



Spline Smoothing: The Equivalent Variable Kernel Method

Author(s): B. W. Silverman

Source: *The Annals of Statistics*, Vol. 12, No. 3 (Sep., 1984), pp. 898-916

Published by: Institute of Mathematical Statistics

Stable URL: <https://www.jstor.org/stable/2240968>

Accessed: 20-04-2020 20:06 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



JSTOR

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve and extend access to *The Annals of Statistics*

SPLINE SMOOTHING: THE EQUIVALENT VARIABLE KERNEL METHOD

BY B. W. SILVERMAN

University of Bath

The spline smoothing approach to nonparametric regression and curve estimation is considered. It is shown that, in a certain sense, spline smoothing corresponds approximately to smoothing by a kernel method with bandwidth depending on the local density of design points. Some exact calculations demonstrate that the approximation is extremely close in practice. Consideration of kernel smoothing methods demonstrates that the way in which the effective local bandwidth behaves in spline smoothing has desirable properties. Finally, the main result of the paper is applied to the related topic of penalized maximum likelihood probability density estimates; a heuristic discussion shows that these estimates should adapt well in the tails of the distribution.

1. Introduction. Consider the nonparametric regression problem of estimating a curve g given observations $Y_i = g(t_i) + \varepsilon_i$, $i = 1, \dots, n$. Assume that the design points t_i are known and not necessarily evenly spaced, and that the ε_i are random errors. The cubic spline estimator \hat{g} of the regression curve is defined to be the minimizer over functions g of

$$(1.1) \quad \lambda \int g''(t)^2 dt + n^{-1} \sum_{i=1}^n \{Y_i - g(t_i)\}^2.$$

The parameter $\lambda > 0$ is a smoothing parameter which controls the trade-off between smoothness, as measured by $\int g''^2$, and goodness of fit to the data, as measured by the sum of squares of deviations between $g(t_i)$ and Y_i . The larger the value of λ , the more the data will be smoothed to produce the curve estimate. This form of the spline smoothing method is due to Schoenberg (1964) and Reinsch (1967) but the basic underlying idea, of penalizing a measure of goodness of fit by one of roughness, was described by Whittaker (1923). For more recent work and bibliography on cubic smoothing splines see, for example, De Boor (1978, Chapter 14), Craven and Wahba (1979), Rice and Rosenblatt (1983), and Silverman (1984b, 1985). These last two papers include discussion of some of the ramifications of the work developed below.

It is well known (cf. equation (2.2) of Wahba, 1975) and easily shown from the quadratic nature of (1.1) that the spline smoother \hat{g} is linear in the observations Y_i , in the sense that there exists a weight function $G(s, t)$ such that

$$(1.2) \quad \hat{g}(s) = n^{-1} \sum_{i=1}^n G(s, t_i) Y_i.$$

The weight function $G(s, t)$ depends on the design points t_1, \dots, t_n and on the

Received June 1983; revised March 1984.

AMS 1980 subject classifications. Primary 62G05, 62J02, 65D10; secondary 46E35.

Key words and phrases. Nonparametric regression, variable kernel, splines, roughness penalty, weight function, adaptive smoothing, Sobolev space, penalized maximum likelihood, curve estimation, density estimation.

smoothing parameter λ , but these dependences will not be expressed explicitly. The main object of this paper is to investigate the form of G in order to establish connections between spline smoothing and kernel (or convolution or moving average) smoothing. These connections give insight into the behaviour of the spline smoother and also show that splines should provide good results *whether or not* the design points are uniformly spaced. For the special case of regularly spaced design points, connections between spline and kernel smoothing have been obtained by Cox (1983) and, under the additional assumption of periodicity, by Cogburn and Davies (1974).

Our study of G will show that, under suitable conditions, the weight function will be approximately of a form corresponding to smoothing by a kernel function κ with bandwidth varying according to the local density f of design points. The kernel κ is given by

$$(1.3) \quad \kappa(u) = \frac{1}{2} \exp(-|u|/\sqrt{2}) \sin(|u|/\sqrt{2} + \pi/4).$$

A graph of κ is given in Figure 1. The effective local bandwidth demonstrated below is $\lambda^{1/4}f(t)^{-1/4}$ asymptotically; thus the smoothing spline's behaviour is intermediate between fixed kernel smoothing (no dependence on f) and smoothing based on an average of a fixed number of neighbouring values (effective local bandwidth proportional to $1/f$). The desirability of this dependence on a low power of f will be discussed in Section 3.

The paper is organized as follows. In Section 2 the main theorem is stated and discussed. In addition, some graphs of actual weight functions are presented and compared with their asymptotic forms. These show that the kernel approximation of the weight function is excellent in practice. Section 3 contains some discussion

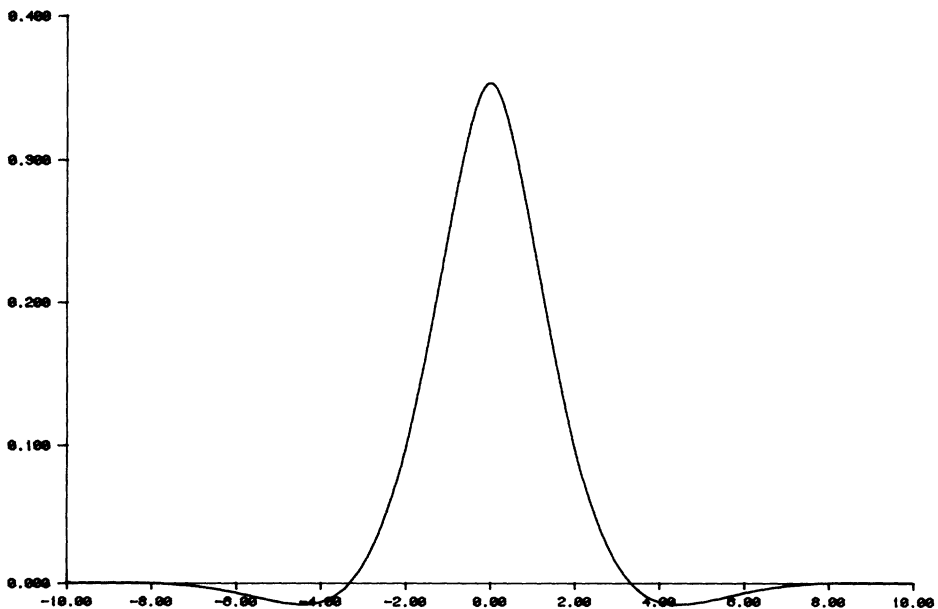


FIG. 1. The effective kernel κ .

of variable bandwidth kernel smoothing methods and their relation to spline smoothers. The proof of the main theorem is given in Section 4. The approximation of $G(s, t)$ by a kernel function deteriorates somewhat near the boundary of the design set, and a modification to correct for boundary effects is developed in Section 5. The case where the data points are weighted is considered in Section 6; the main theory of the paper easily generalizes to this case. Also in Section 6, the results are applied to give an approximation to the diagonal values of the influence matrix (or hat matrix) for spline regression. Finally, in Section 7 we discuss applications to the related subject of penalized maximum likelihood probability density estimation, and demonstrate heuristically that penalized maximum likelihood density estimators should adapt well in the tails of the distribution.

It is the hope that this paper will achieve two main objects. First, the connections demonstrated should help statisticians gain intuition about the way that splines smooth data. Although the implicit definition of the smoothing spline as the solution of a minimization problem is appealing, there are occasions when an explicitly defined estimator is easier to understand fully. Second, the discussion shows that the spline smoother is an excellent choice for nonuniform design points. It provides a single stage procedure which adapts to the nonuniformity in a highly desirable way in balancing the amount of smoothing applied to different parts of the sample.

2. The effective weight function: assumptions and results. It is convenient first to establish some notation and to state the assumptions under which the main result will be proved. Suppose throughout that (a, b) is a fixed finite real interval. Let F_n be a sequence of probability distribution functions with $F_n(a) = 0$ and $F_n(b) = 1$, for all n . In the application to the spline smoothing problem, the function F_n will be the empirical distribution function of the design points, given by

$$(2.1) \quad F_n(s) = n^{-1} \times (\text{number of } t_1, \dots, t_n \leq s).$$

However it is convenient (see Sections 6 and 7) not to restrict attention to this case.

Let $H^2[a, b]$ be the space of functions g on $[a, b]$ for which g and g' are absolutely continuous and $\int_a^b g''(s)^2 ds < \infty$. For t in (a, b) , define a functional $A_t(g)$ by

$$(2.2) \quad A_t(g) = \frac{1}{2} \lambda \int_a^b g''(u)^2 du + \frac{1}{2} \int_a^b g(u)^2 dF_n(u) - g(t).$$

It can be shown that A_t has a unique minimizer in $H^2[a, b]$ provided the support of the probability measure defined by F_n has more than one point. (See Lemma 3 below for a proof of a slightly weaker existence result). Let g_t be this minimizer, and set

$$(2.3) \quad G(s, t) = g_t(s) \quad \text{for all } s \text{ and } t \text{ in } [a, b].$$

To see that the definition (2.3) accords with the implicit definition of G given

in (1.2) above, suppose that all the design points lie in (a, b) and that F_n is defined as in (2.1). Then, by an easy calculation

$$(2.4) \quad 2A_i(g) = \lambda \int g''^2 + n^{-1} \sum_{j \neq i} g(t_j)^2 + n^{-1}(g(t_i) - n)^2 - n.$$

By definition (2.3), the minimizer of (2.4) is $G(\cdot, t_i)$. However (2.4) is, up to a constant, the functional in (1.1) with $Y_i = n$, $Y_j = 0$ otherwise. Substituting into (1.2) shows that G as defined by (2.3) is indeed the required weight function. (The minimizer of (1.1) is the same whatever range of integration is taken in (1.1), provided this range is an interval containing all the design points; this is an easy consequence of the boundary conditions (see Reinsch, 1967) satisfied by smoothing splines.)

The main assumptions of this paper are the following:

(2.5) There exists an absolutely continuous distribution function F on $[a, b]$ such that $F_n \rightarrow F$ uniformly as $n \rightarrow \infty$.

(2.6) Setting $f = F'$,

$$0 < \inf_{[a,b]} f \leq \sup_{[a,b]} f < \infty.$$

(2.7) The density f has bounded first derivative on $[a, b]$.

(2.8) Setting $\alpha(n) = \sup_{[a,b]} |F_n - F|$, the smoothing parameter λ depends on n in such a way that $\lambda \rightarrow 0$ and $\lambda^{-1/4} \alpha(n) \rightarrow 0$ as $n \rightarrow \infty$.

The last assumption in (2.8) ensures that the smoothing parameter does not tend to zero too rapidly. In order to explain and illuminate the assumptions in the spline smoothing context we consider two cases of obvious interest, though the applicability of the results is by no means restricted to these:

(i) Design points “regularly distributed with density f ”; i.e.

$$t_i = F^{-1}((i - 1/2)/n).$$

In this case $\sup |F_n - F| = 1/2n^{-1}$, so that we require $n^4 \lambda \rightarrow \infty$ and $\lambda \rightarrow 0$ for (2.8) to hold.

(ii) Design points randomly and independently distributed with density f . By Serfling (1980) Theorem 2.1.4b for example, we have $\sup |F_n - F| = O(n^{-1/2}(\log \log n)^{1/2})$ almost surely. Thus the assumptions will hold with probability 1 if $\lambda \rightarrow 0$ and $n^{2-\epsilon} \lambda \rightarrow \infty$ for some $\epsilon > 0$.

The main theorem of the paper can now be stated. The idea is to blow up the scale of the s -axis near t and to renormalize to keep the integral of $G(\cdot, t)$ constant. Doing this in such a way as to obtain a nondegenerate limit displays the asymptotic form of the weight function.


THEOREM A. Choose any fixed t such that $a < t < b$. Under assumptions (2.5)–(2.8), defining κ as in (1.3) above,

$$(2.9) \quad \lambda^{1/4} f(t)^{-1/4} G(t + \lambda^{1/4} f(t)^{-1/4} x, t) \rightarrow \kappa(x)/f(t)$$

as $n \rightarrow \infty$, uniformly over all x for which $t + \lambda^{1/4} f(t)^{-1/4} x$ lies in $[a, b]$.

The uniformity over t of the convergence in (2.9) is discussed in Section 4 below. Together with the remarks made there, Theorem A says that for large n and small λ , and t_i not too close to the boundary, **the weight function corresponding to the observation at t_i looks approximately like a kernel function κ centered at t_i with bandwidth $\lambda^{1/4}f(t_i)^{-1/4}$** ; the explicit approximation is

(2.10)
$$G(s, t) \doteq \frac{1}{f(t)} \frac{1}{h(t)} \kappa\left(\frac{s-t}{h(t)}\right) \quad \text{with} \quad h(t) = \lambda^{1/4}f(t)^{-1/4}.$$

 最终

To obtain (2.10), set $s = t + xh(t)$ in (2.9). The reason for the quotient $f(t)$ in (2.10) is discussed in Section 3 below. The proof of Theorem A will be given in Section 4.

In order to illustrate how well the approximation (2.10) works out in practice, some explicit calculations were carried out. The density f was taken to be the normal density with mean $1/2$ and standard deviation $1/4$, truncated at ± 1.96 standard deviations away from the mean. This gives a highly nonuniform density. A hundred design points t_i were placed regularly with density f . The value $\lambda = 10^{-7}$ was used for the smoothing parameter since this was found in other studies (to be reported elsewhere) to be a useful value for spline smoothing with these design points. The weight functions $G(s, t_i)$ for various values of i are shown in Figure 2 together with the approximation (2.10) in terms of κ . (Outside the range displayed, the curves are effectively zero within the range of the design points). It can be seen that the closeness of the approximation is remarkable for points t_i away from the boundary. Even for $i = 99$ ($t_i = 0.94$) the approximation

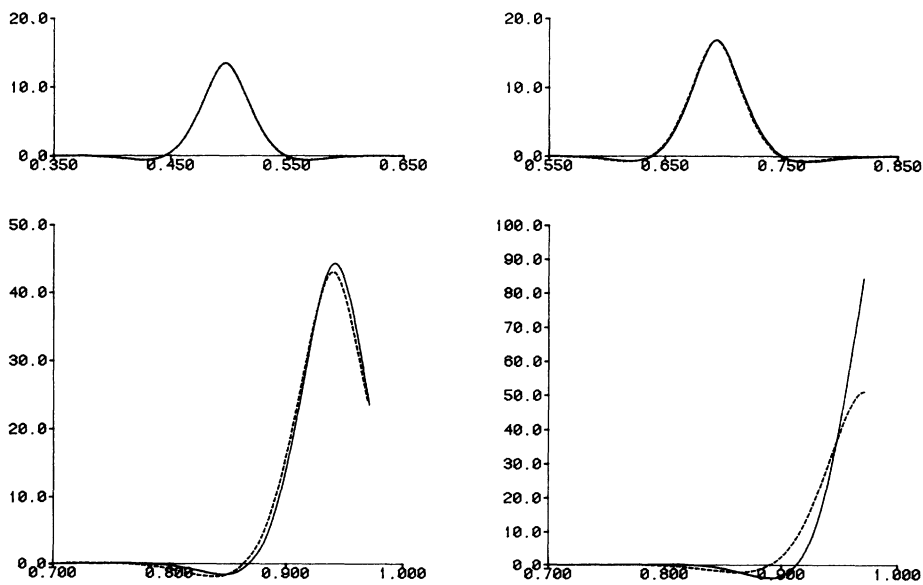


FIG. 2. Comparison between the exact weight function (solid curves) and its asymptotic form (dashed curves), 100 design points regularly spaced with truncated $N(1/2, 1/16)$ density, $\lambda = 10^{-1}$. Top left: $i = 50$, $t_i = 0.497$; top right: $i = 80$, $t_i = 0.693$; bottom left: $i = 99$, $t_i = 0.940$; bottom right: $i = 100$, $t_i = 0.971$. (Note that in the top two figures the curves almost coincide).

is excellent. The approximation is not so good for the very last design point of all, $i = 100$, but even then the discrepancy is only serious in the part of the range near the boundary; this is precisely where the boundary difficulties discussed by Rice and Rosenblatt (1983) will occur. In addition the approximation may be breaking down because the local bandwidth is too small relative to the local spacing between the design points, and the relative value of f is changing rapidly in this region. Further remarks about boundary effects are made in Section 5 below.

3. How would an ideal variable kernel smoother behave? Consider the estimation of the curve g using a kernel estimate with kernel function K . This problem has been considered by Priestley and Chao (1972) and many other authors. Concentrate attention on the case where the t_i are regularly distributed with density f , i.e. $F(t_i) = (i - 1/2)/n$. The Priestley-Chao estimate of $g(s)$ is then essentially of the form

$$(3.1) \quad g_n(s) = \sum \frac{Y_i h^{-1} K\{(s - t_i)/h\}}{nf(t_i)}.$$

In the original formulation, the factor $\{nf(t_i)\}^{-1}$ is replaced by $t_i - t_{i-1}$, but the difference between the two is of negligible magnitude and it is clearer and more convenient to assume that $f(t_i)$ is known. The weight function associated with the Priestley-Chao estimator is thus

$$w_{PC}(s, t_i) = f(t_i)^{-1} h^{-1} K((s - t_i)/h)$$

in the sense that the estimator is then given by

$$g_n(s) = n^{-1} \sum Y_i w_{PC}(s, t_i).$$

Comparison of this weight function with (2.10) demonstrates the sense in which spline smoothing corresponds asymptotically with kernel smoothing with local bandwidth $\lambda^{1/4} f(t)^{-1/4}$ and kernel κ , and accounts for the quotient $f(t)$ on the right hand side in (2.10).

Suppose that the density f and the curve g have bounded derivatives up to second order on $[a, b]$, and that f is bounded below away from zero on $[a, b]$. Suppose that the kernel K has two bounded continuous derivatives, and that the conditions of Section 5 of Benedetti (1977) are satisfied with $r = 2$. Assume that the ϵ_i are uncorrected with mean 0 and variance σ^2 . Assume also that $h \rightarrow 0$ and $n^2 h^5 \rightarrow \infty$ as $n \rightarrow \infty$. Then, by routine analysis it may be shown that

$$Eg_n(t) - g(t) = \frac{1}{2} h^2 g''(t) \int u^2 K(u) du + o(h^2) + O(n^{-2} h^{-3})$$

and

$$\text{var } g_n(t) = \frac{\sigma^2}{nhf(t)} \left\{ \int K^2(u) du + o(1) + O(n^{-2} h^{-3}) \right\}.$$

The usual manipulations (see Benedetti, 1977, and Parzen, 1962, Lemma 4a)

give the optimum value of h (in the sense of minimizing mean square error at t) as

$$(3.2) \quad h_{\text{opt}}(t) = n^{-1/5} \sigma^{2/5} A(K) |g''(t)|^{-2/5} f(t)^{-1/5}$$

where $A(K)$ is a functional of the kernel only.

Now suppose that we wanted to choose h optimally for each value of t . The values of g'' are unknown and difficult to estimate. In the absence of knowledge about g'' , equation (3.2) demonstrates that the local bandwidth should be chosen proportional to $f(t)^{-1/5}$. In other words, speaking qualitatively, the bandwidth should be smaller where the design points are more thickly spread, but the variability in bandwidth should be much slower than that in local density. It should perhaps be pointed out that in (3.2) the bandwidth depends on the point at which g is being estimated while the variable bandwidth Priestley-Chao estimator will have bandwidth depending on the design points. Asymptotically, there is no difference since only design points near t will contribute to the estimate at t .

The effective local bandwidth of spline smoothing is shown in Theorem A above to be proportional to $f(t)^{-1/4}$. This is not quite the same as the rate given in the last paragraph, but the difference in practical terms between one quarter and one fifth power dependence is so slight as to be of little importance.

It must be noted that the kernel κ , unfortunately, does not itself satisfy the conditions of the above discussion, since it has

$$(3.3) \quad \int u^2 \kappa(u) du = 0 \quad \text{and} \quad \int u^4 \kappa(u) du = -1.$$

For such a kernel, the optimum asymptotic dependence of the bandwidth on f would in fact be proportional to $f(t)^{-1/9}$. However the sample sizes required for these higher order asymptotics actually to operate are astronomical; see the remarks of Bartlett (1963). We should conclude that the effective local bandwidth of the spline smoother varies in approximately the same way as that of the ideal variable kernel estimator with nonnegative kernel. The spirit in which this discussion should be taken is summarized by Rosenblatt (1971) page 1818. "The arguments . . . have been of an asymptotic character and it is a mistake to take them too literally from a finite sample point of view. But even asymptotic arguments if used and interpreted with care can yield meaningful ideas."

4. Proof and discussion of the main theorem. The main part of this section is concerned with the proof of Theorem A, using ideas from fairly elementary functional analysis. Readers who are prepared to take the proof on trust should jump to the remarks at the end of this section, which they may find of some interest.

In the proof, it will be assumed without loss of generality that $t = 0$ and hence that $a < 0 < b$. We first define some additional notation. Write

$$\begin{aligned} a_\lambda &= \lambda^{-1/4} f(0)^{1/4} a, & b_\lambda &= \lambda^{-1/4} f(0)^{1/4} b \\ c_\lambda &= \lambda^{-1/4} f(0)^{1/4} \quad \text{and} \quad c'_\lambda &= c_\lambda / \sqrt{2}. \end{aligned}$$

The notation \int_λ and $\sup_{(\lambda)}$ will denote integral and supremum over the interval $[a_\lambda, b_\lambda]$. The space of functions g which have g and g' absolutely continuous on $[a_\lambda, b_\lambda]$ and for which $\int_\lambda g''^2 < \infty$ will be called H_λ .

We shall define, for x in $[a_\lambda, b_\lambda]$

$$F_{n,\lambda}(x) = c_\lambda F_n(c_\lambda^{-1}x)/f(0), \quad F_\lambda(x) = c_\lambda F(c_\lambda^{-1}x)/f(0)$$

and

$$f_\lambda(x) = f(c_\lambda^{-1}x)/f(0).$$

Define norms $\|g\|_\lambda$ and $\|g\|_{n,\lambda}$ on H_λ by $\|g\|_\lambda^2 = \int_\lambda (g''^2 + g'^2 + g^2)$ and

$$\|g\|_{n,\lambda}^2 = \int_\lambda g''^2 + \int_\lambda g^2 dF_{n,\lambda}.$$

(Except where explicitly stated otherwise, integrals are taken with respect to Lebesgue measure.) The norm $\|g\|_\lambda$ is well known as the *Sobolev norm* on H_λ ; see Adams (1975) for a treatment of such norms.

Define, for g in H_λ ,

$$A(g) = \frac{1}{2} \|g\|_{n,\lambda}^2 - g(0)$$

and let

$$\gamma(x) = c_\lambda^{-1}f(0)G(c_\lambda^{-1}x, 0).$$

Both A and γ depend on n and λ , but this dependence will not be expressed explicitly. Note that the definition of γ makes it possible to rewrite the result (2.9) in the much simpler form

$$\gamma(x) \rightarrow \kappa(x) \quad \text{as } n \rightarrow \infty.$$

In the proofs of the various lemmas below, c_1, c_2, \dots denote strictly positive constants which may depend on f . The proof of Theorem A now proceeds in several stages. The first lemma uses properties of Sobolev norms to give important properties of $\|g\|_{n,\lambda}$.

LEMMA 1. *Under assumptions 1 and 2 above, there exist positive constants $c_1(f)$ and $c_2(f)$ such that, for all sufficiently large n , and all g in H_λ ,*

$$(4.1) \quad \begin{aligned} c_1^{-1} \|g\|_\lambda &\leq \|g\|_{n,\lambda} \leq c_1 \|g\|_\lambda \\ \sup_{(\lambda)} |g| &\leq c_2 \|g\|_{n,\lambda}, \quad \sup_{(\lambda)} |g'| \leq c_2 \|g\|_{n,\lambda}. \end{aligned}$$

PROOF. Once (4.1) has been proved, the last two statements follows from the Sobolev embedding theorem (Theorem 5.4C of Adams, 1975). The fact that the same constant c_2 works for all sufficiently large n follows from the remarks about the universality of the embedding constant made at the top of page 97 of Adams (1975). (The property $\lambda \rightarrow 0$ ensures that in Adams's notation, the same cone C can be used for all the intervals (a_λ, b_λ) .)

Define

$$\|g\|_{\lambda}^{*2} = \int_{\lambda} g'^2 + \int_{\lambda} g^2 dF_{\lambda} \quad \text{for all } g \text{ in } H_{\lambda}.$$

Then

$$\begin{aligned} & | \|g\|_{n,\lambda}^2 - \|g\|_{\lambda}^{*2} | \\ &= \left| \int_{\lambda} g^2 d(F_{n,\lambda} - F_{\lambda}) \right| \\ &\leq [g^2(F_{n,\lambda} - F_{\lambda})]_{a_{\lambda}}^{b_{\lambda}} + 2 \int_{\lambda} |gg'| \sup_{(\lambda)} |F_{n,\lambda} - F_{\lambda}| \\ (4.2) \quad &\leq \sup |F_{n,\lambda} - F_{\lambda}| \cdot 2 \left\{ \sup_{(\lambda)} g^2 + \left(\int_{\lambda} g^2 \right)^{1/2} \left(\int_{\lambda} g'^2 \right)^{1/2} \right\} \\ &\leq c_3 \|g\|_{\lambda}^2 \sup |F_{n,\lambda} - F_{\lambda}| \end{aligned}$$

for all n , by using the Sobolev embedding theorem. The fact that $\lambda \rightarrow 0$, so that $|b_{\lambda} - a_{\lambda}|$ is bounded away from zero, ensures that the constant c_3 can be chosen independently of n . Now notice that, by straightforward calculus, making use of Assumption (2.6) above and Lemma 4.10 of Adams (1975),

$$(4.3) \quad c_4 \|g\|_{\lambda}^2 \leq \|g\|_{\lambda}^{*2} \leq c_5 \|g\|_{\lambda}^2 \quad \text{for all } \lambda \leq \lambda_{\max}.$$

Since, by assumption (2.8) above

$$\sup |F_{n,\lambda} - F_{\lambda}| = f(0)^{-3/4} \lambda^{-1/4} \sup |F_n - F| \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

for sufficiently large n , we will have

$$c_3 \sup |F_{n,\lambda} - F_{\lambda}| \leq \min^{1/2}(c_4, c_5)$$

and hence, combining (4.2) and (4.3) will ensure that

$$^{1/2} c_4 \|g\|_{\lambda}^2 \leq \|g\|_{n,\lambda}^2 \leq 2c_5 \|g\|_{\lambda}^2$$

completing the proof of Lemma 1. \square .

The next lemma gives a technical result which will be required later.

LEMMA 2. *Given any u in H_{λ} ,*

$$\int_{\lambda} u''\kappa'' = [u'\kappa'' - u\kappa''']_{a_{\lambda}}^{b_{\lambda}} - \int_{\lambda} u\kappa + u(0).$$

PROOF. It is easily verified by fairly tedious calculus that

$$\kappa^{iv} = \delta - \kappa$$

where δ is the Dirac delta function. Integrating by parts twice and substituting this result completes the proof of the Lemma. \square

Notice that, again by easy but long-winded calculus we have, for all x ,

$$(4.4) \quad \kappa^{(i)}(x) \leq \frac{1}{2} \exp(-|x|/\sqrt{2}) \quad \text{for } i = 2, 3$$

and hence, for all u in H_λ ,

$$(4.5) \quad |[u' \kappa'' - u \kappa''']_{a_\lambda}^{b_\lambda}| \leq (\sup_\lambda |u| + \sup_\lambda |u'|) \exp(-c'_\lambda \min(|a|, |b|)).$$

Combining (4.5) with Lemma 1 then gives the following corollary.

COROLLARY. *Under the assumptions of Lemma 1, there exists a constant c_6 such that, for all sufficiently large n ,*

$$(4.6) \quad |[u' \kappa'' - u \kappa''']_{a_\lambda}^{b_\lambda}| \leq c_6 \|u\|_{n,\lambda} \exp\{-c'_\lambda \min(|a|, |b|)\}.$$

The final lemma is a consequence of the connection between the functionals A and A_0 .

LEMMA 3. *For all sufficiently large n , A has a unique minimizer over H_λ and this minimizer is γ .*

PROOF. The existence and uniqueness of the minimizer of A follows from Lemma 1 by applying Theorem 1.7 of Tapia and Thompson (1978). It is easy to see that, letting $Tg(x) = c_\lambda^{-1} f(0) g(c_\lambda^{-1} x)$ for all g in $H^2[a, b]$, $c_\lambda^{-1} f(0) A_0(g) = A(Tg)$. Hence A_0 will indeed have a unique minimizer. By the definition (2.3), this minimizer is $G(\cdot, 0)$; thus the minimizer of A will be $TG(\cdot, 0)$, which is equal to γ as required. \square

We can now state and prove the key result of this section.

PROPOSITION 1. *Define γ and κ as above. Under assumptions (2.5)–(2.8), there exists a constant $c_9(f)$ such that, for all sufficiently large n ,*

$$\|\gamma - \kappa\|_{n,\lambda} \leq c_9 [\lambda^{1/4} + \lambda^{-1/4} \alpha(n) + \exp\{-c'_\lambda \min(|a|, |b|)\}].$$

PROOF. The functional calculus used in this proof is nicely explained in Appendix 1 of Tapia and Thompson (1978). Since γ is the minimizer of $A(g)$ over g in H_λ , and since the functional derivative A'' is constant over H_λ , it follows that, for all u in H_λ

$$A'(\kappa)(u) = \{A'(\kappa) - A'(\gamma)\}(u) = A''(\gamma)(\kappa - \gamma, u).$$

Setting $u = \kappa - \gamma$ gives

$$A'(\kappa)(\kappa - \gamma) = \|\kappa - \gamma\|_{n,\lambda}^2$$

since (see Example 6, page 155 of Tapia and Thompson, 1978),

$$A''(\gamma)(u, u) = \|u\|_{n,\lambda}^2 \quad \text{for all } u \text{ in } H_\lambda.$$

Thus, if it can be shown that, for all u in H_λ

$$(4.7) \quad |A'(\kappa)(u)| \leq \varepsilon(n) \|u\|_{n,\lambda}$$

for some $\varepsilon(n)$, it will follow that

$$(4.8) \quad \varepsilon(n) \|\kappa - \gamma\|_{n,\lambda} \geq \|\kappa - \gamma\|_{n,\lambda}^2$$

and hence that

$$(4.9) \quad \|\kappa - \gamma\|_{n,\lambda} \leq \varepsilon(n).$$

This is the essence of the remainder of the proof. For any u in H_λ , applying Lemma 2 gives

$$(4.10) \quad \begin{aligned} A'(\kappa)(u) &= \int_\lambda \kappa'' u'' + \int_\lambda \kappa u \, dF_{n,\lambda} - u(0) \\ &= [u' \kappa'' - u \kappa''']_{a_\lambda}^{b_\lambda} + \int_\lambda \kappa u \, dF_{n,\lambda} - \int_\lambda \kappa u. \end{aligned}$$

The first part of this expression is dealt with in (4.6) above. To cope with the second part, set, for all t ,

$$(4.11) \quad \begin{aligned} G_{n,\lambda}(t) &= \int_0^t \kappa(x)(dF_{n,\lambda}(x) - dx) \\ &= \int_0^t \kappa(x)d[F_{n,\lambda}(x) - F_\lambda(x)] + \int_0^t \kappa(x)\{f_\lambda(x) - 1\} \, dx. \end{aligned}$$

Now

$$(4.12) \quad \begin{aligned} &\left| \int_0^t \kappa(x) \, d(F_{n,\lambda}(x) - F_\lambda(x)) \right| \\ &\leq \left| \int_0^t (F_{n,\lambda}(x) - F_\lambda(x))\kappa'(x) \, dx \right| + [\kappa(F_{n,\lambda} - F_\lambda)]_0^t \\ &\leq \left(\int_\infty^\infty |\kappa'| + 2 \sup |\kappa| \right) \sup |F_{n,\lambda} - F_\lambda| \\ &= c_7 \lambda^{-1/4} \sup |F_n - F| = c_7 \lambda^{-1/4} \alpha(n). \end{aligned}$$

On the other hand, the smoothness properties assumed for f ensure that Taylor's theorem can be used to write

$$(4.13) \quad \begin{aligned} \int_0^t |\kappa(x)(f_\lambda(x) - 1)| \, dx &\leq \int_0^t |x\kappa(x)| \, dx \sup |f'_\lambda| \\ &\leq \int_{-\infty}^\infty |x\kappa(x)| \, dx \cdot \lambda^{1/4} f(0)^{-5/4} \sup |f'|. \end{aligned}$$

Substituting the bounds (4.12) and (4.13) into (4.11) gives

$$(4.14) \quad \sup_{(\lambda)} |G_{n,\lambda}| \leq c_8 \{\lambda^{1/4} + \gamma^{-1/4} \alpha(n)\}.$$

Now, for all t in (a_λ, b_λ)

$$(4.15) \quad \left| \int_0^t \kappa u \, dF_{n,\lambda} - \int_0^t \kappa u \right| = \left| \int_0^t u \, dG_{n,\lambda} \right| \leq |G_{n,\lambda} u|_0^t + \left| \int_0^t u' G_{n,\lambda} \right| \\ \leq 3c_2 \|u\|_{n,\lambda} \sup G_{n,\lambda} \quad \text{for sufficiently large } n,$$

by Lemma 1. Substitute (4.14) into (4.15) and combine the result with (4.6) to give, as a consequence of (4.10),

$$(4.16) \quad |A'(\kappa)(u)| \leq c_9 \|u\|_{n,\lambda} [\lambda^{1/4} + \lambda^{-1/4} \alpha(n) + \exp\{-c'_\lambda \min(|a|, |b|)\}]$$

for all sufficiently large n . Inequality (4.16) is precisely (4.7) with $\varepsilon(n)$ equal to $c_9 \times$ the quantity in []. The proof of the proposition follows from (4.9) at once. \square

We can now derive, very easily, a proposition from which Theorem A follows at once. Drop the restriction $t = 0$ and define, for each t in (a, b) and $t + \lambda^{1/4}f(t)^{-1/4}x$ in $[a, b]$, $\gamma_t(x) = \lambda^{1/4}f(t)^{3/4}G(t + \lambda^{1/4}f(t)^{-1/4}x, t)$. Combining Lemma 1 and Proposition 1 then gives the following result.

PROPOSITION 2. *Under assumptions (2.5)–(2.8) the supremum of both $|\gamma_t(x) - \kappa(x)|$ and $|\gamma'_t(x) - \kappa'(x)|$ over*

$$\{x: a \leq t + \lambda^{1/4}f(t)^{-1/4}x \leq b\}$$

are bounded by a constant multiple of

$$(4.17) \quad \lambda^{1/4} + \lambda^{-1/4} \sup |F_n - F| + \exp\{-\lambda^{-1/4}f(t)^{1/4}2^{-1/2}\min(t - a, b - t)\}.$$

The constant depends only on f .

REMARKS.

1. Theorem A is an immediate consequence, because the quantity (4.17) converges to zero for each t in (a, b) . In addition the theorem remains true if both sides of (2.9) are differentiated with respect to x .

2. The form of the exponential term in (4.17) ties in closely with the discussion of boundary effects given by Rice and Rosenblatt (1983). The distortion of the effective smoothing kernel near the boundaries a and b dies away at an exponential rate in just the same way as the bias effect discussed by Rice and Rosenblatt. Furthermore we can use the bound (4.17) to show that the convergence in Theorem A is uniform over all t in $(a + \varepsilon_n, b - \varepsilon_n)$, where ε_n is any sequence for which

$$\lambda^{-1/4}\varepsilon_n \rightarrow \infty \quad \text{as } n \rightarrow \infty.$$

3. The assumption (2.7) that f has bounded first derivative can in fact be weakened somewhat. This assumption is used to get the bound (4.13).

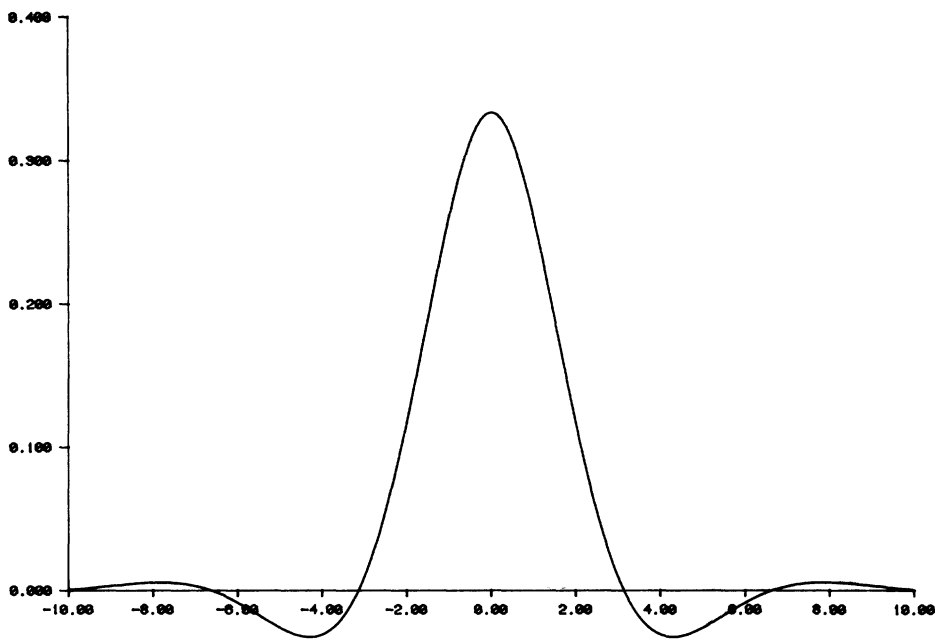


FIG. 3. *The effective kernel for a roughness penalty involving third derivatives.*

If the assumption (2.7) is replaced by a Lipschitz condition of the form $\sup |f(x) - f(y)| = O(|x - y|^\epsilon)$ then a bound like (4.13) can easily be obtained with the power $1/4$ replaced by $1/4 \epsilon$. This will have corresponding repercussions on the $\lambda^{1/4}$ term in Propositions 1 and 2, but the basic conclusions will remain the same.

4. In order to make the argument reasonably clear, and because of the interest in cubic spline smoothing, we have concentrated attention on the roughness penalty $\int g''^2$. Similar techniques can be used to deal with other roughness penalties. If the roughness penalty is of the form

(4.18)
$$\int (g^{(r)} + \text{terms in lower derivatives of } g)^2$$

then the $1/4$ powers will be replaced by $1/2r$ powers throughout the argument. The effective asymptotic kernel will be given by the solution vanishing at $\pm\infty$ of

(4.19)
$$(-1)^r \kappa^{(2r)} + \kappa = \delta$$

where δ is the Dirac delta function. It follows from (4.19) that the kernel will have Fourier transform $(1 + s^{2r})^{-1}$. For $r = 1$ the asymptotic effective kernel is the Laplace density $1/2 \exp(-|u|)$ with local bandwidth $\lambda^{1/2} f(t)^{-1/2}$. For $r = 3$ the kernel becomes

(4.20)
$$1/6 \{ \exp(-|u|) + 2 \exp(-|u|/2) \sin(\pi/6 + \sqrt{3}|u|/2) \}$$

with local bandwidth $\lambda^{1/6} f(t)^{-1/6}$. The plot of the kernel (4.20) given in Figure 3 shows it to be very similar in appearance to κ as defined in (1.3).

5. The equivalent kernel near a boundary. One of the uncomfortable features of the development so far is that the approximation of the weight function $G(s, t)$ deteriorates when t is near the boundary of the design set. In this section we fill this gap by exploring how the equivalent kernel is distorted when t is close to the boundary.

Suppose that all the assumptions of Section 2 still hold. Given any point t in $[a, b]$, define

$$\delta = \min(t - a, b - t),$$

the distance from t to the nearest boundary. In order to state what will be the equivalent kernel when δ is small, it is first necessary to define some additional notation. Let $h(t) = \lambda^{1/4} f(t)^{-1/4}$ as before, and define r and α , both of which depend on t and λ , by

$$r \cos \alpha = 1 - 2 \sin(2^{1/2} \delta / h), \quad r \sin \alpha = 1 + 2 \cos(2^{1/2} \delta / h).$$

Define a kernel $\kappa^*(u)$, again depending implicitly on t and λ , by

$$\kappa^*(u) = -2^{-3/2} r \exp(-|u|/\sqrt{2}) \sin(|u|/\sqrt{2} - \alpha).$$

The definition of κ^* is similar to that of κ given in (1.3) except for a change of amplitude and phase in the trigonometric term.

Let t^* be the reflection of t in the nearest boundary, that is

$$t^* = \begin{cases} t - 2\delta & \text{if } t < \frac{1}{2}(a + b) \\ t + 2\delta & \text{if } t > \frac{1}{2}(a + b). \end{cases}$$

It can then be shown, in a sense to be made clear below, that, for s and t in $[a, b]$,

$$(5.1) \quad G(s, t) \doteq \frac{1}{f(t)} \frac{1}{h(t)} \left\{ \kappa\left(\frac{s-t}{h(t)}\right) + \kappa^*\left(\frac{s-t^*}{h(t)}\right) \right\},$$

and that the approximation is uniformly good for all s and t .

John Rice (personal communication) has pointed out that, if the design points are regularly spaced, plots of the weight function G give a local bandwidth appearing to get smaller near the boundary. Since $h(t)$ is constant for all t , this at first sight does not seem to accord with (5.1). A similar phenomenon appears in Figure 2, where the exact weight function appears to have narrower bandwidth than that predicted by the approximation (2.10) when t is near the boundary.

Some intuition may be gained by considering the limiting case where t is actually on the boundary. Suppose that the boundary is at 0; in the limiting case t and t^* will coincide. Since $\delta = 0$ we have $r \cos \alpha = 1$ and $r \sin \alpha = 3$. After some calculation this yields, for $x \geq 0$,

$$\kappa(x) + \kappa^*(x) = 2^{1/2} \exp(-x/\sqrt{2}) \cos(x/\sqrt{2}).$$

Although $\kappa + \kappa^*$ will die away at the same exponential rate as the uncorrected kernel κ , there are two respects in which it appears to have narrower bandwidth. First, it is the case that $\kappa(0) + \kappa^*(0) = 4\kappa(0)$, so the central value is considerably larger. Second, the first zero-crossing of $\kappa + \kappa^*$ is at $\frac{1}{2}\pi\sqrt{2}$ while that of κ is at

$\frac{3}{4}\pi\sqrt{2}$, so that the distance from t to the first zero-crossing of the weight function is multiplied by $\frac{2}{3}$ near the boundary.

We close this section with a theorem, corresponding to Theorem A, which justifies the approximation (5.1).

THEOREM B. *For each t in $[a, b]$ and for $\lambda > 0$, define*

$$\kappa_c(x) = \kappa(x) + \kappa^*\{x + (t - t^*)/h(t)\}$$

where κ , κ^ , t^* and $h(t)$ are as defined above. Then, provided the assumptions of Section 2 are satisfied, for all sufficiently large n ,*

$$|h(t)G\{t + xh(t), t\} - \kappa_c(x)/f(t)| = O(\lambda^{1/4} + \lambda^{-1/4}\sup |F_n - F|)$$

uniformly over all t in $[a, b]$ and all x for which $t + xh(t)$ lies in $[a, b]$.

REMARK. Substituting $s = t + xh(t)$ yields the approximation (5.1) for $G(s, t)$ given above. The difference between Theorem B and Theorem A is that in Theorem B the approximation holds uniformly over all t as well as over x , thus eliminating boundary problems at the cost of introducing an equivalent kernel κ_c whose form depends on t and λ .

PROOF. The proof closely parallels that of Proposition 2, replacing κ by κ_c throughout. Using the same notation as in Section 4, assume without loss of generality that $t = 0$. The key property of κ_c , which may be checked by rather tedious calculus, is that its second and third derivatives both vanish at the nearer of a_λ and b_λ to zero. This makes it possible to replace $\min(|a|, |b|)$ by $\max(|a|, |b|)$, and hence by $\frac{1}{2}(b - a)$, in the inequalities corresponding to (4.5) and (4.6). The exponential term $\exp\{-\frac{1}{2}c_\lambda(b - a)\}$ is then uniformly dominated, for all sufficiently large n , by $\lambda^{1/4}$, and hence the exponential term can be omitted from the result corresponding to Proposition 1. Theorem B then follows by exactly the argument leading to Proposition 2.

6. Weighted spline smoothing and the hat matrix. The results already obtained can be applied very easily to give the equivalent kernel for weighted spline smoothing, where the objective function (1.1) is replaced by

$$\lambda \int g''(t)^2 dt + \sum_{i=1}^n w_i \{Y_i - g(t_i)\}^2;$$

here w_1, \dots, w_n are positive weights. Note that all the weights will depend on n , but this dependence is not expressed explicitly. Let $W = w_1 + \dots + w_n$. To simplify the notation we shall assume that $W = 1$; if it is not, λ should be replaced by λ/W and w_i by w_i/W for the remainder of this section.

Now replace the distribution function F_n of (2.1) by the weighted version

$$F_n^W(s) = \sum w_i I[t_i \leq s].$$

Suppose F_n^W converges uniformly to a continuous distribution function F as in Section 2.

It is now the case that the smoothing spline is of the form

$$(6.1) \quad \hat{g}(s) = \sum G(s, t_i) w_i Y_i$$

where $G(\cdot, t_i)$ is the minimizer of

$$(6.2) \quad A_t^W(g) = \frac{1}{2} \lambda \int g''(u)^2 du + \frac{1}{2} \int g(u)^2 dF_n^W(u) - g(t).$$

To justify this assertion, the argument following (2.4) is used, except that Y_i is set to w_i^{-1} . Since A_t^W is of exactly the same form as A_t in (2.2), all the approximation arguments follow through, and the limiting form for $G(s, t)$ is exactly as before. The only difference is that the density f is the limiting density of the appropriately weighted design points.

A natural application of the results of the paper is to find an approximation to the so-called *hat matrix* A which maps the vector Y_j of observed values into the vector $\hat{g}(t_j)$ of predicted values. The hat matrix is of great importance in constructing diagnostic checks for regression (see Cook and Weisberg, 1982). The special case of spline regression is discussed by Silverman (1985); there the hat matrix is used both in diagnostic checks and to give posterior confidence intervals for the curve \hat{g} . It follows at once from (6.1) that the hat matrix will be given by

$$A_{ij} = G(t_i, t_j) w_j;$$

the unweighted spline discussed before this section of course corresponds to the case $w_j = n^{-1}$. Of particular interest are the diagonal entries of the hat matrix. The work of Section 5 can be used to give, setting $\delta_i = \min(t_i - a, b - t_i)$

$$\begin{aligned} A_{ii} &= w_i G(t_i, t_i) \\ &\doteq w_i f(t_i)^{-1} h(t_i)^{-1} [\kappa(0) + \kappa^* \{2\delta_i h(t_i)^{-1}\}] \\ (6.3) \quad &= 2^{-3/2} w_i \lambda^{1/4} f(t_i)^{-3/4} [1 + (2 + \cos u_i - \sin u_i) \exp(-u_i)] \end{aligned}$$

where

$$u_i = 2^{1/2} \lambda^{-1/4} f(t_i)^{1/4} \delta_i.$$

The result of Theorem B shows that this approximation for A_{ii} will be uniformly good for all i , provided the conditions of Section 2 hold. Notice that the factor $[\]$ in (6.3) approaches 4 as $u_i \rightarrow 0$; this accords with the discussion just preceding the statement of Theorem B and demonstrates that observations near the boundary have, as may be expected, a considerably larger local effect.

7. Probability density estimation. Suppose we are estimating a probability density function f_0 on a finite interval (a, b) , given independent observations X_1, \dots, X_n from f . This problem is distinct from, but related to, the nonparametric regression problem considered up to now in this paper. An approach closely related to the spline smoothing regression method is *penalized maximum likelihood*, suggested for density estimation by Good and Gaskins (1971). A general discussion of penalized maximum likelihood estimation is given by

Silverman (1984a). The basic idea is to choose a roughness penalty functional Φ and then to maximize

$$n^{-1} \sum \log f(X_i) + \lambda \Phi(f)$$

subject to f being a probability density function with $\Phi(f) < \infty$. Silverman (1982) considered the case where $\Phi(f)$ is of the form (4.18) with $g = \log f$; this formulation has various advantages discussed in that paper.

Set $g = \log f$ from now on and concentrate attention on the roughness penalty $\int g''^2$ discussed above. (The extension to other penalties of the form (4.18) works in the same way as in the discussion following (4.18).) The estimator \hat{g} of g is then the unconstrained minimizer over v in $H^2[a, b]$ of

$$(7.1) \quad \frac{1}{2} \lambda \int v''^2 + \int e^v - n^{-1} \sum v(X_i).$$

An approximation g_1 to \hat{g} is obtained by minimizing

$$(7.2) \quad \frac{1}{2} \lambda \int v''^2 + \frac{1}{2} \int v^2 f - \int v g f + \int v f - n^{-1} \sum v(X_i)$$

over v in $H^2[a, b]$; for details of the derivation of (7.1) and (7.2) and the accuracy of the approximation see Silverman (1982). In the remainder of this section a partly heuristic argument is given to show that $\exp(g_1)$ is approximately a variable bandwidth kernel estimator, and to investigate the way in which the local bandwidth behaves. Assume that f obeys (2.6) and (2.7), and let F be $\int_{-\infty}^x f$. Set $F_n = F$ for all n in (2.2) and define G accordingly in (2.3). Assume that $\lambda \rightarrow 0$ as $n \rightarrow \infty$, and that λ obeys enough conditions (see Silverman, 1982) to ensure that g_1 is a good estimate of g .

The expression (7.2) is of the form of the quadratic part of (2.2) plus a linear term in v . By standard minimization arguments, and the definition of G , it follows that

$$(7.3) \quad g_1(s) = \int G(s, t)g(t)f(t) dt - \int G(s, t)f(t) dt + n^{-1} \sum G(s, X_i).$$

Fix s and let κ_λ be the kernel κ of (1.3) rescaled to have bandwidth $\lambda^{1/4} \cdot f(s)^{-1/4}$. Appealing to Theorem A,

$$(7.4) \quad \begin{aligned} \int G(s, t)f(t)g(t) dt &\doteq \int \kappa_\lambda(s - t)g(t) dt \\ &= g(s) + \int \kappa_\lambda(s - t)\{g(t) - g(s)\} dt \\ &\doteq g(s) + \int \kappa_\lambda(s - t)\{f(t) - f(s)\}f(s)^{-1} dt, \end{aligned}$$

since most of the weight of the integral is near s , making use of a linear expansion of $\log y$ near $y = f(s)$. The expression (7.4) is equal to

$$(7.5) \quad g(s) + f(s)^{-1} \int \kappa_\lambda(s - t)f(t) dt - 1 \doteq g(s) - 1 + \int G(s, t)f(t) dt$$

appealing to Theorem A again. Substitute (7.5) into (7.3) to give

$$(7.6) \quad g_1(s) \doteq g(s) - 1 + n^{-1} \sum G(s, X_i).$$

Now use the consistency results of Silverman (1982) to ensure that $g_1 - g$ is small and that $\hat{g} - g_1$ is relatively negligible. Setting $f_1 = \exp(g_1)$, and $\hat{f} = \exp(\hat{g})$,

$$\begin{aligned} \hat{f}(s) &\doteq f_1(s) = f(s) \exp\{g_1(s) - g(s)\} \\ &\doteq f(s)\{1 + g_1(s) - g(s)\} \\ &\doteq n^{-1} \sum f(s)G(s, X_i) \end{aligned}$$

by substituting (5.6). Appealing to Theorem A again gives

$$(7.7) \quad \hat{f}(s) \doteq n^{-1} \sum \kappa_\lambda(s - X_i).$$

Thus the roughness penalty density estimator $\hat{f} = \exp(\hat{g})$ constructed by minimizing (5.1) is approximately equal to an adaptive kernel estimator with kernel κ and local bandwidth $\lambda^{1/4}f(s)^{-1/4}$. Adaptive kernel estimators of this general kind were discussed by Breiman et al. (1977) and Abramson (1982).

It follows from (4.14) of Parzen (1962) that, if a nonnegative symmetric kernel K is used, then the optimum choice of an adaptive bandwidth would be, for a constant $c(K)$ depending on the kernel,

$$(7.8) \quad h(s) = c(K)n^{-1/5}|f''(s)|^{-2/5}f(s)^{1/5}.$$

This formula depends both on f and on f'' . If we presume that the likely values of f'' are proportional to those of f (i.e. lower values of f'' in the tails) then (7.8) gives the ideal local bandwidth proportional to $f^{-1/5}$. If we use the appropriate formula corresponding to (7.8) for the kernel κ , which satisfies (3.3), then the ideal local bandwidth comes out proportional to $(f/|f^{iv}|^2)^{1/9} \sim f^{-1/9}$ if a similar argument concerning the likely values of f^{iv} is used. Another approach entirely is given by Abramson (1982) who recommends use of a local bandwidth proportional to $f^{-1/2}$. As a compromise between $-1/2$, $-1/5$ and $-1/9$ as the power dependence on f , the $f^{-1/4}$ dependence given asymptotically by the roughness penalty estimate is certainly appealing.

Acknowledgements. The author is most grateful to A. Robinson, J. F. Toland and J. R. Willis for useful discussions; to G. W. Watters for computational assistance; to J. Rice for stimulating correspondence; to the referees for helpful suggestions; and to the British Science and Engineering Council for research facilities.

REFERENCES

- ABRAMSON, I. S. (1982). On bandwidth variation in kernel estimates—a square root law. *Ann. Statist.* **10** 1217–1223.
 ADAMS, R. A. (1975). *Sobolev Spaces*. Academic, New York.
 BARTLETT, M. S. (1963). Statistical estimation of density functions. *Sankhyā A* **25** 245–254.
 BENEDETTI, J. K. (1977). On the nonparametric estimation of regression functions. *J. Roy. Statist. Soc. B* **39** 248–253.

- BREIMAN, L., MEISEL, W. and PURCELL, E. (1977). Variable kernel estimates of probability densities. *Technometrics* **19** 135–144.
- COGBURN, R. and DAVIES, H. T. (1974). Periodic splines and spectral estimation. *Ann. Statist.* **2** 1108–1126.
- COOK, R. D. and WEISBERG, S. (1982). *Residuals and Influence in Regression*. Chapman, New York.
- COX, D. D. (1983). Asymptotics of *M*-type smoothing splines. *Ann. Statist.* **11** 530–551.
- CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.
- DE BOOR, C. (1978). *A Practical Guide to Splines*. Springer, New York.
- GOOD, I. J. and GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277.
- PARZEN, E. (1962). On the estimation of the probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.
- PRIESTLEY, M. B. and CHAO, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. B* **34** 385–392.
- REINSCH, C. (1967). Smoothing by spline functions. *Numer. Math.* **10** 177–183.
- RICE, J. and ROSENBLATT, M. (1983). Smoothing splines: regression, derivatives and deconvolution. *Ann. Statist.* **11** 141–156.
- ROSENBLATT, M. (1971). Curve estimates. *Ann. Math. Statist.* **42** 1815–1842.
- SCHOENBERG, I. J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci. U.S.A.* **52** 947–950.
- SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.
- SILVERMAN, B. W. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810.
- SILVERMAN, B. W. (1984a). Penalized maximum likelihood. In *Encyclopedia of Statistical Sciences* **6**. (S. Kotz and N. L. Johnson, eds.) Wiley, New York.
- SILVERMAN, B. W. (1984b). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Amer. Statist. Assoc.* **79** to appear.
- SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. B* **47** 1, to appear.
- TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation*. Johns Hopkins University Press, Baltimore, Maryland.
- WAHBA, G. (1975). Smoothing noisy data with spline functions. *Numer. Math.* **24** 383–393.
- WHITTAKER, E. (1923). On a new method of graduation. *Proc. Edinburgh Math. Soc.* **41** 63–75.

SCHOOL OF MATHEMATICS
UNIVERSITY OF BATH
BATH BA2 7AY
UNITED KINGDOM