Equivalence of smoothing parameter selectors in density and intensity estimation

by

Peter Diggle
CSIRO Division of Mathematics
and Statistics, Canberra

J.S. Marron[1]
Australian National University
and University of North Carolina

June 1987

AMS 1980 subject classification: primary 62G05, secondary 60G35.

Key words and phrases: Cox process, Cross-validation, Kernel estimators, Nonparametric density estimation, Poisson intensity estimation, Smoothing

[1]Research partially supported by NSF Grant DMS-8400602.

Summary

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, i.e. 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 which minimizes an estimate of the mean squared error under the assumption that the data are generated by a stationary Cox process. This paper 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 esimation is that this equivalence makes it clear how the Cox process method may be applied to kernels which 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 to the intensity setting. A benefit for density estimation is that it motivates an analogue 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 which 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.

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 $\lambda(x)$.

A reasonable estimate of either f or λ should be a function which takes on large values in regions where the data are dense, and values close to zero where the data are sparse. The kernel smoothing method of constructing such a function is to let

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

for density estimation, or

$$\hat{\lambda}_t(x) = \sum_{i=1}^{n} \delta_t(x-X_i),$$

for intensity estimation, where $\delta_t(\cdot) = \frac{1}{t}\,\delta(\cdot/t)$ 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.

Figure 1

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 $\hat{\lambda}_t(x)$. The data are the times of 191 coal-mining disasters

in a total time-period of 40550 days, as compiled by Jarrett (1979). Rudemo (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,

$$\delta(x) = \begin{cases} 0.9375(1-x^2)^2 & : -1 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Figure (1) shows three estimates $\hat{\lambda}_t(x)$ with a four-fold increase in t between each pair of estimates and the intermediate value t = 5957.3 the optimum value according to the method of bandwidth selection described in Diggle (1985).

Note that the estimate with t = 1489.3 oscillates wildly and contains many features that one can not expect to distinguish with only 191 observations. On the other hand, the curve with t = 23829.2 has lost features which 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.

It is generally the case that 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).

For density estimation, Rudemo (1982) and Bowman (1984) have proposed choosing t by the method of least squares cross-validation. This essentially entails letting $\hat{t}_{cv}$ denote the minimizer of the cross-validation score

$$CV(t) = \int[\hat{f}_t(x)]^2 dx - 2\, n^{-1} \sum_{j=1}^{n} \hat{f}_{t,j}(X_j), \tag{1.1}$$

where $\hat{f}_{t,j}(x)$ is the leave one out estimator

$$\hat{f}_{t,j}(x) = n^{-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} \delta_t(x - X_i). \tag{1.2}$$

The motivation for the form of CV(t) is that it provides an intuitively reasonable estimate of the first two terms of

$$\int \hat{f}_t^2 - 2\int \hat{f}_t f + \int f^2 = \int (\hat{f}_t - f)^2.$$

See Stone (1984), Marron (1987) and Hall & Marron (1987a,b) for properties of $\hat{t}_{CV}$ and an access to additional literature. This particular form of CV(t) is treated in Stone (1984). Several slightly different, but essentially equivalent forms may be found elsewhere.

For intensity estimation, Diggle (1985) has proposed choosing t by the following empirical Bayesian method. First suppose that the intensity $\lambda(x)$ is a realization of a stationary, nonnegative valued random process $\{\Lambda(x) : x \in \mathbb{R}\}$, with $E[\Lambda(x)] = \mu$ and $E[\Lambda(x)\Lambda(y)] = \upsilon(|x-y|)$. Then, $X_1,\ldots,X_n$ form a partial realization of a Cox process, or doubly stochastic Poisson process. See, for example, Cox & Isham (1980, pp. 70–75). Letting E denote expectation over both the randomness in $X_1,\ldots,X_n$ and in $\Lambda(x)$, Diggle shows that, for x more than t units form the boundary of [0,T],

$$(1.3) \qquad MSE(t) \equiv E([\hat{\lambda}_t(x) - \Lambda(x)]^2)$$

$$= \upsilon(0) + \mu[1 - 2\mu K(t)]/2t + (\mu/2t)^2 \int_0^{2t} K(y)dy,$$

where

$$K(t) = 2\mu^{-2} \int_0^t \upsilon(x)dx,$$

and where the special case of the uniform kernel

$$(1.4) \qquad \delta(\cdot) = \frac{1}{2} 1_{[-1,1]}(\cdot),$$

has been used in the estimator $\hat{\lambda}_t$. The function K(t) is very useful for the analysis of spatial point processes; see Ripley (1981) and Diggle (1983). In the one-dimensional case it can be estimated by

$$\hat{K}(t) = T\, n^{-2} \sum_{i \neq j} \sum 1_{[-t,t]}(X_i - X_j).$$

Hence, the bandwidth which minimizes the mean squared error (1.3) should be fairly well approximated by the bandwidth, $\hat{t}_M$, which minimizes

$$\hat{M}(t) = (2\hat{\mu}t)^{-1} - t^{-1}\hat{K}(t) + (2t)^{-2}\int_0^{2t} \hat{K}(y)dy,$$

where μ has been estimated by

(1.5) $$\hat{\mu} \equiv n/T.$$

It should be noted that in all of the above, no attempt has been made to adjust for boundary effects. These adjustments are very important, but 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 $\hat{t}_{CV}$ and $\hat{t}_M$ are both well defined in either setting. The surprising fact that they are exactly the same, under no additional assumptions is contained in

Theorem 2.1: *In the case of the uniform kernel, (1.4),*

$$\hat{t}_{CV} = \hat{t}_M,$$

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

Lemma 2.2: *For the uniform kernel,*

$$\hat{M}(t) = T \cdot CV(t).$$

The proof of Lemma 2.2 is in the appendix.

The first consequence of Theorem 2.1 is that it shows each method is more natural than previously suspected, because it can be derived from much different ideas than those originally used. However, the benefits go far deeper than this,

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 $\hat{t}_M$ appeared to apply only to this case. However, the motivation for $\hat{t}_{CV}$ works for any kernel, so this suggests using Lemma 2.2 to find an appropriate $\hat{M}(t)$ for the general case. In particular define

$$\hat{M}(t) \equiv T \cdot CV(t)$$

A form of $\hat{M}(t)$ which will allow a Mean Square Error interpretation, is given by:

Lemma 3.1 *For a general* δ,

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

where

$$\hat{K}_{\delta}(t) = n^{-2}\, t\, T \sum_{i \neq j} \sum \delta_t(X_i - X_j),$$

$$\delta*\delta(\cdot) = \int \delta(\cdot - x)\delta(x)dx,$$

and $\hat{\mu}$ *is defined at* (1.5).

The proof of Lemma 3.1 is in the appendix.

Next it should be verified that choosing $\hat{t}_n$ 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, whose proof is also in the appendix. Recall that boundary effects are ignored until section 5.

Lemma 3.2:

$$MSE(t) = \mu t^{-1}[\int \delta^2] + \mu^2 t^{-1} K_{\delta*\delta}(t) - \mu^2 2t^{-1} K_\delta(t) + \nu(0),$$

where

$$K_\delta(t) = 2t \mu^{-2} \int_0^\infty \delta_t(u)\nu(u)du.$$

Since $\hat{K}_\delta(t)$ is a reasonable estimate of $K_\delta(t)$, the minimizer of $\mu^2\hat{M}(t) + \nu(0)$ (which also minimizes $M(t)$) should be close to the minimizer of $MSE(t)$. Note that when $\delta(\cdot)$ is the uniform kernel, $K_\delta(t)$ differs from $K(t)$ by a factor of two.

A very useful tool in the study of density estimation is asymptotics with $n\to\infty$. The above strong connection between bandwidth selectors in the two settings motivates looking for an analogue in the intensity estimation case. The first type of asymptotics that one might consider is $T\to\infty$. However, as pointed out at the end of section 2.2 of Diggle (1985), this will only add new information at the right hand 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 μ was the mean of the underlying stochastic process introduced at the end of section 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 $\nu(x)$ (defined in section 1) to be of the form

$$\nu(x) = \mu^2 \nu_0(x), \tag{3.1}$$

for some fixed function $\nu_0(x)$. For insight into the effects of $\mu\to\infty$ with ν 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

$$\Lambda(x) = \sum_{i=1}^{\infty} h(x-Z_i), \tag{3.2}$$

where the Z_i are the points of a homogeneous Poisson process and $h(\cdot)$ is a non-negative 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, i.e.

the fact that t too small admits too much sample noise while t too large smooths away features of the underlying curve. In particular, see Rosenblatt (1971) for example, various squared error criteria can be expanded into simple and easily interpreted variance and squared bias components which become large when t is small and large respectively.

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

$$\mu^{-2}MSE(t) = \mu^{-1}t^{-1}[\int\delta^2] + t^{-1}K_{\delta*\delta}(t) - 2t^{-1}K_{\delta}(t) + v_0(0).$$

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

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

$$b^2(t) = t^{-1}K_{\delta*\delta}(t) - 2t^{-1}K_{\delta}(t) + v_0(0).$$

To gain more insight into the behaviour of the squared bias, consider the expansion summarized in the following lemma, whose proof is in the appendix.

Lemma 3.3: *If $v_0(|x|)$ has a continuous fourth derivative at the origin, then as* $t \to 0$,

$$b^2(t) = t^4 v_0^{(4)}(0)[\int u^2\delta(u)du/2]^2 + o(t^4).$$

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 $v_0^{(4)}(0)$ replaces $[f^{(2)}(x)]^2$. This is not surprising in view of the well known relationship between the k-th derivative of a process $\Lambda(x)$ and the 2k-th 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 μ. 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 (1962) demonstrated that if a density has more than 2 derivatives, say k, then a faster rate of convergence of $\hat{f}_t$ to f can be obtained by using a "higher order kernel", i.e., assuming that

$$\int x^{\ell}\delta(x)dx = \begin{cases} 1 & \text{if } \ell=0 \\ 0 & \text{if } \ell=1,2,\ldots,k-1 \\ C>0 & \text{if } \ell=k \end{cases} \tag{3.3}$$

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

Lemma 3.4: *If δ satisfies (3.3) and $v_0(|x|)$ has a continuous 2k-th derivative at the origin, then as t→0,*

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

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

The above 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, that will require much more work than can be done in this paper, include analogs of:

(i) the optimal rates ideas of Farrell (1972) and Stone (1980) for example.

(ii) the asymptotic optimality results of Stone (1984) and Marron (1987) for example.

(iii) The noise in bandwidth selection ideas of Hall and Marron (1987a,b).

(iv) the location dependent smoothing ideas of Abramson (1982) and Hall and Marron (1986a).

(v) the kernel selection ideas of Epanechnikov (1969) and Hall and Marron (1986b).

4. Benefits for density estimation

The equivalence described in Section 2 motivates consideration of the density estimation problem from the empirical Bayes type of viewpoint considered in Diggle (1985). 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 Λ, with $E[\Lambda(x)] = T^{-1}$ and $E[\Lambda(x)\Lambda(y)] = \upsilon(|x-y|)$. An example of such a process Λ is a normalized vesion of the "linear Cox process" described in Section 2.2 of Diggle (1985) and at (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 square error, $MSE^*_t(t)$, of the estimator $\hat{f}_t(x)$ is contained in

Lemma 4.1: *In the Cox process density estimation setting*

$$MSE^*(t) = (ntT)^{-1}[\int \delta^2] + (1-n^{-1})t^{-1}T^{-2}K^*_{\delta*\delta}(t) - 2t^{-1}T^{-2}K^*_\delta(t) + \upsilon(0),$$

where

$$K^*_\delta(t) = 2tT^2\int_0^\infty \delta_t(u)\upsilon(u)du.$$

The proof of Lemma 4.1 is in the appendix. Note that when μ 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, $K_\delta^*(t)$ is, by considerations similar to those above, reasonably estimated by

$$\hat{K}_\delta^*(t) = n^{-1}(n-1)^{-1} t T \sum_{i \neq j} \sum \delta_t(X_i - X_j).$$

Hence, the minimizer of

$$\hat{M}^*(t) = (nt)^{-1} T[f\delta^2] + (1-n^{-1})t^{-1} \hat{K}_{\delta * \delta}(t) - 2t^{-1} \hat{K}_\delta^*(t)$$

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

Lemma 4.2:

$$\hat{M}(t) = T \cdot CV^*(t),$$

where

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

$$\hat{f}_{t,j}^*(x) = (n-1)^{-1} \sum_{\substack{i=1 \\ i \neq j}}^{n} \delta_t(x - X_i).$$

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, $\hat{K}_\delta^*$ and $\hat{f}_{t,j}^*$, and their original analogs $\hat{K}_\delta$ and $\hat{f}_{t,j}$, 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 $CV^*(t)$ appears more often in the literature than $CV(t)$, see Rudemo (1982), Marron (1987), Hall and Marron (1987a). The calculations

done in this section make 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 $\hat{\lambda}_t(x)$ 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 $\hat{\lambda}_t(0)$ and $\hat{\lambda}_t(T)$ 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 zero on an interval and equal to zero off it. See Rice (1984), Schuster (1985), Gasser, Muller 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" type of adjustment considered by Schuster (1985) and Cline and Hart (1986). We illustrate the method by showing how $\hat{\lambda}_t$ can be adjusted at 0. Adjustments at T and for density estimation are similar. The basic idea is that the piece of each $\delta_t(\cdot - X_i)$ 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

$$\hat{\lambda}_t^+(x) = \sum_{i=1}^{n} [\delta_t(x-X_i) + \delta_t(x+X_i)]\, 1_{[0,T]}(x). \tag{5.1}$$

Another possible boundary adjustment is given in (1.1) of Diggle (1985). 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 adjustments 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 2.3 of Diggle (1985). Hence, one sensible method of adjusting for boundary effects is to use the estimator $\hat{\lambda}_t^+$, with the bandwidth t chosen to minimize $\hat{M}(t)$ as defined in Diggle (1985). Similar remarks hold for density estimation.

Figure 2

A drawback of the above recipe is that for the bandwidth selection part, the boundary will still have a tendency to introduce a bias towards 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, Muller and Mammitzsch (1985).

Another means of approaching the bias towards 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

$$\int_{t_0}^{T-t_0} \hat{f}_t^2 - 2\int_{t_0}^{T-t_0} \hat{f}_t f + \int_{t_0}^{T-t_0} f^2 = \int_{t_0}^{T-t_0} [\hat{f}_t - f]^2,$$

by

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

It is important to note that all of $X_1, \ldots, X_n$ are used to construct the estimates $\hat{f}_t$ and $\hat{f}_{t,j}$, but only those inside $[t_0, T-t_0]$ appear in the sum inside CV^+. 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 2.2 motivates looking for another interpretation of

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

Much of the work in this is summarized in:

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

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

$$E\left[\sum_{i\neq j}\sum \int_{t_0}^{T-t_0} \delta_t(x-X_i)\,\delta_t(x-X_j)dx\right] = (T-2t_0)\,\mu^2 t^{-1}\,K_{\delta*\delta}(t),$$

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

The proof of Lemma 5.1 is in the appendix.

It now follows from Lemma 3.2 that $\hat{M}^+(t)$ is essentially estimating, with proper attention paid to boundary effects

$$(T-2t_0)T^{-1}\mu^{-2}[MSE(t) - \nu(0)].$$

To get some feeling for the performance of CV^+, consider Table 1, which shows the approximate value of $\hat{t}_M^+$, the minimizer of $\hat{M}^+(t)$, for several values of t_0.

Table 1

t_0	0	2000	4000	6000	8000	10,000
$\hat{t}_M^+$	6000	6100	6200	5800	6000	5,300

Note that the values are quite stable until t_0 becomes 10,000. To understand this, note that in Figure 1, when t_0 approaches 10,000 two quite pronounced peaks are removed from the interval $[t_0, T-t_0]$, which has a large effect on the amount of information available to $\hat{t}_M^+$. The practical difference between the bandwidths in Table 1 is demonstrated by Figure 3. The fact that the curve corresponding to t_0=10,000 represents a materially different amount of smoothing is clearly demonstrated by the extra small mode near t=2000.

Figure 3

Appendix

Proof of Lemma 2.2: First note that

$$\int_0^{2t} \hat{K}(y)dy = T\, n^{-2} \sum_{i\neq j} \sum \int_0^{2t} 1_{[-y,y]}(X_i - X_j)dy$$

$$= T^{-1} n^{-2} \sum_{i\neq j} \sum (2t - |X_i - X_j|) 1_{[-2t,2t]}(X_i, X_j).$$

Hence

$$(A.1)\quad \int [\hat{f}_t(x)]^2 dx = n^{-2}\left[\sum_{i=1}^{n} \int \delta_t(x-X_i)^2 dx + \sum_{i\neq j} \sum \int \delta_t(x-X_i)\, \delta_t(x-X_j)dx\right]$$

$$= \frac{1}{4} n^{-2} t^{-2}\left[\sum_{i=1}^{n} \int 1_{[X_i-t, X_i+t]}(x)dx + \sum_{i\neq j} \sum \int 1_{[X_i-t, X_i+t]}(x)\, 1_{[X_j-t, X_j+t]}(x)dx\right]$$

$$= \frac{1}{4} n^{-2} t^{-2}\left[2\, nt + \sum_{i\neq j} \sum (2t - |X_i - X_j|) 1_{[-2t,2t]}(X_i - X_j)\right]$$

$$= (2nt)^{-1} + T^{-1}(2t)^{-2} \int_0^{2t} \hat{K}(y)dy$$

Thus,

$$CV(t) = \int [\hat{f}_t(x)]^2 dx - 2\, n^{-2} \sum_{i\neq j} \sum (2t)^{-1} 1_{[-t,t]}(X_i - X_j)$$

$$= (2nt)^{-1} + T^{-1}(2t)^{-2} \int_0^{2t} \hat{K}(y)dy - (tT)^{-1} \hat{K}(t)$$

$$= T^{-1} \hat{M}(t).$$

Proof of Lemma 3.1: First note that (A.1) becomes

$$\int [\hat{f}_t(x)]^2 dx = (nt)^{-1}[\int \delta^2] + n^{-2}\, t^{-1} \sum_{i\neq j} \sum \delta * \delta((X_i - X_j)/t)$$

$$= (nt)^{-1}[\int \delta^2] + (tT)^{-1}\,\hat{K}_{\delta*\delta}(t).$$

and that, by (1.2)

$$-2n^{-1}\sum_{j=1}^{n}\hat{f}_{t,j}(X_j) = -2(tT)^{-1}\hat{K}_{\delta}(t).$$

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 Λ,

$$(E[\hat{\lambda}_t(x)\,|\Lambda] - \Lambda(x))^2 = (\int \delta_t(x-u)\Lambda(u)du - \Lambda(x))^2 =$$

$$= \iint \delta_t(x-u)\delta_t(x-v)\Lambda(u)\Lambda(v)dudv$$

$$-2\int \delta_t(x-u)\Lambda(u)\Lambda(x)du + \Lambda(x)^2,$$

and

$$\mathrm{var}[\hat{\lambda}_t(x)\,|\Lambda] = \int \delta_t(x)^2\Lambda(x)dx.$$

Thus

$$\mathrm{MSE}(t) = E(\mathrm{var}[\hat{\lambda}_t(x)\,|\Lambda] + (E[\hat{\lambda}_t(x)\,|\Lambda] - \Lambda(x))^2)$$

$$= \mu\int \delta_t(x)^2dx + \iint \delta_t(x-u)\delta_t(x-v)v(|u-v|)dudv$$

$$-2\int \delta_t(x-u)v(|x-u|)du + v(0)$$

$$= \mu\, t^{-1}[\int \delta^2] + \mu^2\, t^{-1}\, K_{\delta*\delta}(t) - \mu^2\, 2t^{-1}K_{\delta}(t) + v(0).$$

Proof of Lemma 3.3: Note that

$$b^2(t) = t^{-1}\left[2\int_0^{\infty} \delta*\delta(u/t)v_0(u)du\right] - 2t^{-1}\left[2\int_0^{\infty} \delta(u/t)v_0(u)du\right] + v_0(0)$$

$$= \int_{-\infty}^{\infty} [\delta*\delta(v) - 2\delta(v)]\; v_0(|vt|)dv + v_0(0)$$

$$= \int_{-\infty}^{\infty} [\delta * \delta(v) - 2\delta(v)] \left[\sum_{j=1}^{4} v_j(0)(vt)^j/j! + o(t^4) \right] dv + v_0(0)$$

where

$$v_j(x) = \frac{d^j}{dx^j} v(|x|)$$

Thus

$$b^2(t) = \sum_{j=0}^{4} C_j t^j v_j(0)/j! + v_0(0) + o(t^4) \tag{A.2}$$

where

$$C_j = \int v^j [\delta * \delta(v) - 2\delta(v)] dv$$

$$= \iint v^j \delta(v-z)dv \, \delta(z)dz - 2 \int v^j \delta(v)dv$$

$$= \iint (u+z)^j \delta(u) \, \delta(z)dudz - 2 \int v^j \delta(v)dv$$

$$= \sum_{\ell=1}^{j} \binom{j}{\ell} \left[\int u^\ell \delta(u)du \right] \left[\int z^{j-\ell} \delta(z)dz \right] - 2 \int v^j \delta(v)dv.$$

But δ is a probability density which is symmetric about the origin, so

$$C_0 = -1,$$

$$C_j = 0, \quad j=1,2,3,$$

$$C_4 = 6\left[\int u^2 \delta(u)du\right]^2 .$$

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 above, we first condition on the process Λ and get the same expression for the conditional squared bias. This time the conditional variance is

$$\mathrm{var}[\hat{f}_t(x)|\Lambda] = n^{-1} \int \delta_t(x-u)^2 \Lambda(u)du - n^{-1}\left[\int \delta_t(x-u)\Lambda(u)du\right]^2$$

Hence,

$$MSE^{*}(t) = (nT)^{-1} \int \delta_t(v)^2 dv + (1-n^{-1}) \iint \delta_t(x-u)\delta_t(x-v)v(|u-v|)dudv$$
$$-2 \int \delta_t(x-u)v(|x-u|)du + v(0)$$

$$= (ntT)^{-1}[\int \delta^2] + (1-n^{-1})\, t^{-1} K_{\delta*\delta}^{(*)}(t) - 2t^{-1} K_{\delta}^{(*)}(t) + v(0).$$

Proof of Lemma 5.1: Since δ is supported on $[-1,1]$ and $t < t_0$,

$$E\left[\sum_i \int_{t_0}^{T-t_0} \delta_t(x-X_i)^2 dx\right] = E\left[\int_0^T \int_{t_0}^{T-t_0} \delta_t(x-y)^2 \Lambda(y)dxdy\right]$$

$$= \int_{t_0}^{T-t_0} \int_{-\infty}^{\infty} \delta_t(x-y)^2\, \mu\, dydx$$

$$= (T-t_0)\, \mu\, t^{-1}[\int \delta^2],$$

which is the first part of Lemma 5.1.

To prove the second part, note that

$$E\left[\sum_{i\neq j} \sum \int_{t_0}^{T-t_0} \delta_t(x-X_i)\delta_t(x-X_j)dx\right] =$$

$$= E\left[\int_0^T \int_0^T \int_{t_0}^{T-t_0} \delta_t(x-y)\delta_t(x-z)\Lambda(y)\Lambda(z)dx\, dy\, dz\right]$$

$$= \int_{t_0}^{T-t_0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta_t(x-y)\delta_t(x-z)v(|y-z|)dy\, dz\, dx$$

$$= \int_{t_0}^{T-t_0} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta_t(x-z-u)\delta_t(x-z)v(|u|)du\, dz\, dx$$

$$= (T-2t_0)\ 2 \int_0^\infty (\delta * \delta)_t(u)\ v(u)du.$$

Finally observe that

$$E\left[\sum_{i \neq j} \sum \delta_t(X_i - X_j) 1_{[t_0, T-t_0]}(X_j)\right] = E\left[\int_{t_0}^{T-t_0} \int_0^T \delta_t(x-y)\Lambda(x)\Lambda(y)dx\ dy\right]$$

$$= \int_{t_0}^{T-t_0} \int_{-\infty}^{\infty} \delta_t(x-y)v(|x-y|)dx\ dy$$

$$= (T-2t_0)\ 2 \int_0^\infty \delta_t(u)v(u)\ du,$$

which completes the proof of Lemma 5.1.

REFERENCES

Abramson, I. S. (1982), "On bandwidth variation in kernel estimates - a square root law," *Annals of Statistics*, 10, 1217-1223.

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

Cline, D. and Hart, J. (1986), "Kernel estimation of densities with discontinuities or discontinuous derivatives," unpublished manuscript.

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

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

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

Diggle, P. (1985), "A kernel method for smoothing point process data," *Applied Statistics*, 34, 138-147.

Ellis, S.P. (1986), "A limit theorem for spatial point processes," *Advance in Applied Probabilities*, 18, 649-659.

Epanechnikov, V. A. (1969) "Nonparametric estimation of a multivariate probability density", *Theory of Probability and Its Applications*, 14, 153-158.

Farrell, R. H. (1972), "On the best obtainable rates of convergence in estimation of a density function at a point," *Annals of Mathematical Statistics*, 43, 170-180.

Gasser, T., Müller, H. G. and Mammitzsch, V. (1985), "Kernels for nonparametric curve estimation," *Journal of the Royal Statistical Society, Series B*, 47, 238-252.

Hall, P. and Marron, J. S. (1986a), "Choice of kernel order in density estimation", unpublished manuscript.

Hall, P. and Marron, J. S. (1987a), "Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation," *Theory of Probability and Related Fields*, to appear.

Jarrett, R.G. (1979), A note on the intervals between coal-mining disasters. *Biometrika*, 66, 191-3.

Leadbetter, M.R. and Wold, D. (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, J. S. (1987), "A comparison of cross-validation methods in density estimation," *Annals of Statistics*, 15 (to appear).

Parzen, E. (1962), "On estimation of a probability density function and mode," *Annals of Mathematical Statistics*, 33, 1065-1076.

Rice, J. (1984) "Boundary modification for kernel regression," *Communications in Statistics, Series A*, 13, 893-900.

Ripley, (1981), *Spatial Statistics*, New York: Wiley.

Rosenblatt, M. (1971), "Curve estimates," *Annals of Mathematical Statistics*, 42, 1815-1842.

Rudemo, M. (1982), "Empirical choice of histograms and kernel density estimators," *Scandanavian Journal of Statistics*, 9, 65-78.

Schuster, E.A. (1985) "Incorporating support constraints into nonparametric estimators of densities," *Communications in Statistics, Series A*, 14, 1123-1136.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

Stone, C. J. (1980), "Optimal convergence rates for nonparametric estimators," *Annals of Statistics*, 8, 1348-1360.

Stone, C. J. (1984), "An asymptotically optimal window selection rule for kernel density estimates," *Annals of Statistics*, 12, 1285-1297.

Figure 1 : Effect of band-width on kernel estimates for coal-mining disaster data

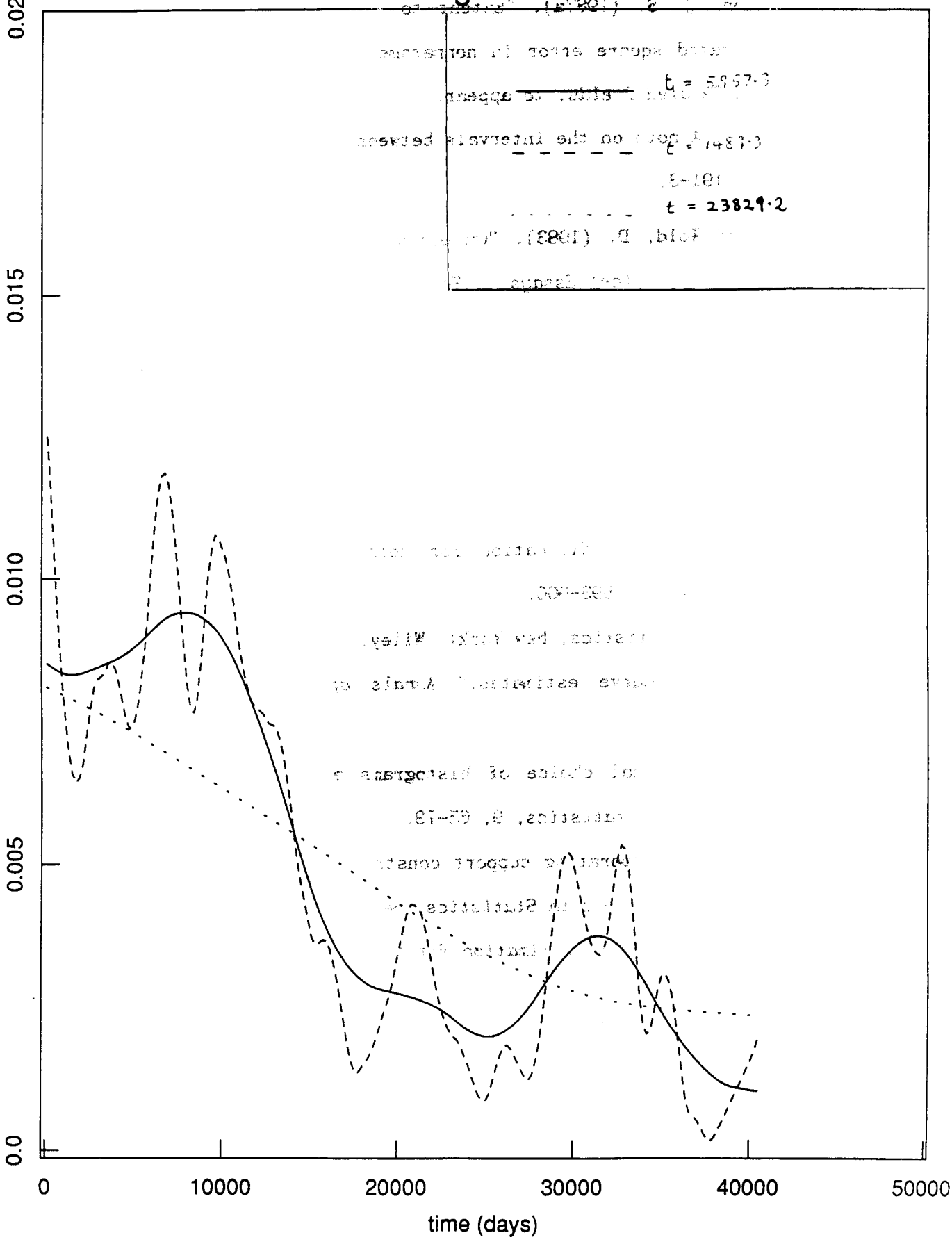

Figure 2 : Effect of boundary adjustments on kernel estimates for coal-mining disaster data

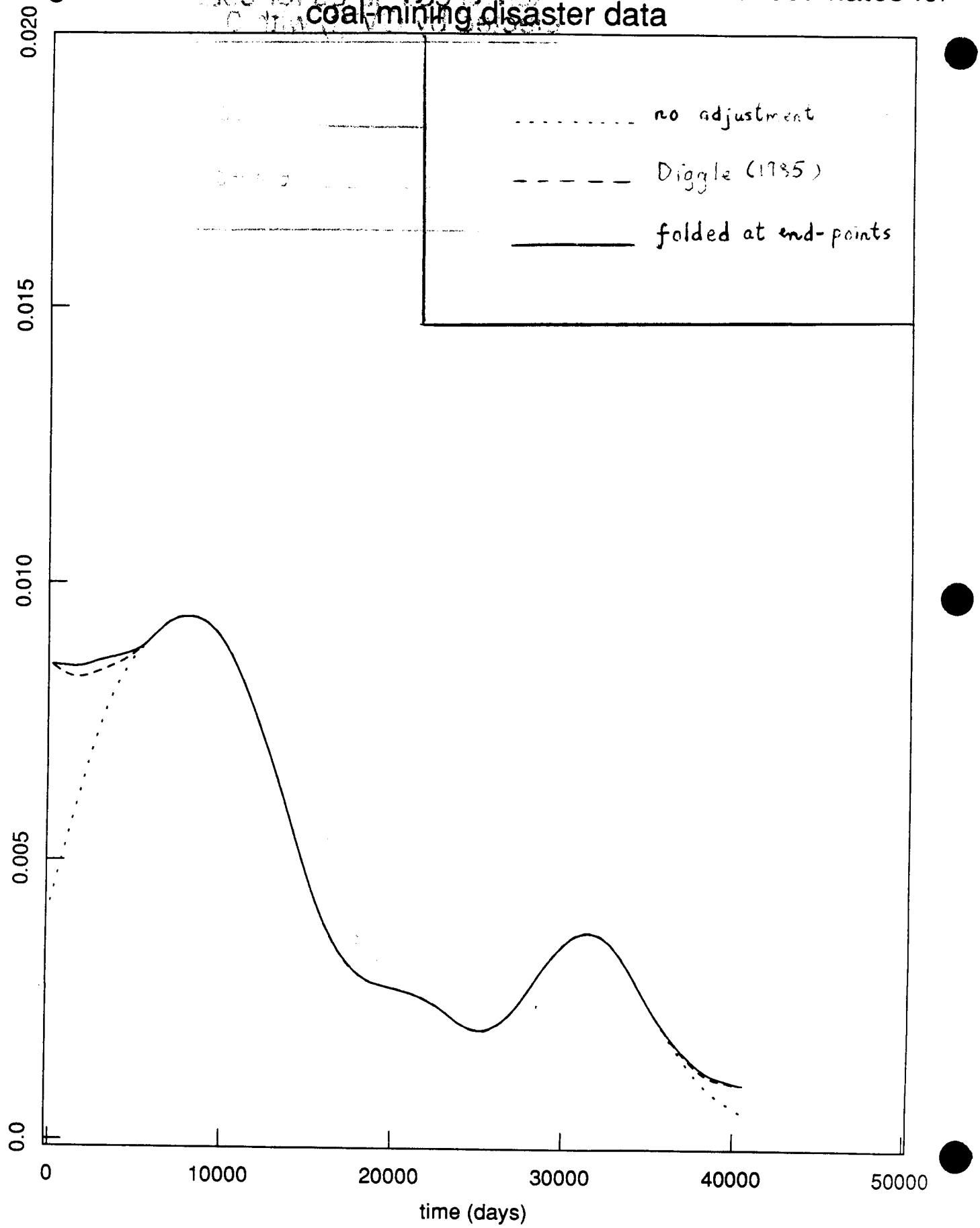

Figure 3 : Extreme kernel estimates for coal-mining disaster data, as selected by $CV^{*}(t_0)$ with $0 \leq t_0 \leq 10000$