

where  $V(\cdot)$  and  $v(\cdot)$  are defined as in (4.76). Moreover, in the intensity estimation situation, we know that conditional on  $N$ , the  $Y_i$ 's are distributed with density  $\alpha(x)I_{[0, T]}(x)$ . Now,

$$\text{Var}[S_1 | N] = \text{Var}[c_s^{-2} \sum_{i < j} V(Y_i, Y_j) | N] = c_s^{-4} (N^2 - N) \text{Var}[V(Y_i, Y_j) | N].$$

Thus, by (4.80),

$$\begin{aligned} \text{Var}[S_1 | N] &= c_s^{-4} (N^2 - N) [2h_0^{-1} (\int_0^T \alpha^2) \{ \int [\int K(y+z) \{K(z) - L(z)\} dz]^2 dy \} + o(h_0^{-1})] \\ &= c_s^{-2} h_0^{-1} N^2 c_s^{-2} 2(\int_0^T \alpha^2) \{ \int [\int K(y+z) \{K(z) - L(z)\} dz]^2 dy \} + o(c_s^{-2} h_0^{-1}). \end{aligned}$$

This means that

$$\text{Var}[S_1 | N] \rightarrow c_s^{-9/5} N^2 c_s^{-2} 2a_0^{-1} (\int_0^T \alpha^2) \{ \int [\int K(y+z) \{K(z) - L(z)\} dz]^2 dy \}.$$

In addition, (4.81) leads to

$$\begin{aligned} \text{Var}[S_2 | N] &= \text{Var}[c_s^{-1} \sum_i v(Y_i) | N] \\ &= c_s^{-2} N \text{Var}[v(Y_i) | N] \\ &= c_s^{-2} N [4h_0^4 k^2 \{ \int_0^T (\alpha'')^2 \alpha - (\int_0^T \alpha'' \alpha)^2 \} + o(h_0^4)] \\ &= c_s^{-1} h_0^4 N c_s^{-1} 4k^2 \{ \int_0^T (\alpha'')^2 \alpha - (\int_0^T \alpha'' \alpha)^2 \} + o(c_s^{-1} h_0^4). \\ \text{Var}[S_2 | N] &\rightarrow c_s^{-9/5} N c_s^{-1} 4k^2 a_0^4 \{ \int_0^T (\alpha'')^2 \alpha - (\int_0^T \alpha'' \alpha)^2 \}. \end{aligned}$$

The limits of  $\text{Var}[S_1 | N]$  and  $\text{Var}[S_2 | N]$  above confirm that (4.78) and (4.79) hold.

Finally, given that (4.75) is true, we will prove (4.74) and hence the main result of Lemma 4.4. Let  $\bar{\Phi}(\cdot)$  be the standard normal cumulative distribution function, and let  $\varphi(x, a)$  be the normal density function at  $x$  with mean zero and standard deviation  $a$ . Recall that  $s \rightarrow \infty$  implies that  $c_s \rightarrow \infty$ . Then,

$$\begin{aligned} \lim_{s \rightarrow \infty} P[c_s^{9/10}(S_1 + S_2) \leq x] &= \lim_{s \rightarrow \infty} P[(S_1 + S_2) \leq c_s^{-9/10} x] \\ &= \lim_{s \rightarrow \infty} \left\{ \sum_{n=0}^{\infty} P[(S_1 + S_2) \leq c_s^{-9/10} x \mid N=n] P[N=n] \right\}. \end{aligned}$$

Since the summands are positive numbers, we can move the limit inside of the summation.

$$\begin{aligned} \lim_{s \rightarrow \infty} P[c_s^{9/10}(S_1 + S_2) \leq x] &= \sum_{n=0}^{\infty} \lim_{s \rightarrow \infty} \{ P[(S_1 + S_2) \leq c_s^{-9/10} x \mid N=n] P[N=n] \} \\ &= \sum_{n=0}^{\infty} [ \lim_{s \rightarrow \infty} \{ P[(S_1 + S_2) \leq c_s^{-9/10} x \mid N=n] \} \lim_{s \rightarrow \infty} \{ P[N=n] \} ]. \end{aligned}$$

By (4.76), conditional on  $N$ ,  $(S_1 + S_2)$  is asymptotically normal with variance  $[c_s^{-9/5} (N^2 c_s^{-2} \sigma_1^2 + N c_s^{-1} \sigma_2^2)]$ . Moreover, by (4.25),  $N$  is asymptotically normal with mean  $c_s$  and variance  $c_s$ . Hence,

$$\begin{aligned} \lim_{s \rightarrow \infty} P[c_s^{9/10}(S_1 + S_2) \leq x] &= \sum_{n=0}^{\infty} \left[ \lim_{s \rightarrow \infty} \left\{ \Phi \left[ \frac{c_s^{-9/10} x}{c_s^{-9/10} (n^2 c_s^{-2} \sigma_1^2 + n c_s^{-1} \sigma_2^2)^{1/2}} \right] \right\} \right. \\ &\quad \left. \lim_{s \rightarrow \infty} \left\{ \Phi \left[ \frac{(n+.5) - c_s}{c_s^{1/2}} \right] - \Phi \left[ \frac{(n-.5) - c_s}{c_s^{1/2}} \right] \right\} \right] \\ &= \lim_{s \rightarrow \infty} \left\{ \sum_{n=0}^{\infty} \left[ \Phi \left[ \frac{c_s^{-9/10} x}{c_s^{-9/10} (n^2 c_s^{-2} \sigma_1^2 + n c_s^{-1} \sigma_2^2)^{1/2}} \right] \right. \right. \\ &\quad \left. \left. \left[ \Phi \left[ \frac{(n+.5) - c_s}{c_s^{1/2}} \right] - \Phi \left[ \frac{(n-.5) - c_s}{c_s^{1/2}} \right] \right] \right] \right\} \\ &= \lim_{s \rightarrow \infty} \left\{ \int \Phi \left[ \frac{c_s^{-9/10} x}{c_s^{-9/10} (n^2 c_s^{-2} \sigma_1^2 + n c_s^{-1} \sigma_2^2)^{1/2}} \right] \frac{1}{\sqrt{2\pi}} e^{-(n-c_s)^2/2c_s} dn \right\} \\ &= \lim_{s \rightarrow \infty} \left\{ \int \Phi \left[ \frac{x}{(n^2 c_s^{-2} \sigma_1^2 + n c_s^{-1} \sigma_2^2)^{1/2}} \right] \frac{1}{\sqrt{2\pi}} e^{-(n-c_s)^2/2c_s} dn \right\}. \end{aligned}$$

Using the substitution  $w=c_s^{-1/2}(n-c_s)$  (i.e.  $nc_s^{-1}=[wc_s^{-1/2}+1]$ ), we get

$$\begin{aligned} & \lim_{s \rightarrow \infty} P[-c_s^{9/10}(S_1+S_2) \leq x] \\ &= \lim_{s \rightarrow \infty} \left\{ \int \Phi \left[ \frac{x}{((wc_s^{-1/2}+1)^2 \sigma_1^2 + (wc_s^{-1/2}+1)^2 \sigma_2^2)^{1/2}} \right] \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw \right\} \quad (4.82) \end{aligned}$$

(4.82) implies that the asymptotic cumulative distribution function of  $(S_1+S_2)$  is the limit of a weighted normal cumulative distribution function with normal weights. We show below that this weighted function actually is a normal distribution function. First, by the monotone convergence theorem, it follows that

$$\begin{aligned} & \lim_{s \rightarrow \infty} P[-c_s^{9/10}(S_1+S_2) \leq x] \\ &= \int \lim_{s \rightarrow \infty} \left\{ \Phi \left[ \frac{x}{((wc_s^{-1/2}+1)^2 \sigma_1^2 + (wc_s^{-1/2}+1)^2 \sigma_2^2)^{1/2}} \right] \right\} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} dw \\ &= \int \lim_{s \rightarrow \infty} \left\{ \Phi \left[ \frac{x}{((wc_s^{-1/2}+1)^2 \sigma_1^2 + (wc_s^{-1/2}+1)^2 \sigma_2^2)^{1/2}} \right] \right\} \varphi(w,1) dw. \end{aligned}$$

Hence,

$$\begin{aligned} \lim_{s \rightarrow \infty} P[c_s^{9/10}(S_1+S_2) \leq x] &= \int \Phi \left[ \frac{x}{(\sigma_1^2 + \sigma_2^2)^{1/2}} \right] \varphi(w,1) dw \\ &= \Phi \left[ \frac{x}{(\sigma_1^2 + \sigma_2^2)^{1/2}} \right] \\ &= \Phi \left[ \frac{x}{a_4 \sigma_0} \right]. \quad (4.83) \end{aligned}$$

Thus,  $c_s^{9/10}(S_1+S_2) \xrightarrow{D} N(0, a_4^2 \sigma_0^2)$ , and this completes the proof of the Lemma 4.4. \*\*\*

**Lemma 4.5:**  $c_s^{7/10} \delta'(h_0) \xrightarrow{D} N(0, (a_0 a_3)^2 \sigma_{cu}^2)$  as  $s \rightarrow \infty$ .

proof of Lemma 4.5:

First, define:

$$\begin{aligned}\delta_1(h) &\equiv -(h/2)\delta'(h) = T_1 + T_2 + T_3 \\ \sigma_1^2 &\equiv (2/a_0) (J_0^T \alpha^2) (JL^2) \\ \sigma_2^2 &\equiv 4a_0^4 k^2 [J_0^T (\alpha'')^2 \alpha - (J_0^T \alpha'' \alpha)^2]\end{aligned}$$

and observe that  $\sigma_{cv}^2 \equiv a_4^{-2}(\sigma_1^2 + \sigma_2^2)$ .

Consider  $T_3$ . Recall from (4.64) that,

$$T_3(h_0) = [(N-1)c_s^{-1} - 1] W(h_0)$$

where  $W(h_0) = o(h^2)$ . Thus, one can show that  $T_3(h_0) = o_p(c_s^{-9/10})$ .

Therefore, in order to prove Lemma 4.5, it is sufficient to show that

$$c_s^{9/10} (T_1 + T_2) \xrightarrow{D} N(0, a_4^2 \sigma_{cv}^2) \quad (4.84)$$

Given  $N$ ,  $T_1$  and  $T_2$  are terms that are found in the density setting. Therefore, we use procedure similar to the one seen in Lemma 4.4 for  $T_1$  and  $T_2$ . The density setting result of Lemma 3.5 in Hall and Marron (1987) is utilized to show that conditional on  $N$ ,  $aT_1 + bT_2$  is asymptotically normal and

$$\begin{aligned}c_s^{9/5} \text{var}(T_1|N) &\rightarrow N^2 c_s^{-2} \sigma_1^2 = N^2 c_s^{-2} 2a_0^{-1} (J_0^T \alpha^2) JL^2, \\ c_s^{9/5} \text{var}(T_2|N) &\rightarrow Nc_s^{-1} \sigma_2^2 = Nc_s^{-1} 4a_0^4 k^2 \{J_0^T (\alpha'')^2 \alpha - (J_0^T \alpha'' \alpha)^2\}.\end{aligned}$$

Then, the Cramer-Wold device implies that conditional on  $N$ ,

$$(c_s^{9/10} T_1, c_s^{9/10} T_2) \xrightarrow{D} (Z_1, Z_2) \text{ as } c_s \rightarrow \infty \quad (4.85)$$

where  $Z_1$  and  $Z_2$  are independent normal variables with zero mean,

$$\text{var}(Z_1) = N^2 c_s^{-2} \sigma_1^2 \text{ and } \text{var}(Z_2) = Nc_s^{-1} \sigma_2^2.$$

Statement (4.84) and hence Lemma 4.5 follow from (4.85) using a similar argument to the one found in Lemma 4.4. \*\*

**Lemma 4.6:** Given assumption 4) as well as assumptions 1), 2) and 3),  
for any  $0 < a < b < \infty$ ,

$$\sup_{a \leq t \leq b} |D''(c_s^{-1/5} t)| = o_p(c_s^{-2/5}) \text{ as } s \rightarrow \infty.$$

*proof of Lemma 4.6:*

We begin by creating a decomposition of  $D_2(h) \equiv (h^2/2)D''(h)$ .

Then, as in Lemma 4.1, we can prove that

$$\sup_{s; a \leq t \leq b} E[|c_s^{1/2} D^{*''}(c_s^{-1/5} t)|^{2\ell}; N > 0] \leq A(a, b, \ell).$$

Moreover, using the same methods as those found in Lemma 4.2, it can be shown that for some  $\epsilon > 0$ ,

$$\begin{aligned} \sup_{a \leq t \leq b} |D^{*''}(c_s^{-1/5} t)| &= o_p(c_s^{-2/5-\epsilon}) \text{ as } s \rightarrow \infty, \\ \sup_{a \leq t \leq b} |R^{*''}(c_s^{-1/5} t)| &= o_p(c_s^{-2/5-\epsilon}) \text{ as } s \rightarrow \infty, \end{aligned}$$

and hence,

$$\sup_{a \leq t \leq b} |D''(c_s^{-1/5} t)| = o_p(c_s^{-2/5}) \text{ as } s \rightarrow \infty.$$

This completes the proof of Lemma 4.6. \*\*

CHAPTER 5  
AN EXAMPLE OF KERNEL INTENSITY ESTIMATION WITH THE LEAST-SQUARES  
CROSS-VALIDATION BANDWIDTH APPLIED TO COFFEE SALES DATA

In general, there are two ways of viewing kernel estimation. First of all, our goal might be to estimate a "true" underlying intensity function. Certainly, this goal motivates the theory that was presented in the previous chapters. The alternative is to view the kernel estimator as an exploratory data analysis tool. That is, we use smoothing techniques in order to see important features in data. This is typically the big advantage for using kernel estimation in practice.

We saw in Chapters 2 and 4 that the least-squares cross-validation bandwidth has desirable asymptotic properties. Asymptotic analysis helps us see the basic structure of an estimator; however, since data sets always have a finite number of observations, it is essential that an estimator also performs reasonably well for moderately large data sets. We will show how the kernel intensity estimator with the cross-validation bandwidth can be utilized to analyze data by presenting an actual example with coffee sales purchase data.

*5.a. Kernel Estimation of the Coffee Data.*

In this chapter, we aim to describe the rate of purchases for various brands of coffee. In particular, we are interested in the

sales of regular coffee (i.e. not decaffeinated and not instant) in standard sized packages (between 13 and 32 ounces) at one given store. The data were obtained from the Academic Research Data Base at Information Resources Inc. Two thousand families were observed over a period of one hundred and eight weeks (April, 1980 through April, 1982). Each time a participating family member purchased coffee, the time of the sale and the coffee brand were recorded. Therefore, each data set represents the sales of one brand of coffee at one store, and the observations are the ordered times that the particular brand was purchased by any of the participants in the study.

For a process such as the sales of coffee, we do not expect the rate of observations to be constant over time. For instance, we expect coupons, price changes and advertising campaigns to have an effect on sales. Thus, we are interested in presenting a curve which describes the purchase rate for a given brand of coffee. If we assume that the purchase times form a nonhomogeneous Poisson process, then the intensity function describes the rate of coffee purchases over the relevant time. In addition, we are interested in observing the relationship between sales and marketing efforts. By overlaying the promotional activities above the intensity estimate, we hope to explore the connection between promotions and sales.

A kernel estimator of the intensity function is appealing since it does not restrict the researcher to any prespecified family of functions. This method allows one to see the existing structure in a data set since the kernel estimate is high at times when there are many observations and is low where there are few observations. In the

purchase rate context, the area under the curve in the interval  $[r,s]$  is approximately the number of coffee sales that occurred between time  $r$  and time  $s$ .

Let  $t$  be time measured in weeks; then, for this example we are interested in the interval  $[0,T]$  where  $T=108$ . Let  $N$  be the number of purchases and  $X_1, X_2, \dots, X_N$  be the purchase times that we observe. The kernel smoothed purchase rate is:

$$\hat{\lambda}_h(t) = \sum_{i=1}^N K_h(t-X_i) \quad \text{for } t \in [0,T] \quad (5.1)$$

where  $h$  is the bandwidth or smoothing parameter and  $K(\cdot)$  is the kernel function.

Since we are estimating an intensity function on a finite interval it is important to consider boundary effects. In particular, there are no observations outside of the interval  $[0,T]$ , and hence, the kernel estimator is artificially low at the endpoints. In fact, if  $\lambda(0)$  and  $\lambda(T)$  are positive, it is known that  $\hat{\lambda}_h$  is inconsistent at the two endpoints. This is parallel to the problem in density estimation when there are discontinuities in the density function. Several solutions have been studied by Rice (1984), Gasser, Müller and Marrwitzsch (1985), Schuster (1985) and Cline and Hart (1986). Furthermore, when boundary corrections are not used, the kernel intensity estimate has mass outside of the relevant interval, and hence,  $\int_0^T \hat{\lambda}_h(t) dt$  is smaller than the number of observations in  $[0,T]$ .

In this chapter, we use the kernel estimator with "mirror image" corrections at the two endpoints which was presented in Schuster (1985) and Cline and Hart (1986). With this method, the area from the kernel functions that extends beyond the boundaries of the interval are folded

over at 0 and T so that all of their area is inside of the interval [0,T]. Therefore, the resulting kernel estimator with boundary adjustments is defined by

$$\hat{\lambda}_h^*(t) = \sum_{i=1}^N [K_h(t-X_i) + K_h(t+X_i) + K_h(2T-t-X_i)] \text{ for } t \in [0,T]. \quad (5.2)$$

Thus, when we use the kernel estimator with end corrections,  $\int_0^T \hat{\lambda}_h^*(t) dt$  is equal to the number of observations in the interval [0,T].

The kernel function  $K(\cdot)$  in (5.2) is generally a symmetric probability density function. In our analysis, for computational simplicity we use the biweight kernel:

$$K(x) = \begin{cases} 0.9375(1-x^2)^2 & \text{for } -1 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

With kernel smoothing, the main question is not what shape should the kernel be, but rather how wide should the kernel be. The smoothing parameter,  $h$ , quantifies the width of the kernel. Thus, choosing the smoothing parameter for a kernel estimator is an important concern.

Consider figures 5.1, 5.2 and 5.3. These are kernel intensity estimates of the purchase rate for the coffee sales of Brand 4 with three different bandwidths. The vertical axis represents the number of sales per week, and this brand has a total of 1506 sales between week 0 and week 108. In figure 5.1,  $h=2.82$  weeks is used. This kernel estimate displays a number of obvious peaks, but the curve is fairly smooth over the range of the 108 weeks. Figure 5.2 shows the curve obtained from using a very small bandwidth ( $h=.75$ ). The purchase rate curve has many sharp increases and decreases. These peaks occur precisely when there are a relatively large number of sales in a small time period; however, we believe that this is an undersmoothed version

Figure 5.1

Brand 4,  $h=2.82$

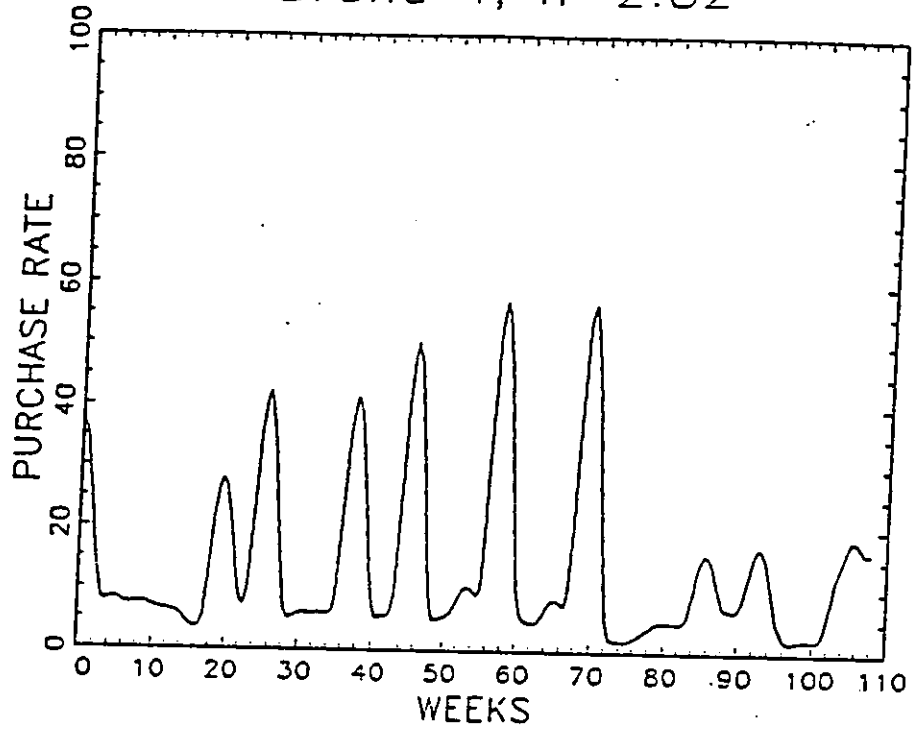


Figure 5.2

Brand 4,  $h=.75$

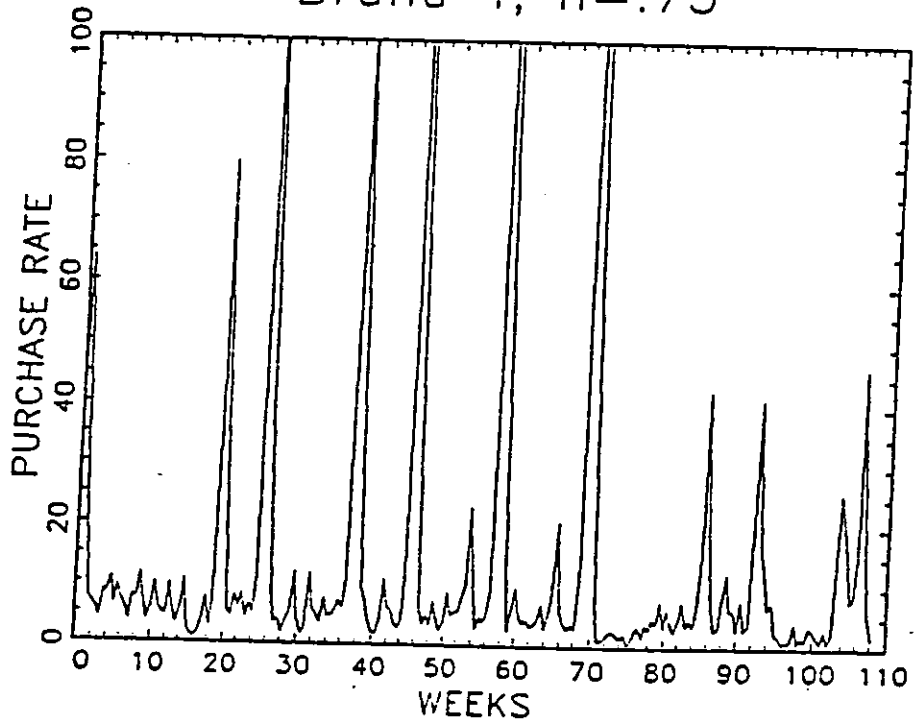
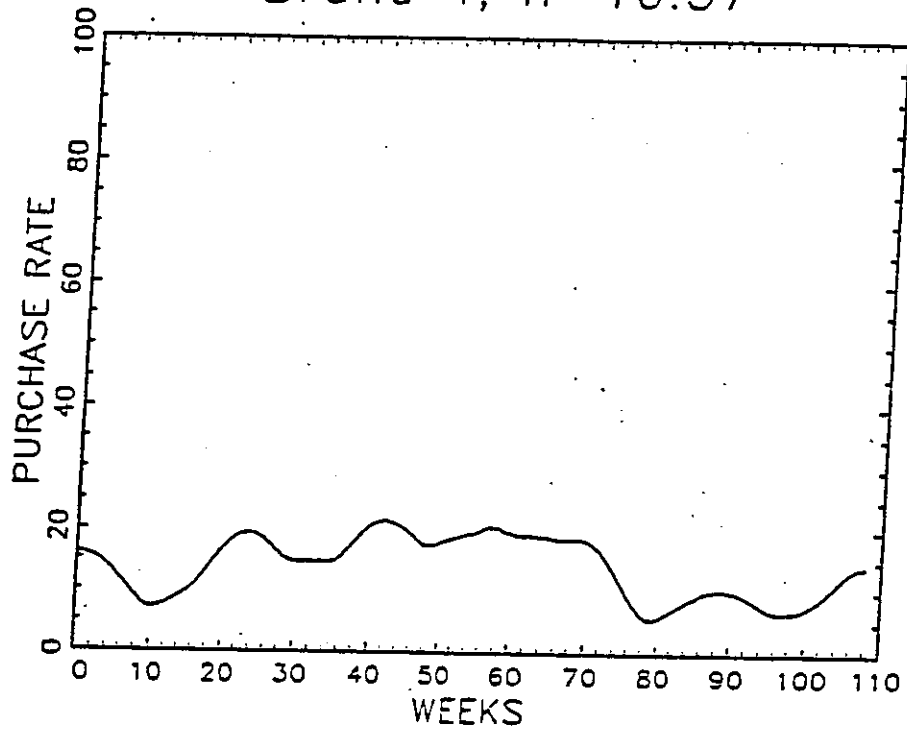


Figure 5.3

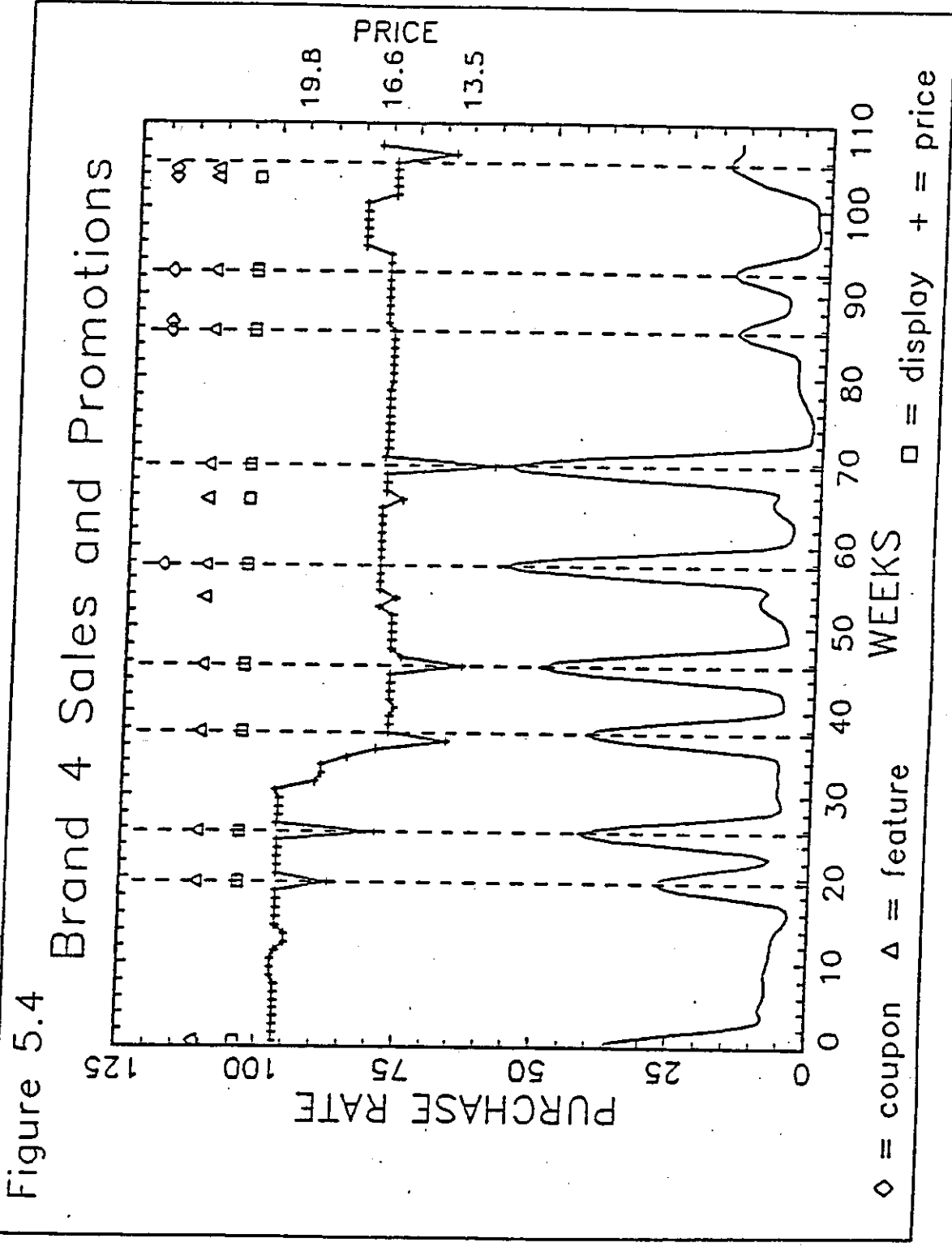
Brand 4,  $h=10.57$



of the purchase rate since the little peaks probably are due to noise in the data set. For example, consider the intensity estimate from week 28 to 35 and from week 48 to 56. Finally, in figure 5.3, a large bandwidth is used ( $h=10.57$ ). With this oversmoothed curve, we lose some information about the purchase rate. Notice how increasing the bandwidth results in one mild peak where there had been two dramatic peaks in figure 5.1 (e.g. near weeks 23 and 43).

In general, small bandwidths allow one to pinpoint exactly when the purchase rate is high and low for each brand, however, it is important to keep in mind that these pictures also display changes in the rates that may be due to random variation rather than identifiable causes. Alternatively, with large bandwidths, the resulting curve may summarize the purchase rate, but this might be an oversimplified view. It is crucial, therefore, to choose a bandwidth so that the resulting curve displays the important features of the data without letting the random variation of the data set effect the curve disproportionately.

For Brand 4, the middle bandwidth,  $h=2.82$ , might be a reasonable choice. Fortunately, we also have information about several covariates that may influence the rate of coffee sales. For each brand, we have weekly data that gives the average price of the coffee for that week and indicates what type or types of marketing efforts were employed during the week. See figure 5.4. First, we overlay the price data; the "+" signs indicate the average price per ounce of Brand 4 coffee. It appears that decreasing the price does effect the purchase rate; however, several of the peaks are not explained by price changes. We also include the information regarding three additional covariates. A



diamond is printed at each week that Brand 4 coupons were distributed; a triangle marks the weeks when the brand was advertised in the local circular; and a square is placed at the weeks when the brand was on "special display" (i.e. either the coffee was displayed on a rack at the end of the store aisle or else there was a big sign advertising the coffee). We can now see that most of these large peaks are associated with marketing efforts. Thus, the purchase rate kernel intensity estimate, allows a researcher to decide whether a promotion influenced the sales rate, and she can determine the size of the effect since the area under a curve is approximately equal to the number of purchases in that period.

#### 5.b. *The Least-Squares Cross-Validation Bandwidth.*

We would like to choose the bandwidth,  $h$ , such that the resulting kernel intensity estimator reflects the important features (and only those) of the data. The cross-validation bandwidth selection method can be motivated by thinking about choosing the bandwidth to minimize the distance between the kernel smooth and the "true" purchase rate function, denoted by  $\lambda(t)$ . In this context,  $\lambda(t)$  is the noiseless ideal purchase rate underlying the sampling that is used in this study. In order to quantify the error of  $\hat{\lambda}_h^*$ , we use the integrated square error (ISE) which is defined by

$$ISE(h) = \int_0^T [\hat{\lambda}_h^*(t)]^2 dt - 2 \int_0^T \hat{\lambda}_h^*(t) \lambda(t) dt + \int_0^T \lambda^2(t) dt$$

We would like to find the bandwidth that minimizes  $ISE(h)$ , but since  $\lambda$  is unknown, this is not possible. In Chapter 2, we saw that the cross-validation score function has roughly the same behavior as

ISE(h). Thus, we choose h by minimizing the cross-validation score function:

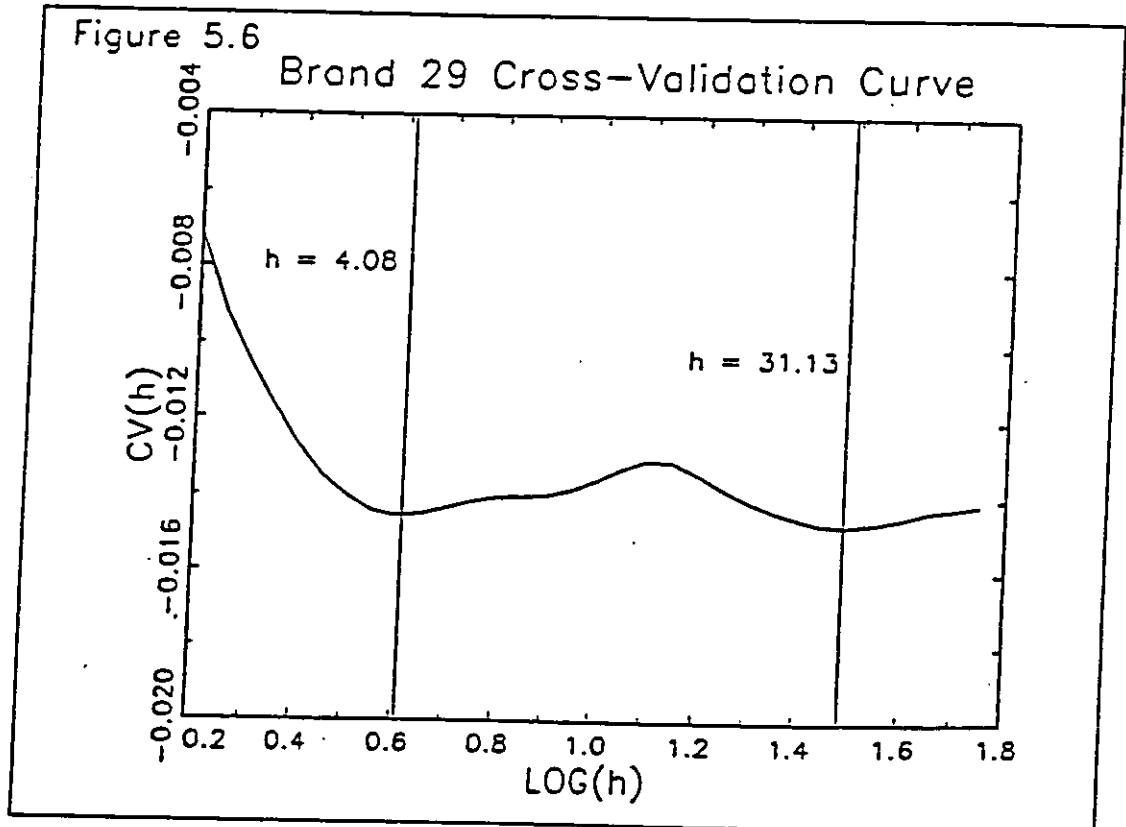
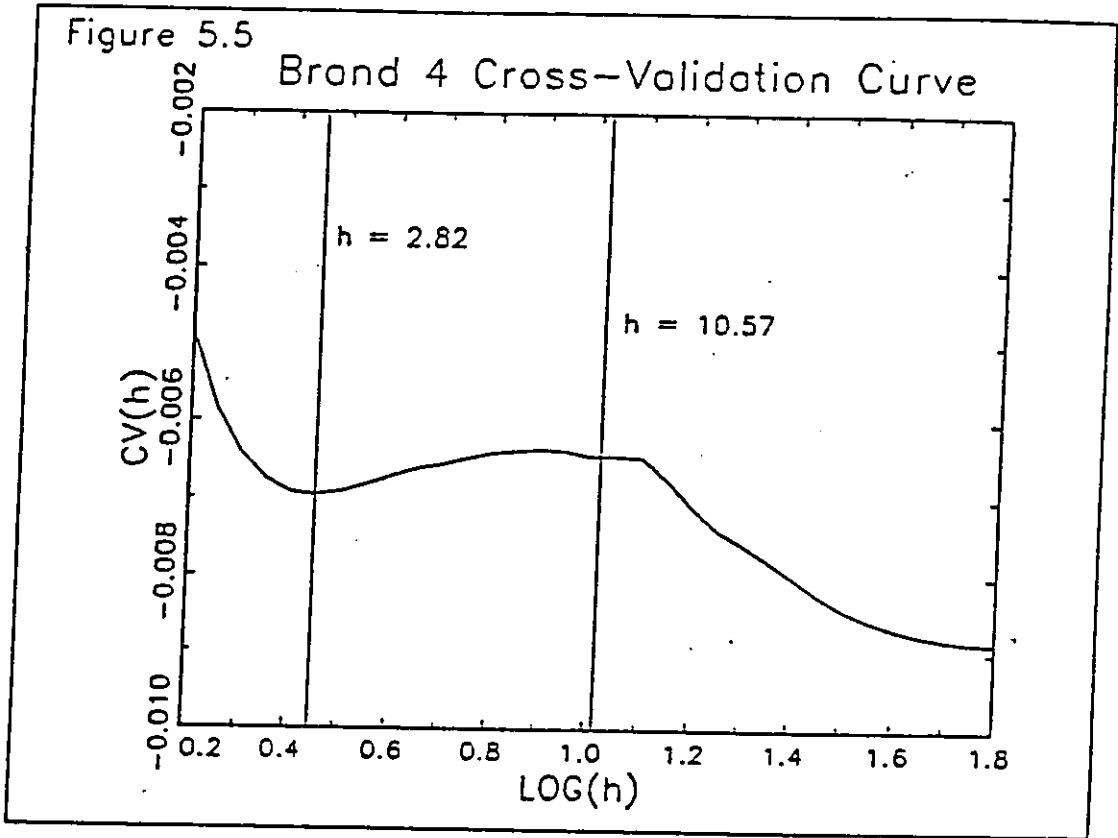
$$CV(h) = \int_0^T [\hat{\lambda}_h^*(t)]^2 dt - 2 \sum_{i=1}^N \hat{\lambda}_{hi}^*(X_i)$$

where  $\hat{\lambda}_{hi}^* = \sum_{\substack{j=1 \\ j \neq i}}^N [K_h(t-X_j) + K_h(t+X_j) + K_h(2T-t-X_j)]$ . Thus, the

cross-validation bandwidth should be close to the bandwidth that minimizes ISE(h). In particular, we saw in Chapters 2 and 4, that the cross-validation bandwidth has some good theoretical properties, for instance asymptotic optimality. Therefore, we used the cross-validation bandwidth for the following purchase rate kernel intensity estimators.

In general, the cross-validation bandwidth selection method performed well for the coffee data. Unfortunately, there are some practical problems to be dealt with. Chief among them is the fact that the cross-validation score function, CV(h), often has several local minima.

For example, consider for Brand 4. The cross-validation score function appears in figure 5.5. There are two local minima at  $h=2.82$  and  $h=10.57$ ; note  $\log(2.82)=.45$  and  $\log(10.57)=1.02$ . The global minimum of the cross-validation score function is larger than 100 weeks; however, since the entire study lasted 108 weeks, this possibility was considered unreasonable. The kernel purchase rate estimator that is created by using the smallest local minimum bandwidth ( $h=2.82$ ) is shown in figure 5.4, and the estimator that is created using the larger local minimum bandwidth ( $h=10.57$ ) is seen in figure



5.3. As mentioned above, the bandwidth  $h=2.82$  seems to perform better because the larger bandwidth,  $h=10.57$ , tends to smooth the data such that two distinct clusters of observations are sometimes combined to form one peak. We believe that the rather small change in  $CV(h)$  at the second local minimum ( $\log(h)=1.02$ ) may indicate that the first local minimum ( $\log(h)=.45$ ) gives the better bandwidth.

It is known that the cross-validation bandwidth is strongly affected by clusters in the data. From simulated examples, data that have large clusters which contain smaller clusters of observations often produce cross-validation score functions with local minima. In addition, Hall and Marron (1991) have shown that for density estimation the expected number of local minima in  $CV(h)$  is proportional to  $\rho = [ \int f^2 ] [ \int (f'')^2 ]^{-1/5}$ . Thus,  $\rho$ , a ratio of two measures of the roughness, indicates the likelihood for having multiple minima. The existence of several minima in  $CV(h)$  is not desirable; on the other hand, the curves that result from the various cross-validation bandwidths often provide complementary and relevant insights into the data. Hence, for a complete data analysis, looking at the curve estimators for each of the local minima is recommended.

Next, look at Brand 29, a brand that was introduced in the middle of the study and had 513 observations. The cross-validation curve in figure 5.6 displays two significant minima at  $h=4.08$  and  $h=31.13$  ( $\log(4.08)=.61$  and  $\log(31.13)=1.49$ ). The kernel intensity estimate with the smaller bandwidth,  $h=4.08$ , is seen in figure 5.7. With this curve, the peaks occur at or near the time of the promotions and thus suggest a strong correlation between the marketing efforts and the

Figure 5.7

Brand 29,  $h=4.08$

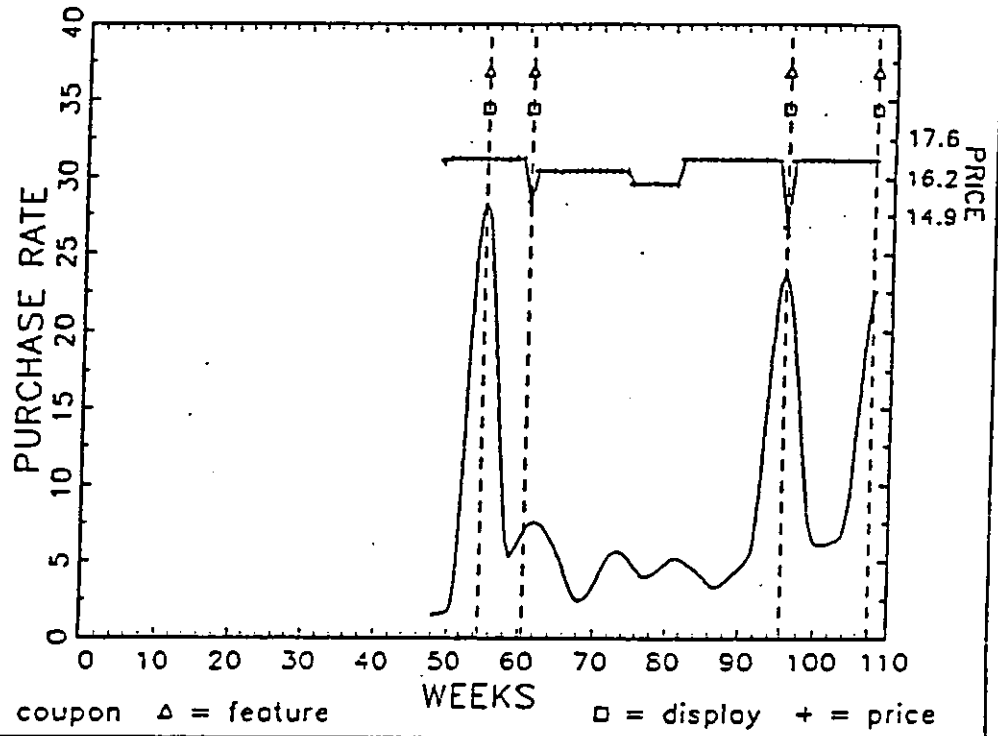
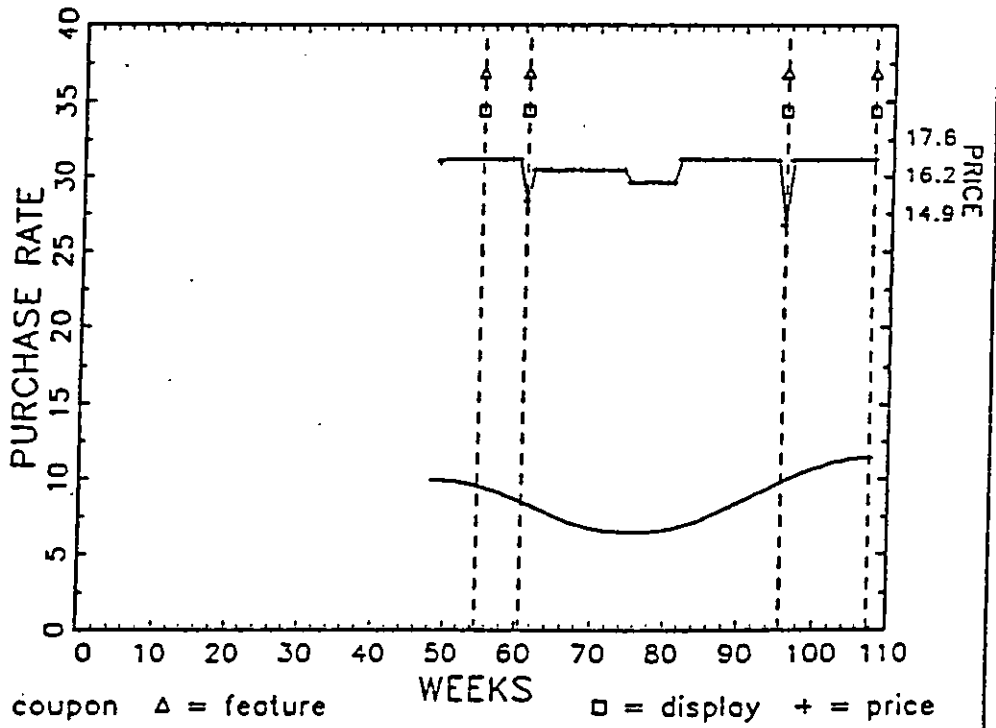


Figure 5.8

Brand 29,  $h=31.13$



purchase rate. In figure 5.8, the intensity estimate with the larger bandwidth,  $h=31.13$ , displays the overall trend of the purchase rate over time. This brand demonstrates how the local minima in the cross-validation curve each provide interesting and valid viewpoints for studying the data. That is, combining the two pictures shows very clearly what is happening to the sales of Brand 29. First of all, sales responded very well to the marketing techniques employed (seen in figure 5.6), and looking macroscopically, the purchase rate decreased after the introduction of the brand and then increased after week 80 (seen in figure 5.7).

Finally, consider Brand 20 a relatively small brand with only 217 observations. In figure 5.9, the cross-validation curve for Brand 20 has only one minimum at  $h=55.56$  ( $\log(55.56)=1.74$ ). The resulting purchase rate curve in figure 5.10 is extremely smooth. This kernel intensity estimate indicates that the purchase rate gently rises in the early weeks and then declines in the later weeks, but remains fairly steady during the entire study. In figure 5.11, we consider a smaller bandwidth ( $h=4.00$ ). When  $h=4$ , peaks are evident at the times that marketing efforts were used; however, by looking at the purchase rate curve alone, it is difficult to distinguish the explained peaks from the unexplained ones. Note that the intensity estimator represents between one and seven sales per week; therefore, these peaks may simply represent Poisson variability in the data. As a consequence, the smooth kernel purchase rate estimate given by the cross-validation bandwidth ( $h=55.56$ ) in figure 5.10 seems more reasonable in this situation.

Figure 5.9

Brand 20 Cross-Validation Curve

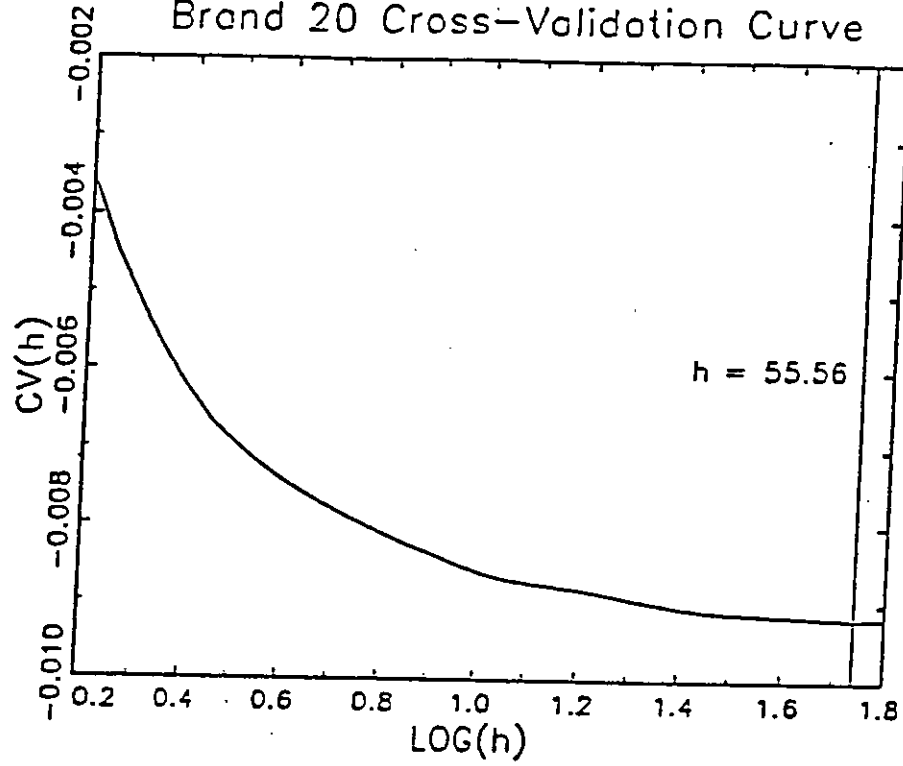


Figure 5.10

Brand 20,  $h=55.56$

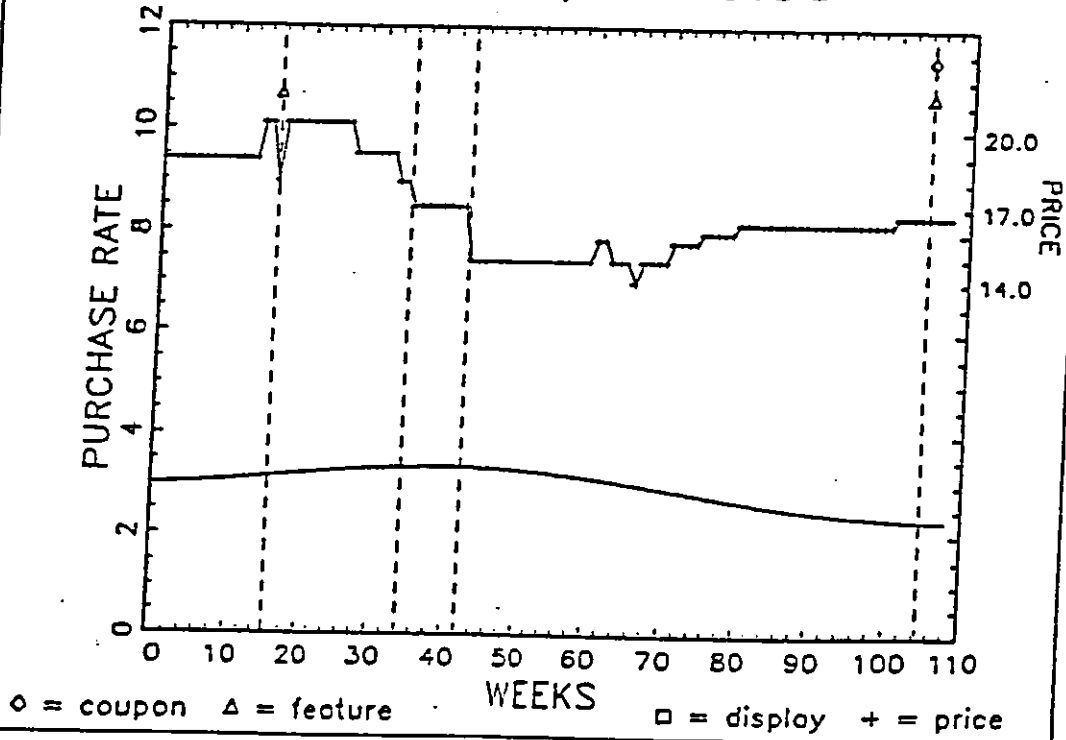
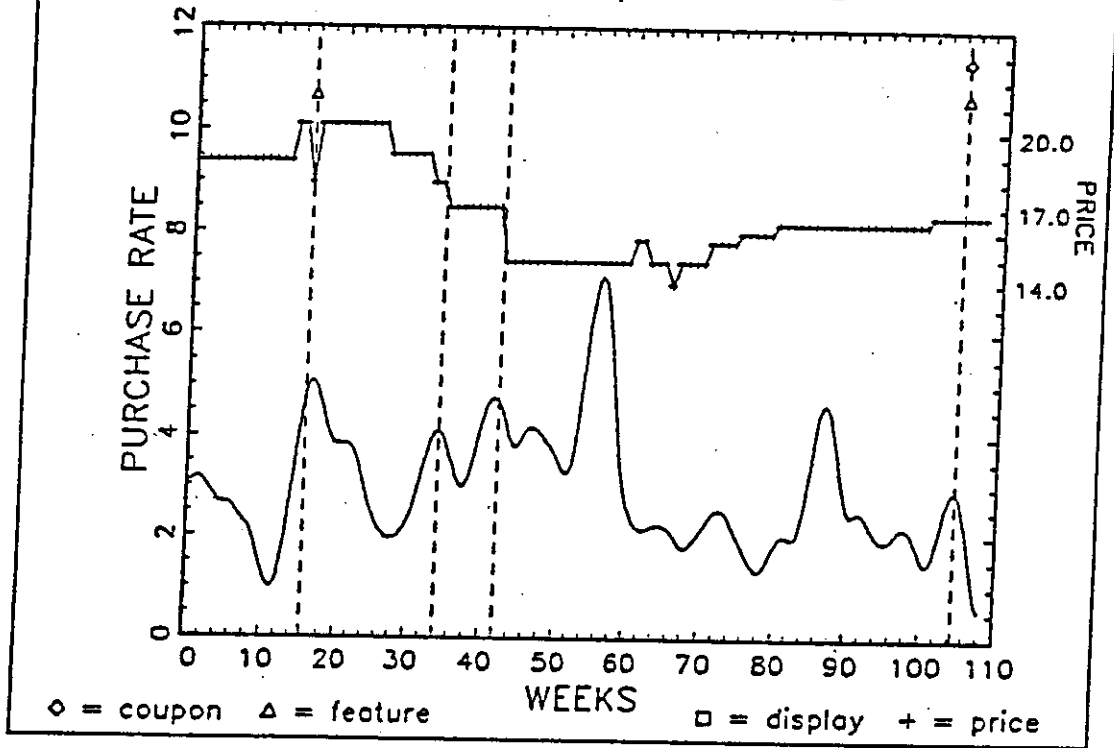


Figure 5.11

Brand 20,  $h=4.00$



The fairly constant sales rate for Brand 20 can be justified by the marketing theory that small brands may have a loyal group of customers who continue to buy their coffee with or without promotions. Combining the marketing theory and the purchase rate kernel estimator with the cross-validation bandwidth leads us to believe that promotions are not very effective for increasing the Brand 20 sales rate. Recall that for the large brand, Brand 4, there appeared to be a strong relationship between promotions and the purchase rate. Thus, one can hypothesize that large brands derive greater benefits from marketing efforts compared to small brands.

The purchase data for the three coffee brands discussed above are examples where the kernel estimator with the cross-validation bandwidth seems to perform quite well. In particular, this bandwidth is quite effective in distinguishing between increases in sales that are due to promotions rather than random variations in the data. Therefore, the kernel estimator with the least-squares cross-validation bandwidth assists us in interpreting the relationship between promotions and the purchase rate and thus provides greater insight into the coffee data. This example demonstrates that the cross-validation bandwidth can be a practical bandwidth for analyzing data from a counting process.

## REFERENCES

- Aalen, Odd (1978), "Nonparametric Inference for a Family of Counting Processes," *The Annals of Statistics*, 6, 701-726.
- Andersen, Per Kragh and Borgan, Ornulf (1985), "Counting Process Models for Life History Data: A Review," *Scandinavian Journal of Statistics*, 12, 97-158.
- Bowman, Adrian W. (1984), "An Alternative of Cross-Validation for the Smoothing of Density Estimates," *Biometrika*, 71, 363-60.
- Burkholder, D. L. (1973), "Distribution Function Inequalities for Martingales," *The Annals of Probability*, 1, 19-42.
- Bürman, Prabir (1985), "A Data Dependent Approach to Density Estimation," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 69, 609-628.
- Castellana, J. V. and Leadbetter, M. R. (1986), "On Smoothed Probability Density Estimation for Stationary Processes," *Stochastic Processes and their Applications*, 21, 179-193.
- Clevenson, M. Lawrence and Zidek, James V. (1977), "Bayes Linear Estimators of the Intensity Function of the Nonstationary Poisson Process," *Journal of the American Statistical Association*, 72, 112-120.
- Cline, D. and Hart, J. (1986), "Kernel Estimation of Densities with Discontinuities or Discontinuous Derivatives," unpublished manuscript.
- Cox, D. R. and Isham, Valerie (1980), *Point Processes*, Chapman and Hall, London.
- Devroye, Luc and Györfi, László (1984), *Nonparametric Density Estimation: The  $L_1$  View*, John Wiley and Sons, New York.
- Diggle, Peter J. (1983), *Statistical Analysis of Spatial Point Patterns*, Academic Press, London.
- Diggle, Peter (1985), "A Kernel Method for Smoothing Point Process Data," *Applied Statistics*, 34, 138-147.
- Diggle, Peter and Marron, J. S. (1988), "Equivalence of Smoothing Parameter Selectors in Density and Intensity Estimation," *Journal of the American Statistical Association*, 83, 793-800.
- Duin, Robert P. W. (1976), "On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions," *IEEE Transactions on Computers*, C-25, 1175-1179.

- Ellis, Steven (1986). "A Limit Theorem for Spatial Point Processes," *Advances in Applied Probability*, 18, 646-65.
- Epanechnikov, V. A. (1969). "Non-parametric Estimation of a Multivariate Probability Density," *Theory of Probability and Its Applications*, 14, 153-158.
- Farrell, R. H. (1972). "On the Best Obtainable Asymptotic Rates of Convergence in Estimation of a Density Function at a Point", *The Annals of Mathematical Statistics*, 43, 170-180.
- Fiskel, Thomas (1988). "Edge-Corrected Density Estimators for Point Processes", *Statistics*, 1, 67-75.
- Foldes, A. and Revesz, P. (1974). "A General Method for Density Estimation," *Studia Scientiarum Mathematicarum Hungarica*, 9, 81-90.
- Gasser, T., Mueller, H. G. and Mammitzsch, V. (1985). "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society, Series B*, 47, 238-252.
- Habbema, J. D. F., Hermans, J. and van den Broek, K. (1974). "A Stepwise Discriminant Analysis Program Using Density Estimation," *Compstat 1974: Proceedings in Computational Statistics*, Physica Verlag, Vienna, 101-110.
- Hall, Peter (1983). "Large Sample Optimality of Least Squares Cross-Validation in Density Estimation," *The Annals of Statistics*, 11, 1156-1174.
- Hall, Peter (1984). "Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators," *Journal of Multivariate Analysis*, 14, 1-16.
- Hall, Peter (1987a). "On Kullback-Leibler Loss and Density Estimation," *The Annals of Statistics*, 15, 1491-1519.
- Hall, Peter (1987b). "On the Use of Compactly Supported Density Estimates in Problems of Discrimination," *Journal of Multivariate Analysis*, 23, 131-158.
- Hall, Peter and Marron, J. S. (1987a). "Extent to which Least-Squares Cross-Validation Minimises Integrated Squares Error in Nonparametric Density Estimation," *Probability Theory and Related Fields*, 74, 567-581.
- Hall, Peter and Marron, J. S. (1987b). "Choice of Kernel Order in Density Estimation," *The Annals of Statistics*, 16, 161-173.
- Hall, Peter and Marron, J. S. (1991). "Local Minima in Cross-validation Functions," *Journal of the Royal Statistical Society, Series B*, 53, 245-252.

- Hardle, Wolfgang, Hall, Peter and Marron, J. S. (1988), "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?" *Journal of the American Statistical Association*, 83, 86-101.
- Hardle, Wolfgang, and Marron, J. S. (1985), "Optimal Bandwidth Selection in Nonparametric Regression Estimation," *The Annals of Statistics*, 13, 1465-1481.
- Hardle, Wolfgang, Marron, J. S. and Wand, M. P. (1990), "Bandwidth Choice for Density Derivatives," *Journal of the Royal Statistical Society, Series B*, 52, 223-232.
- Leadbetter, M. R. and Wold, Diane (1983), "On Estimation of Point Process Intensities." *Contributions to Statistics: Essays in Honour of Norman L. Johnson* (P.K. Sen, ed.), North-Holland, Amsterdam, 299-312.
- Marron, James Stephen (1985) "An Asymptotically Efficient Solution to the Bandwidth Problem of Kernel Density Estimation," *The Annals of Statistics*, 13, 1011-1023.
- Marron, J. S. (1987) "A Comparison of Cross-Validation Techniques in Density Estimation," *The Annals of Statistics*, 15, 152-162.
- Marron, James Stephen and Hardle, Wolfgang (1986), "Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation," *Journal of Multivariate Analysis*, 20, 91-113.
- Parzen, Emanuel (1962), "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, 33, 1065-1076.
- Ramlau-Hansen, Henrik (1983a), "Smoothing Counting Process Intensities by Means of Kernel Functions," *The Annals of Statistics*, 11, 453-466.
- Ramlau-Hansen, Henrik (1983b), "The Choice of a Kernel Function in the Graduation of Counting Process Intensities," *Scandinavian Actuarial Journal*, 165-182.
- Rice, John (1984), "Boundary Modification for Kernel Regression," *Communications in Statistics, Series A*, 13, 893-900.
- Ripley, B. D. (1981), *Spatial Statistics*, John Wiley and Sons, New York.
- Rosenblatt, Murray (1956), "Remarks on Some Nonparametric estimates of a Density Function," *The Annals of Mathematical Statistics*, 27, 832-837.

- Rosenblatt, Murray (1971), "Curve Estimates," *The Annals of Mathematical Statistics*, 42, 1815-1842.
- Rudemo, Mats (1982), "Empirical Choice of Histograms and Kernel Density Estimators," *Scandinavian Journal of Statistics*, 9, 65-78.
- Schucany, W. R. and Sommers, J. P. (1977), "Improvement of Kernel Type Density Estimators," *Journal of the American Statistical Association*, 72, 420-423.
- Schuster, Eugene F. (1985), "Incorporating Support Constraints into Nonparametric Estimators of Densities," *Communications in Statistics, Series A*, 14, 1123-1136.
- Scott, David W., Tapia, Richard A. and Thompson, James R. (1977), "Kernel Density Estimation Revisited," *Nonlinear Analysis, Theory, Methods and Applications*, 1, 339-372.
- Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.
- Stone, Charles (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348-1360.
- Stone, Charles (1984), "An Asymptotically Optimal Window Selection Rule For Kernel Density Estimates," *The Annals of Statistics*, 12, 1285-97.
- Tapia, Richard A. and Thompson, James R. (1978), *Nonparametric Probability Density Estimation*, The Johns Hopkins University Press, Baltimore.
- Walter, G. and Blum, J. (1979), "Probability Density Estimation Using Delta Sequences," *The Annals of Statistics*, 7, 328-340.
- Watson, G. S. and Leadbetter, M. R. (1963), "On the Estimation of the Probability Density. I," *The Annals of Mathematical Statistics*, 34, 480-491.
- Whittle, P. (1958), "On the Smoothing of Probability Density Functions," *Journal of the Royal Statistical Society, Series B*, 20, 334-343.