Equivalence of Smoothing Parameter Selectors in
Density and Intensity Estimation

PETER DIGGLE and J. S. MARRON*

Kernel smoothing is an attractive method for the nonparametric estimation of either a probability density function or the intensity function of a nonstationary Poisson process. In each case the amount of smoothing, controlled by the bandwidth, that is, smoothing parameter, is crucial to the performance of the estimator. Bandwidth selection by cross-validation has been widely studied in the context of density estimation. A bandwidth selector in the intensity estimation case has been proposed that minimizes an estimate of the mean squared error under the assumption that the data are generated by a stationary Cox process. This article shows that these two methods each select the same bandwidth, even though they are motivated in much different ways. In addition to providing further justification of each method, this equivalence of smoothing parameter selectors yields new insights for both density and intensity estimation. A benefit for intensity estimation is that this equivalence makes it clear how the Cox process method may be applied to kernels that are nonuniform, or even of higher order. Another benefit is that this duality between problems makes it clear how to apply the well-developed asymptotic methods for understanding density estimation in the intensity setting. A benefit for density estimation is that it motivates an analog of the Cox process method, which provides a useful nonasymptotic means of studying that problem. The specific forms of the estimators and smoothing parameter selectors are introduced in Section 1. The basic equivalence result is stated in Section 2. Sections 3 and 4 describe new insights that follow for intensity and density estimation, respectively. Section 5 discusses modification of these ideas to take boundary effects into consideration and shows how they can be used to motivate new boundary adjustments in intensity estimation.

KEY WORDS: Cox process; Cross-validation; Kernel estimators; Nonparametric density estimation; Poisson intensity estimation; Smoothing.

1. THE ESTIMATORS

The raw data for both density and intensity estimation consist of a set of points \( X_1, \ldots, X_n \in \mathbb{R} \). For density estimation, these are thought of as realizations of \( n \) independent random variables, all having probability density function \( f(x) \). For intensity estimation these are thought of as a realization on an interval \([0, T]\) of a nonstationary Poisson process with intensity function \( \lambda(x) \). This function has also been called the rate function or mean density function. See any standard textbook, such as Parzen (1962a), for a definition.

A reasonable estimate of either \( f \) or \( \lambda \) should be a function that takes on large values in regions where the data are dense and values close to 0 where the data are sparse. The kernel smoothing method of constructing such a function is to let

\[
\hat{f}(x) = n^{-1} \sum_{i=1}^{n} \delta_i(x-X_i)
\]

for density estimation or

\[
\hat{\lambda}(x) = \sum_{i=1}^{n} \delta_i(x-X_i)
\]

for intensity estimation, where \( \delta_i(\cdot) = (1/t)\delta(\cdot/t) \) for \( t > 0 \) and \( \delta(\cdot) \) is a symmetric probability density. The parameter \( t \) controls the amount of smoothing that is done and is called the bandwidth. The normalization factor of \( n^{-1} \) makes \( \hat{f}(x) \) a probability density. For access to the literature on theoretical properties of these estimators see Devroye and Györfi (1984), Leadbetter and Wold (1983), and Ellis (1986).

In Figure 1 we show how varying the bandwidth \( t \) affects the smoothness of the intensity estimator \( \hat{\lambda}(x) \). The data are the times of 191 coal-mining disasters in a total time period of 40,550 days, as compiled by Jarrett (1979). Rudosko (1982) used an earlier, incomplete version of these data to illustrate the use of least squares cross-validation for bandwidth selection. We use a quartic kernel,

\[
\delta(x) = 0.9375(1 - x^2)^2, \quad -1 \leq x \leq 1
\]

\[
= 0, \quad \text{otherwise.}
\]

Figure 1 shows three estimates \( \hat{\lambda}(x) \) with a fourfold increase in \( t \) between each pair of estimates and the intermediate value \( t = 5.9573 \), the optimum value according to the method of bandwidth selection described in Diggle (1985).

Note that the estimate with \( t = 1.4893 \) oscillates wildly and contains many features that one cannot expect to distinguish with only 191 observations. On the other hand, the curve with \( t = 23.8292 \) has lost features that are more likely to be reflected by the underlying intensity function. This is the essence of the smoothing problem: to smooth enough to keep sample noise at acceptable levels, while not smoothing so much that interesting features disappear.

Generally the choice of \( t \) is far more important for the effective performance of a kernel estimator than is the choice of \( \delta(\cdot) \). For further discussion and theoretical analysis of this issue see, for example, Silverman (1986).

* Peter Diggle is Chief, Division of Mathematics and Statistics, Commonwealth Scientific and Industrial Research Organisation, Canberra, Australian Capital Territory 2600, Australia. J. S. Marron is Assistant Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27514. This research, supported by National Science Foundation Grants DMS-8400602 and DMS-87071201, was carried out while Marron was visiting the Australian National University.

© 1988 American Statistical Association
Journal of the American Statistical Association
September 1988, Vol. 83, No. 403, Theory and Methods
793
process or doubly stochastic Poisson process. See, for example, Cox and Isham (1980, pp. 70–75). Letting $E$ denote expectation over both the randomness in $X_1, \ldots, X_n$ and in $\Lambda(x)$, Diggle showed that, for $x$ more than $t$ units from the boundary of $[0, T]$, 
\[
\text{MSE}(t) = E[(\hat{\lambda}_n(x) - \Lambda(x))^2] = v(0) + \mu[1 - 2\mu K(t)]/2t + (\mu/2t)^2 \int_0^T K(y) dy,
\] (1.3)
where $K(t) = 2\mu^{-2} \int_0^t v(x) dx$ for the uniform kernel $\delta(\cdot) = \frac{1}{b} I_{[-1,1]}(\cdot)$ (1.4) has been used in the estimator $\hat{\lambda}_n$. The function $K(t)$ is very useful for the analysis of spatial point processes (see Diggle 1983; Ripley 1981). In the one-dimensional case it can be estimated by 
\[
\hat{K}(t) = Tn^{-2} \sum_{i \neq j} I_{[-1,1]}(X_i - X_j).
\]
Hence the bandwidth that minimizes the mean squared error (MSE) (1.3) should be fairly well approximated by the bandwidth, $\hat{\theta}_M$, which minimizes 
\[
\hat{M}(t) = (2\mu t)^{-1} - t^{-1} \hat{K}(t) + (2t)^{-2} \int_0^T \hat{K}(y) dy,
\]
where $\mu$ has been estimated by 
\[
\hat{\mu} = h/T.
\] (1.5)

Ideas related to this were also developed by Clevenson and Zidek (1977). Note that no attempt has been made to adjust for boundary effects. These adjustments are very important, but they have a tendency to obscure the main points being made. Hence we ignore boundary effects and adjustments until Section 5, where a detailed treatment is given.

## 2. THE EQUIVALENCE

Note that the bandwidths $\hat{\theta}_C$ and $\hat{\theta}_M$ are both well defined in either setting. The surprising fact that they are exactly the same under no additional assumptions is contained in the following theorem.

**Theorem 2.1.** In the case of the uniform kernel (1.4), $\hat{\theta}_C = \hat{\theta}_M$, in the sense that each minimizer of $CV(t)$ is a minimizer of $\hat{M}(t)$ and vice versa.

**Proof.** Theorem 2.1 follows immediately from the following lemma.

**Lemma 2.1.** For the uniform kernel, $\hat{M}(t) = T \cdot CV(t)$. (The proof is in the Appendix.)

The first consequence of Theorem 2.1 is that it shows that each method is more natural than previously suspected, because it can be derived from much different ideas than those originally used. The benefits go far deeper than this, however, because ideas in one setting can now
be used to obtain more insight into the other setting. In Section 3 we show how the well-developed asymptotic theory of density estimation can point the way to new ideas in intensity estimation by discussing different kernel functions, including those of higher order. In Section 4, an empirical Bayes approach is developed for density estimation.

3. BENEFITS FOR INTENSITY ESTIMATION

Note that the results in Section 2 are only stated in terms of the uniform kernel (1.4). This was because in Diggle (1985) the motivation for \( \hat{I}_n \) appeared to apply only to this case. The motivation for \( \hat{I}_{CV} \), however, works for any kernel, so this suggests using Lemma 2.1 to find an appropriate \( \hat{M}(t) \) for the general case. In particular, define \( \hat{M}(t) = T \cdot CV(t) \). A form of \( \hat{M}(t) \) that will allow an MSE interpretation is given by the following lemma.

**Lemma 3.1.** For a general \( \delta_i \),

\[
\hat{M}(t) = T \cdot CV(t) = (\hat{\mu} t)^{-1} \left[ \int \delta^2 \right] + t^{-1} \hat{K}_{\delta}(t) - 2t^{-1} \hat{K}_d(t),
\]

where

\[
\hat{K}_d(t) = n^{-2}T \sum_{i=1}^{n} \delta(X_i - X_t),
\]

\[
\delta \ast \delta(\cdot) = \int \delta(\cdot - x) \delta(x) \ dx,
\]

and \( \hat{\mu} \) is defined at (1.5). (The proof is in the Appendix.)

Next it should be verified that choosing \( \hat{I}_n \) to be the minimizer of this general \( \hat{M}(t) \) still makes sense in the context of Diggle (1985). Much of the work for this is summarized in the following lemma, the proof of which is also in the Appendix. Recall that boundary effects are ignored until Section 5.

**Lemma 3.2.**

\[
\text{MSE}(t) = \mu^{-1} \left[ \int \delta^2 \right] + \mu^2 t^{-1} \hat{K}_{\delta}(t) - \mu^2 t^{-1} \hat{K}_d(t) + v(0),
\]

where \( \hat{K}_d(t) = 2t^{-2} \int_0^t \delta(u) v(u) \ du \). Since \( \hat{K}_d(t) \) is a reasonable estimate of \( K_d(t) \), the minimizer of \( \mu^2 \hat{M}(t) + v(0) \) [which also minimizes \( \hat{M}(t) \)] should be close to the minimizer of MSE(t). Note that when \( \delta(\cdot) \) is the uniform kernel, \( K_d(t) \) differs from \( K(t) \) by a factor of 2.

A very useful tool in the study of density estimation is asymptotics with \( n \to \infty \). The aforementioned strong connection between bandwidth selectors in the two settings motivates looking for an analog in the intensity-estimation case. The first type of asymptotics that one might consider is \( T \to \infty \). As pointed out at the end of section 2.2 of Diggle (1985), however, this will only add new information at the right endpoint, instead of everywhere as for \( n \to \infty \) in density estimation. A way to add information everywhere is to allow \( \mu \to \infty \) (where \( \mu \) was the mean of the underlying stochastic process introduced at the end of Sec. 1). Note that care must be taken to avoid changing the relative shape of the curve \( \Lambda(x) \) in the limiting process. This is accomplished by taking the expected product function \( v(x) \) (defined in Sec. 1) to be of the form

\[
v(x) = \mu^{-1} v_0(x)
\]

for some fixed function \( v_0(x) \). For insight into the effects of \( \mu \to \infty \) with \( v \) of the form (3.1), consider the special case of the “linear Cox process,” described in section 2.2 of Diggle (1985). For this process \{\( \Lambda(x) \)} can be represented as

\[
\Lambda(x) = \sum_{i=1}^{\infty} h(x - Z_i),
\]

where the \( Z_i \) are the points of a homogeneous Poisson process and \( h(\cdot) \) is a nonnegative-valued function.

A very important application of \( n \to \infty \) asymptotics in density estimation is that they allow a very simple and elegant quantification of the smoothing problem, that is, the fact that too small admits too much sample noise and too large smooths away features of the underlying curve. In particular (see, e.g., Rosenblatt 1971), various squared error criteria can be expanded into simple and easily interpreted variance and squared bias components that become large when \( t \) is small and large, respectively.

To apply these ideas to intensity estimation, we work with \( \mu^{-2} \text{MSE}(t) \), where the normalization factor \( \mu^{-2} \) may be thought of as adjusting for the difference between the estimators \( \hat{f}_n \) and \( \hat{f}_n \). From Lemma 3.2,

\[
\mu^{-2} \text{MSE}(t) = \mu^{-1} t^{-1} \left[ \int \delta^2 \right] + t^{-1} \hat{K}_{\delta}(t) - 2t^{-1} \hat{K}_d(t) + v(0).
\]

An inspection of the proof of Lemma 3.2 shows that the first term can be thought of as a type of variance, and the remaining terms are the corresponding squared bias. Hence we define

\[
v(t) = \mu^{-1} t^{-1} \left[ \int \delta^2 \right]
\]

and

\[
b^2(t) = t^{-1} \hat{K}_{\delta}(t) - 2t^{-1} \hat{K}_d(t) + v(0).
\]

To gain more insight into the behavior of the squared bias, consider the expansion summarized in the following lemma, the proof of which is in the Appendix.

**Lemma 3.3.** If \( v_0([x]) \) has a continuous fourth derivative at the origin, then as \( t \to 0 \),

\[
b^2(t) = t^4 v_0''(0) \left[ \int u^2 \delta(u) \ du/2 \right]^2 + o(t^4).
\]

It is interesting to compare this with the expression for the squared bias in density estimation; see, for example, (1.6) of Rosenblatt (1971). The only difference is that \( v_0''(0) \) replaces \( f_0''(x) \). This is not surprising in view of
the well-known relationship between the kth derivative of a process \( \Lambda(x) \) and the 2kth derivative at the origin of its covariance function, \( v(x) - \mu^2 \).

In density estimation, the variance term also admits a simple asymptotic expansion; see (9) of Rosenblatt (1971). Note that the dominant term of this expansion is in fact the same as \( v(t) \), if we identify \( n \) with \( \mu \). The fact that no expansion is required in the present context seems related to the fact that the variance of the Poisson distribution has a simpler form than the binomial.

Parzen (1962b) demonstrated that if a density has more than two derivatives, say \( k \), then a faster rate of convergence of \( \hat{f}_k \) to \( f \) can be obtained by using a “higher-order kernel,” that is, assuming that

\[
\int x^l \delta(x) \, dx = 1 \quad \text{if } l = 0
\]

\[
= 0 \quad \text{if } l = 1, 2, \ldots, k - 1
\]

\[
= C > 0 \quad \text{if } l = k.
\]

(3.3)

It is straightforward to extend the foregoing computations to this case. The answer is summarized as follows.

**Lemma 3.4.** If \( \delta \) satisfies (3.3) and \( v_0(|x|) \) is a continuous 2kth derivative at the origin, then as \( t \to 0 \),

\[
b^2(t) = t^{2k} v_0^{(2k)}(0) \left( \int u^k \delta(u) \, du / k! \right)^2 + o(t^{2k}).
\]

The proof of Lemma 3.4 is omitted because it is essentially the same as the proof of Lemma 3.3.

The foregoing results are a very small part of what can be done in terms of finding intensity estimation analogs of what is already known about density estimation. Other possibilities, which would require much more work than can be done in this article, include analogs of the optimal rates ideas of Farrell (1972) and Stone (1980), the asymptotic optimality results of Stone (1984) and Marron (1987), the noise in bandwidth selection ideas of Hall and Marron (1987a, b), the location-dependent smoothing ideas of Abramson (1982), and the kernel selection ideas of Epanechnikov (1969) and Hall and Marron (1988).

### 4. Benefits for Density Estimation

The equivalence described in Section 2 motivates consideration of the density estimation problem from the empirical Bayes viewpoint considered in Diggle (1985). A related approach was taken by Whittle (1958). An analog of the Cox process may be defined where \( X_1, \ldots, X_n \) is an iid sample from a probability density \( \Lambda(x) \) on \([0, T]\), which is a realization of a stationary stochastic process \( \Lambda \), with \( E[\Lambda(x)] = T^{-1} \) and \( E[\Lambda(x)\Lambda(y)] = v(|x - y|) \). An example of such a process \( \Lambda \) is a normalized version of the “linear Cox process” described in section 2.2 of Diggle (1985) and Equation (3.2). (Note that in this context a kernel density estimate seems especially appropriate, since the underlying density itself has the same form.)

An expression for the mean squared error, \( \text{MSE}^*(t) \), of the estimator \( \hat{f}_n(x) \) is contained in the following lemma.

**Lemma 4.1.** In the Cox-process density-estimation setting

\[
\text{MSE}^*(t) = (nt)^{-1} \left( \int \delta^2 \right) + (1 - n^{-1}) t^{-1} T^{-2} K^*_s(t) - 2t^{-1} T^{-2} K^*_s(t) + v(0).
\]

where \( K^*_s(t) = 2t^{-2} \int_0^t \delta_1(u) v(u) \, du \).

The proof of Lemma 4.1 is in the Appendix. Note that when \( \mu \) is identified with \( T^{-1} \), this is very similar to Lemma 3.2, the only difference being that a factor of \( n \) now appears in the first term and a negligible factor of \( 1 - n^{-1} \) is now in the second term. These differences reflect the fact that the Poisson setting has been exchanged for one closer to the binomial.

Now, \( K^*_s(t) \) is, by similar considerations, reasonably estimated by

\[
\hat{K}^*_s(t) = n^{-1}(n - 1)^{-1} t T \sum_{i \neq j} \delta_i(X_i - X_j).
\]

Hence the minimizer of

\[
\hat{M}^*(t) = (nt)^{-1} T \left( \int \delta^2 \right) + (1 - n^{-1}) t^{-1} \hat{K}^*_s(t) - 2t^{-1} \hat{K}^*_s(t),
\]

say \( \hat{M}^*_s \), should be close to the minimizer of \( \text{MSE}^*(t) \). The ideas of this article are brought full circle by showing that \( \hat{M}^*_s \) has a representation in terms of cross-validation.

**Lemma 4.2.** \( \hat{M}^*(t) = T \cdot \text{CV}^*(t) \), where

\[
\text{CV}^*(t) = \int [\hat{f}_n(x)]^2 \, dx - 2n^{-1} \sum_{j=1}^n \hat{f}^*_n(X_j)
\]

and

\[
\hat{f}^*_n(x) = (n - 1)^{-1} \sum_{i=1}^{n} \delta_i(x - X_i).
\]

The proof of Lemma 4.2 is omitted because it is very similar to the proof of Lemma 2.2.

The difference between the present estimators, \( \hat{K}_s^* \) and \( \hat{f}^*_n \), and their original analogs, \( K_s \) and \( f_n \), is a negligible factor of \( 1 - n^{-1} \). The reason for introducing new notation instead of simply remarking that they are approximately the same is that the cross-validation score \( \text{CV}^*(t) \) appears more often in the literature than \( \text{CV}(t) \) (see Hall and Marron 1987b; Marron 1987; Rudemo 1982). The calculations done in this section make \( \text{CV}^* \) seem slightly more natural; however, there is essentially no difference in practice.

### 5. Boundary Adjustments

Boundary effects can be a problem in intensity estimation when \( \lambda(0) \) and \( \lambda(T) \) are positive. If the definition of \( \lambda(x) \) is extended outside of \([0, T]\) by taking it to have the value 0, then \( \lambda(x) \) is discontinuous at 0 and \( T \). In this case, the continuous estimate \( \hat{\lambda}(x) \) will perform poorly on
neighborhoods of 0 and $T$. By studying asymptotics of the type discussed in Section 3, this difficulty can be quantified. In particular, it can be shown that $\hat{\lambda}_i(0)$ and $\hat{\lambda}_i(T)$ are inconsistent.

The same problem exists in density estimation when there are discontinuities in the density, such as happens if the density is bounded above 0 on an interval and equal to 0 off it. See Rice (1984), Schuster (1985), Gasser, Müller, and Mammitzsch (1985), and Cline and Hart (1986) for further discussion on this topic.

A simple method of correcting this problem is the “mirror image” adjustment considered by Schuster (1985) and Cline and Hart (1986). This correction takes those kernel functions that extend beyond the boundary and “folds” them at the boundary, so all of their mass is inside the interval. We illustrate the method by showing how $\hat{\lambda}_i$ can be adjusted at 0. Adjustments at $T$ and for density estimation are similar. The basic idea is that the piece of each $\delta_i(\cdot - X_i)$ that extends to the left of 0 should be “folded” at 0 so that all of its mass is inside $[0, T]$. This is done by defining

$$
\hat{\lambda}_i^*(x) = \sum_{i=1}^{n} [\delta_i(x - X_i) + \delta_i(x + X_i)]I_{[0,T]}(x). \quad (5.1)
$$

Another possible boundary adjustment is given in (1.1) of Diggle (1985). For this type of correction, those kernels that extend beyond the boundary are truncated and then rescaled to have the correct mass. A simple asymptotic expansion of the type in Lemma 3.3 shows that Diggle’s estimator will typically have slightly more bias than (5.1).

To see the practical implications of this, consider Figure 2, which shows the two edge-corrected estimates, which are in close agreement, together with the uncorrected version. Clearly, some form of boundary adjustment is vital.

Note that (5.1) is exactly the type of boundary adjustment that is being done by the estimator $\hat{K}(t)$ in Section 3.2 of Diggle (1985). Hence one sensible method of adjusting for boundary effects is to use the estimator $\hat{\lambda}_i^*$, with the bandwidth $t$ chosen to minimize $\tilde{M}(t)$ as defined in Diggle (1985). Similar remarks hold for density estimation.

A drawback of the foregoing recipe is that for the bandwidth selection part, the boundary will still have a tendency to introduce a bias toward undersmoothing (see Cline and Hart 1986). One means of approaching this problem is to use the more sophisticated boundary adjustments proposed by Rice (1984) and Gasser et al. (1985).

Another means of approaching the bias toward undersmoothing can be motivated from density estimation (see Marron 1987). Suppose that $\delta(x)$ is supported on $[-1,1]$. Find $t_0$ so that reasonable values of $t$ are smaller than $t_0$. Now, as in Section 1, consider estimating the first two terms of

$$
\int_{t_0}^{T-t_0} \tilde{f}; \int_{t_0}^{T-t_0} \tilde{f}_t f + \int_{t_0}^{T-t_0} f^2 = \int_{t_0}^{T-t_0} \tilde{f}; - f^2
$$

by

$$
CV^+(t) = \int_{t_0}^{T-t_0} \tilde{f}; - 2n^{-1} \sum_{j=1}^{n} \hat{f}_t(X_j)I_{[t_0,T-t_0]}(X_j).
$$

It is important to note that not all of $X_1, \ldots, X_n$ are used to construct the estimates $\tilde{f}_t$ and $\tilde{f}_t$, but only those inside $[t_0, T-t_0]$ appear in the sum inside CV+2. This avoids moving the problem of the boundaries at 0 and $T$ to $t_0$ and $T-t_0$.

To find the analog of this in the intensity estimation setting, Lemma 2.1 motivates looking for another interpretation of

$$
\tilde{M}(t) = T \cdot CV^+(t).
$$

Much of the work in this is summarized in the following lemma.

**Lemma 5.1.** If $\delta$ is supported on $[-1,1]$ and $t < t_0$, then, in the intensity estimation setting,

$$
E \left[ \sum_i \int_{t_0}^{T-t_0} \delta_i(x - X_i) \ dx \right] = (T - 2t_0)\mu t^{-1} \left[ \int \delta^2 \right],
$$

$$
E \left[ \sum_{i,j} \int_{t_0}^{T-t_0} \delta_i(x - X_i) \delta_j(x - X_j) \ dx \right] = (T - 2t_0)\mu t^{-1}K_{2\delta_2}(t),
$$

and

$$
E \left[ \sum_{i,j} \delta_i(X_i - X_j)I_{[t_0,T-t_0]}(X_j) \right] = (T - 2t_0)\mu t^{-1}K_{\delta}(t).
$$

The proof of Lemma 5.1 is in the Appendix.
Hence
\[
\int \left[ f_i(x) \right]^2 \, dx \\
= n^{-2} \left[ \sum_{i=1}^{n} \delta(X_i - X) \right] \, dx \\
+ \sum_{i \neq j} \left[ \delta(X_i - X) \delta(X_i - X_j) \right] \, dx \\
= \frac{1}{4} \, n^{-2} \sum_{i = 1}^{n} \left[ I_{[X_{i-1}, X_i)}(x) \right] \, dx \\
+ \sum_{i \neq j} \left[ I_{[X_{i-1}, X_i)}(X_i - X_j, 2n) \right] \, dx \\
= \frac{1}{4} \, n^{-2} \left[ 2nt + \sum_{i \neq j} (2t - |X_i - X_j|) \right. \\
\times \left. I_{[-2, 2]}(X_i - X_j) \right] \\
= (2nt)^{-1} + T^{-1}(2t)^{-1} \int_{0}^{2t} \hat{K}(y) \, dy. \tag{A.1}
\]
Thus
\[
\text{CV} = \int \left[ \hat{f}_i(x) \right]^2 \, dx - 2n^{-2} \sum_{i \neq j} (2t)^{-1} I_{[-1, 1]}(X_i - X_j) \\
= (2nt)^{-1} + T^{-1}(2t)^{-1} \int_{0}^{2t} \hat{K}(y) \, dy - (tT)^{-1} \hat{M}(t) \\
= T^{-1} \hat{M}(t).
\]

Proof of Lemma 3.1. First note that (A.1) becomes
\[
\left( \int \left[ \hat{f}_i(x) \right]^2 \, dx \right)^2 = (nt)^{-1} \left[ \int \delta^2 \right] + n^{-2} t^{-1} \\
\times \sum_{i \neq j} \delta^2((X_i - X_j)/t) \\
= (nt)^{-1} \left[ \int \delta^2 \right] + (tT)^{-1} \hat{K}_{\text{cum}}(t)
\]
and that, by (1.2),
\[
-2n^{-1} \sum_{i = 1}^{n} \hat{f}_i(X_i) = -2(tT)^{-1} \hat{K}_{\text{cum}}(t).
\]

Lemma 3.1 now follows from the definitions (1.1) and (1.5).

Proof of Lemma 3.2. First observe that, conditioned on the process \( \Lambda \),
\[
\text{E}[\hat{\lambda}_i(x) \mid \Lambda] - \Lambda(x) \Bigl[ \text{E}[\hat{\lambda}_i(x) \mid \Lambda] - \Lambda(x) \Bigr]^2
\]
\[
= \left( \int \delta(x - u) \Lambda(u) \, du - \Lambda(x) \right)^2 \\
= \int \delta(x - u) \Lambda(u) \Lambda(u) \, du \\
- 2 \int \delta(x - u) \Lambda(u) \Lambda(u) \, du + \Lambda(x)^2 \\
\]
and var\[\hat{\lambda}_i(x) \mid \Lambda] = \int \delta(x - u)^2 \Lambda(u) \, du. \text{ Thus}
\]
\[
\text{MSE}(t) = \text{E}(\text{var}[\hat{\lambda}_i(x) \mid \Lambda] + (\text{E}[\hat{\lambda}_i(x) \mid \Lambda] - \Lambda(x))^2) \\
= \mu \int \delta(x)^2 \, dx \\
+ \int \delta(x - u) \delta(x - v) \nu(|u - v|) \, du \, dv
\]
where

\[
\begin{align*}
\delta(x - u|v(x - u)) &= \mu^{-1} \left[ \delta^2 + \mu t^{-1} K_{\gamma}(t) \right] + \mu^2 t^{-1} K_{\delta}(t) + v(0) \\

\text{Proof of Lemma 3.3.} & \quad \text{Note that} \\
b^2(t) &= t^{-1} \left[ 2 \int_0^\infty \delta^2(u/t) v_0(u) \, du \right] \\
&\quad - 2 t^{-1} \left[ 2 \int_0^\infty \delta(u/t) v_0(u) \, du \right] + v_0(0) \\
&\quad = \int_\infty^\infty [\delta^2(v) - 2 \delta(v)] v_0(\{ut\}) \, dv + v_0(0) \\
&\quad = \int_\infty^\infty [\delta(v) - 2 \delta(v)] \\
&\quad \times \left[ \sum_{j=0}^4 \nu_j(0) v_i(0) \right]^{ij} + o(t^4) \, dv + v_0(0), \\
\text{where } v(x) &= (d/dx)v(|x|). \text{ Thus} \\
b^2(t) &= \sum_{j=0}^4 C_i \nu_j(0) \nu_i(0) + o(t^4), \quad (A.2) \\
\end{align*}
\]

where

\[
C_i = \int \left[ \delta^2(v) - 2 \delta(v) \right] \, dv \\
= \int \left[ \delta(v - z) \, dv \delta(z) \, dz - \int \delta(v) \, dv \right] \\
= \int \left[ \delta(v - z) \, dv \delta(z) \, dz \right] - \int \delta(v) \, dv \\
= \sum_{j=0}^4 \left( \frac{1}{j!} \right) \int \left[ \delta^j(v) \, dv \right] \int \left[ \delta^j(z) \, dz \right] - \int \delta(v) \, dv.
\]

But \( \delta \) is a probability density that is symmetric about the origin, so \( C_0 = -1, C_j = 0 \) \( (j = 1, 2, 3) \), and \( C_4 = 6! \int \delta^4(u) \, du \). Lemma 3.3 now follows on applying this to (A.2).

**Proof of Lemma 4.1.** Working as in the proof of Lemma 3.2, we first condition on the process \( \Lambda \) and get the same expression for the conditional squared bias. This time the conditional variance is

\[
\text{var} \left[ f_j(x) \mid \Lambda \right] = n^{-1} \int \delta_j(x - u) \Lambda(u) \, du \\
- n^{-1} \left[ \int \delta_j(x - u) \Lambda(u) \, du \right]^2.
\]

Hence,

\[
\text{MSE}^*(t) = (nT)^{-1} \int \delta(v)^2 \, dv + (1 - n^{-1}) \\
\times \int \delta_j(x - u) \delta_j(x - v) \nu(|u - v|) \, du \, dv \\
- 2 \int \delta_j(x - u) \nu(|x - u|) \, du + v(0) \\
= (nT)^{-1} \int \delta^2 \left[ (1 - n^{-1}) t^{-1} K_{\gamma}(t) \right] \\
- 2 t^{-1} K_{\delta}(t) + v(0).
\]

**Proof of Lemma 5.1.** Since \( \delta \) is supported on \([-1, 1]\) and \( t < t_0 \),

\[
E \left[ \sum_{i=1}^{n_i(t)} \int_0^{T-t_0} \delta(x - X_i) \lambda(x) \, dx \right] \\
= E \left[ \sum_{i=1}^{n_i(t)} \int_0^{T-t_0} \delta(x - y) \lambda(y) \, dy \right] \\
= \int_0^{T-t_0} \int_{-\infty}^{\infty} \delta(x - y) \lambda(y) \lambda(z) \, dx \, dy \\
= E \left[ \sum_{i=1}^{n_i(t)} \int_0^{T-t_0} \delta(x - y) \lambda(y) \lambda(z) \, dx \, dy \right] \\
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x - y) \lambda(y) \lambda(z) \, dx \, dy \\
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x - y) \lambda(y) \lambda(z) \, dx \, dy \\
= (T - 2t_0) \int_0^{\infty} \delta(u) \nu(u) \, du \\
\]

which is the first part of Lemma 5.1.

To prove the second part, note that

\[
E \left[ \sum_{i=1}^{n_i(t)} \int_0^{T-t_0} \delta(x - X_i) \delta(x - X_j) \, dx \right] \\
= E \left[ \sum_{i=1}^{n_i(t)} \int_0^{T-t_0} \delta(x - y) \delta(x - z) \lambda(y) \lambda(z) \, dx \, dy \right] \\
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x - y) \delta(x - z) \lambda(y) \lambda(z) \, dx \, dy \\
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x - y) \delta(x - z) \lambda(y) \lambda(z) \, dx \, dy \\
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x - y) \delta(x - z) \lambda(y) \lambda(z) \, dx \, dy \\
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x - y) \delta(x - z) \lambda(y) \lambda(z) \, dx \, dy \\
= (T - 2t_0) \int_0^{\infty} \delta(u) \nu(u) \, du.
\]

Finally, observe that

\[
E \left[ \sum_{i=1}^{n_i(t)} \delta(x - X_i) \lambda_0(t_0 - t_0(X_i)) \right] \\
= E \left[ \sum_{i=1}^{n_i(t)} \int_0^{T-t_0} \delta(x - y) \lambda(x) \lambda(y) \, dx \, dy \right] \\
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x - y) \lambda(x) \lambda(y) \, dx \, dy \\
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(x - y) \lambda(x) \lambda(y) \, dx \, dy \\
= (T - 2t_0) \int_0^{\infty} \delta(u) \nu(u) \, du,
\]

which completes the proof of Lemma 5.1.

[Received July 1987. Revised December 1987.]

**REFERENCES**


