THE INSTITUTE OF STATISTICS

THE UNIVERSITY OF NORTH CAROLINA
CHapel Hill, North Carolina

ON LOCAL SMOOTHING OF NONPARAMETRIC CURVE ESTIMATORS

by

J. Fan	P. Hall
M. Martin	P. Patil

September 1993

Mimeo Series #2309

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

MIMEO J.Fan P.Hall
SERIES M.Martin P.Patil
#2309 ON LOCAL SMOOTHING OF
NONPARAMETRIC CURVE ESTIMATOR

NAME	DATE

Department of Statistics Library

ON LOCAL SMOOTHING OF NONPARAMETRIC CURVE ESTIMATORS

JIANQING FAN¹ PETER HALL²
MICHAEL MARTIN³ PRAKASH PATIL²

ABSTRACT. We begin by analyzing the local adaptation properties of wavelet-based curve estimators. It is argued that while wavelet methods enjoy outstanding adaptability in terms of the manner in which they capture irregular episodes in a curve, they are not nearly as adaptive when considered from the viewpoint of tracking more subtle changes in a smooth function. We point out that while this problem may be remedied by modifying wavelet estimators, simple modifications are typically not sufficient to properly achieve adaptive smoothing of a relatively highly differentiable function. In that case, local changes to the primary level of resolution of the wavelet transform are required. While such an approach is feasible, it is not an attractive proposition on either practical or aesthetic grounds. Motivated by this difficulty, we develop local versions of familiar smoothing methods, such as cross-validation and smoothed cross-validation, in the contexts of density estimation and regression. It is noted that these new methods have the ability to capture irregular episodes in the curve, and that unlike wavelet methods they enjoy a high degree of adaptability to subtle changes in a smooth curve.

KEY WORDS AND PHRASES. Adaptive smoothing, bandwidth, bootstrap, cross-validation, curve estimation, density estimation, kernel methods, local smoothing, nonparametric regression, wavelets.

AMS (1991) SUBJECT CLASSIFICATION. Primary 62G07, Secondary 62G20.

SHORT TITLE. Local smoothing.

¹ Department of Statistics, University of North Carolina at Chapel Hill.

² Centre for Mathematics and its Applications, Australian National University.

³ Department of Statistics, Stanford University.

1. Introduction. Recent interest in wavelet methods for statistical smoothing has sharpened the focus of research on locally adaptive methods for curve estimation. Wavelet transforms are a device for representing functions in a way that is “local in [both] time and frequency” (Strang 1989). In the case of nonparametric curve estimation, “time” represents “location” and “frequency” is a measure of roughness. Thus, wavelet transforms achieve local adaptivity with respect to both location and roughness. This fact has been very elegantly and convincingly established in recent work of Donoho, Johnstone, Kerkyacharyan and Picard, who have demonstrated that wavelet-based nonparametric function estimators adapt remarkably well to irregular episodes in a curve.

However, we argue in the present paper that the very laudable success with which wavelet methods adapt to irregular features in the curve tends to obscure the relative difficulty that they encounter in modelling more subtle changes in a smooth curve. In particular, they do not readily adapt to the different degrees of smoothing that are necessary near a mode, near a point of inflection or in the tails of a very smooth density function. We appreciate that to some readers this point may have an element of controversy associated with it, and we hasten to add that we are not asserting that wavelet methods cannot be modified to achieve greater adaptability. Indeed, in Section 2 we suggest simple modifications to the terms in an empirical wavelet expansion, which achieve a degree of local adaptability to subtle fluctuations in a curve. However, greater modifications are required if true local adaptability is to be attained. Much of the appeal of any curve estimation method is lost if it is complicated by too many modifications. In particular, its intuitive appeal can be obscured to the point where it virtually disappears.

These considerations motivate us to take another look at some standard methods of global, non-locally-adaptive curve estimation, and to develop mild and intuitively appealing modifications that are sensitive to relatively subtle local changes in curves. Our aim is to develop techniques that are as familiar and widely-applicable as the global methods

from which they are derived — for example, general methods for local cross-validation, with or without smoothing, in the context of general curve estimation problems such as density estimation and nonparametric regression. Our local cross-validation approach is entirely different from those which have been considered by earlier authors, such as Hall and Schucany (1989) and Mielniczuk, Sarda and Vieu (1989). Those methods demand that the smoothing parameter be chosen continuously as one moves through the continuum of possible locations. Therefore, a continuum of minimization operations needs to be performed. By way of contrast, the techniques developed here are computationally expedient, requiring only a single minimization, over a finite number of parameters, at the outset of the algorithm. A recently introduced transformation-based method, proposed by Wand, Marron and Ruppert (1991) in the context of density estimation, is a result of somewhat more complex modifications to standard techniques than the method proposed here. Jones's (1990) excellent review of different approaches to local smoothing should also be mentioned here, as should recent work of Brockmann, Gasser and Hermann (1993) and Fan and Gijbels (1993) on local smoothing in the context of nonparametric regression. The latter two papers address locally adaptive, but essentially linear, non-wavelet approaches to curve estimation.

Section 3 introduces our local smoothing methods based on cross-validation. Numerical implementation is described and discussed in Section 4, and proofs are presented in Section 5.

2. Wavelets and adaptive smoothing.

2.1. Introduction and summary. Wavelet methods, introduced to statistics by Donoho (1992), Donoho and Johnstone (1992a-c) and Kerkyacharyan and Picard (1992, 1993a-c), have very important adaptive properties. However, we argue that they generally do not capture the subtle differences in the degrees of smoothing that are appropriate at different locations of a relatively highly differentiable curve.

We motivate our ideas in the context of density estimation, although they have immediate analogues for nonparametric regression. Subsection 2.2 introduces wavelet expansions and wavelet density estimators of classical form, and subsection 2.3 considers in detail the thresholding device that is most commonly used to construct those estimators. In the work there and in subsection 2.4 we assume that the underlying density is smooth, with a number of derivatives equal to the order of the wavelet employed in its construction, and we show in that context that the usual thresholding method used for constructing wavelet estimators is unsuccessful in adapting to subtle features such as different rates of “curvature” in different parts of the curve. Subsection 2.4 presents modifications of classical wavelet-based density estimators that overcome this problem to some extent. These techniques are founded on more accurate estimates of the coefficients of the (mother) wavelet terms in the wavelet expansion of a smooth density, and they involve the use of shrinkage techniques to achieve a level of local adaptability. However, none of our modified wavelet methods provide outstanding locally adaptive smoothing, since they do not address the problem of changing the primary resolution level in a locally adaptive way. The variance of the estimator depends mainly on primary resolution, and unless that quantity is varied with location, the familiar trade-off between variance and squared bias in an expansion of integrated mean squared error cannot be achieved at a local level by a wavelet estimator. Instead, wavelet methods — even our modified ones, which are demonstrably more adaptive than their classical counterparts — rely on the “multiresolution” property to achieve local adaptability, by adjusting levels of high-order resolution in a location-dependent way. That is, they rely on adjustments to the bias term, essentially ignoring variance. We argue that while such an approach confers excellent adaptability against relatively major irregularities such as discontinuities in the curve or its derivatives, it does not adapt well to more subtle variations. In particular, it is clear that there is no satisfactory substitute for a genuine local trade-off of stochastic error against systematic error. It is technically possible to adjust the level of primary resolution in a location-dependent way, but such an

approach yields complicated modifications to wavelet methods that are computationally demanding, aesthetically unappealing and philosophically counter to our desire for simple, widely-applicable techniques.

All the results in this section are described only in outline, with details of mathematical rigour omitted for the sake of brevity.

2.2. Definition of wavelet-based density estimators. We begin by outlining relevant properties of wavelets. Let ϕ and ψ denote compactly supported father and mother wavelet functions, respectively, the former orthonormal under integer translations. Necessarily, they satisfy the properties that for all $k \geq 0$ and $-\infty < l, l_1, l_2 < \infty$,

$$\begin{aligned} \int \psi(x + l_1) \psi(2^k x + l_2) dx &= \delta_{0k} \delta_{l_1 l_2}, \quad \int \phi(x) \phi(x + l) dx = \delta_{0l}, \\ \int \phi(x + l_1) \psi(2^k x + l_2) dx &= 0, \end{aligned} \quad (2.1)$$

where δ_{ij} denotes the Kronecker delta. Assume that the wavelets are of order r , i.e. for some $r \geq 1$ and $\kappa \neq 0$, we have $\int \phi = 1$, $\int \psi = 0$,

$$\begin{aligned} \int x^k \phi(x) dx &= \int x^k \psi(x) dx = 0 \quad \text{for } 1 \leq k \leq r-1, \\ \int x^r \psi(x) dx &= (r!)^{-1} \kappa; \end{aligned}$$

see, for example, Strang (1989, 1993).

Any square-integrable function f admits a generalized Fourier expansion in terms of wavelets:

$$f = \sum_l b_l \phi_l + \sum_{k=0}^{\infty} \sum_l b_{kl} \psi_{kl}, \quad (2.2)$$

where, for $p > 0$ and $p_k = p 2^k$, we define

$$\begin{aligned} \phi_l(x) &= p^{1/2} \phi(px + l), \quad \psi_{kl} = p_k^{1/2} \psi(p_k x + l), \\ b_l &= \int f \phi_l, \quad b_{kl} = \int f \psi_{kl}. \end{aligned}$$

Analytical details are summarized concisely by Strang (1989, 1993) and Daubechies (1992). The generalized Fourier series in (2.2) converges in mean square.

When f is a probability density, and data from the associated distribution are available, an estimator of f may be constructed by developing an empirical version of the expansion (2.2). For example, if $\{X_1, \dots, X_n\}$ denotes a random sample from the distribution with density f then unbiased estimators of b_l and b_{kl} are given by $\hat{b}_l = n^{-1} \sum_i \phi_l(X_i)$ and $\hat{b}_{kl} = n^{-1} \sum_i \psi_{kl}(X_i)$. One may directly substitute $\hat{b}_l$ for b_l in (2.1). This produces the linear wavelet density estimator

$$\tilde{f} = \sum_l \hat{b}_l \phi_l, \quad (2.3)$$

which has many of the properties of a kernel-type estimator; see Hall and Patil (1993a). However, direct substitution of $\hat{b}_{kl}$ for b_{kl} in (2.2) produces a double series which, with probability one, does not converge. To overcome this problem, some sort of inclusion-exclusion rule is typically utilized, for example involving the inclusion of the (k, l) 'th term if and only if (k, l) is in a set $\mathcal{S}$ of integer pairs:

$$\hat{f} = \sum_l \hat{b}_l \phi_l + \sum_{(k,l) \in \mathcal{S}} \hat{b}_{kl} \psi_{kl}. \quad (2.4)$$

Usually one selects those $\hat{b}_{kl}$'s that are particularly large, and terminates the sum over k at some point to ensure convergence. This gives a nonlinear wavelet density estimator, based on a thresholding/truncation rule:

$$\hat{f} = \sum_l \hat{b}_l \phi_l + \sum_{k=0}^{q-1} \sum_l \hat{b}_{kl} I(|\hat{b}_{kl}| > \delta) \psi_{kl}, \quad (2.5)$$

where $\delta > 0$ and $q \geq 1$ are the threshold and truncation parameters, respectively; see for example Hall and Patil (1993a). The value of δ might depend on k , although it generally depends only on n . Our results in this paper concern only the case where δ does not depend on the resolution level k , since this case is of most practical interest. If δ were to

depend on l , i.e. on location, then it should be a function of the way in which f changes with location. This in turn implies that some sort of pilot estimator of f needs to be constructed.

2.3.1. Linear estimator. Assume that f has r bounded, continuous derivatives, and that $f^{(r)}$ is monotone in the extreme tails. Recall definition (2.3) of the linear estimator, $\tilde{f}$. In view of the orthogonality properties noted at (2.1), the integrated squared error of $\tilde{f}$ may be shown to equal

$$\begin{aligned} \int E(\tilde{f} - f)^2 &= \sum_l E(\hat{b}_l - b_l)^2 \\ &\sim n^{-1} p + p^{-2r} (1 - 2^{-2r})^{-1} \kappa^2 \int f^{(r)^2}, \end{aligned} \quad (2.6)$$

the last line following from asymptotic arguments of the type described by Hall and Patil (1993a). It is assumed that $n, p \rightarrow \infty$, with $p/n \rightarrow 0$. (The notation “ $\sim$ ” in (2.6) and below means that the ratio of the left- and right-hand sides converges to 1 as $n \rightarrow \infty$.) If we replace p by h^{-1} then (2.6) has the form typically associated with kernel density estimators:

$$\int E(\tilde{f} - f)^2 \sim (nh)^{-1} C_1 + h^{2r} C_2, \quad (2.7)$$

where C_1 and C_2 are constants immediately recognizable from (2.6). Compare this formula with, for example, equation (3.20), p. 40 of Silverman (1986). Thus, the asymptotically optimal choice of p is $p \sim \text{const. } n^{1/(2r+1)}$.

2.3.2. Nonlinear estimator. We begin with the inclusion-exclusion estimator defined at (2.4), and note that in a mean square sense its optimal construction involves including the (k, l) 'th term in that series if and only if

$$\text{var } \hat{b}_{kl} < b_{kl}^2. \quad (2.8)$$

When the index set $\mathcal{S}$ is non-stochastic, this recommendation follows straightforwardly by

taking expectations in Parseval's identity:

$$\int E(\hat{f} - f)^2 = \sum_l E(\hat{b}_l - b_l)^2 + \sum_{(k,l) \in \mathcal{S}} E(\hat{b}_{kl} - b_{kl})^2 + \sum_{(k,l) \notin \mathcal{S}} b_{kl}^2. \quad (2.9)$$

When $\mathcal{S}$ is stochastic, a relation of the form (2.9) is generally true in an approximate, asymptotic sense, and so again the rule (2.8) emerges as an (approximately) optimal one. Thresholding, introduced at (2.5), might be thought of as an attempt to implement the rule (2.8) approximately, by selecting those terms $\hat{b}_{kl}$ which are large. When f is smooth (i.e., in the present context, has r piecewise continuous derivatives), the estimator $\hat{f}$ defined at (2.5) admits the mean integrated squared error formulae at (2.6) and (2.7):

$$\int E(\hat{f} - f)^2 \sim n^{-1} p + p^{-2r} (1 - 2^{-2r})^{-1} \kappa^2 \int f^{(r)2}, \quad (2.10)$$

for appropriate choice of p (near the optimal rate $n^{1/(2r+1)}$), q and δ . Details of proofs are given by Hall and Patil (1993b). Minimum mean integrated squared error is of order $n^{-2r/(2r+1)}$. The fact that (2.10) is so similar to its counterpart for mean integrated squared error of a globally smoothed kernel estimator is one manifestation of the phenomenon we observe that wavelet estimators do not provide adequate local smoothing when the underlying curve is smooth. To provide adequate smoothing we would have to vary the level of primary resolution, p , with location, l , in the same way that a locally adaptive kernel density estimator requires bandwidth to be varied with location. This approach is feasible for a wavelet estimator, but it is nonetheless very impractical and, we contend, quite aesthetically unappealing.

The root cause of the failure of wavelet methods to be truly locally adaptive is that thresholding fails to adequately identify all but extremely large b_{kl} 's, since for smooth curves and optimally chosen thresholds, the stochastic error $\hat{b}_{kl} - b_{kl}$ is of the same size as b_{kl} itself. To appreciate why, observe that $\hat{b}_{kl}$ is unbiased for b_{kl} and has asymptotic variance given by $n^{-1} f(l/p_k)$, as p_k tends to infinity. Furthermore, by Taylor expansion,

for $k \geq 0$,

$$\begin{aligned} b_{kl} &= p_k^{-1/2} \int \psi(x) f\{(x+l)/p_k\} dx \\ &\sim p_k^{-(2r+1)/2} \kappa f^{(r)}(l/p_k), \end{aligned} \quad (2.11)$$

which is of size $n^{-1/2}$ when p_k is of size $n^{1/(2r+1)}$. (The order $n^{1/(2r+1)}$ for $p_k = p2^k$ is optimal in the sense of minimizing mean squared error — in particular, the optimal p is, by (2.10), of size $n^{1/(2r+1)}$, and for large k the contribution of b_{kl} to mean squared error is negligibly small.) Since $\hat{b}_{kl}$ is unbiased for b_{kl} then this argument demonstrates that the error about the mean of $\hat{b}_{kl}$ is of the same order as b_{kl} , verifying our claim.

The threshold δ has to be of a larger order of magnitude than $n^{-1/2}$, in fact at least $(n^{-1} \log n)^{1/2}$, if the accumulation of stochastic errors is not to seriously degrade the performance of the estimator $\hat{f}$. The minimal order of magnitude of δ derives from the fact that the chance that $\hat{b}_{kl} - b_{kl}$ exceeds $C(n^{-1} \log n)^{1/2}$ is not larger than $n^{-\varepsilon}$, where the value of ε depends on C . Hall and Patil (1993a) addressed the issue of smallest possible threshold in considerable detail. Since for a smooth, r -times differentiable function f the size of b_{kl} is no greater than $n^{-1/2}$ (see (2.11)), it follows that thresholding alone cannot hope to be effective in assessing the true size of b_{kl} in the context of curvature changes for a smooth curve, where fluctuations of b_{kl} with varying l are only by a constant factor, rather than by an order of magnitude. Thresholding only reacts to relatively dramatic increases in b_{kl} — specifically, those involving a factor of $(\log n)^{1/2}$ or greater — associated with unusual episodes in the curve, such as discontinuities. This observation holds whether or not δ depends on k .

Purely threshold-based wavelet methods have been suggested in the literature, primarily being promoted on the basis of their adaptivity. These methods involve fixing p (specifically, p does not then depend on sample size), and then employing the threshold δ alone to adjust the estimator. In this setting an analogue of formula (2.10) may be proved under appropriate regularity conditions on f : if $\delta \geq C(n^{-1} \log n)^{1/2}$, for C sufficiently

large, and if $\delta \rightarrow 0$, then

$$\int E(\hat{f} - f)^2 \sim \delta^{4r/(2r+1)} (1 - 2^{-2r})^{-1} |\kappa|^{2/(2r+1)} \int |f^{(r)}|^{2/(2r+1)}. \quad (2.12)$$

(This result is related to Proposition 2.1 of Hall and Patil (1993b), and may be proved in similar fashion.) Note particularly that the optimal convergence rate for the purely thresholded estimator is only $(n^{-1} \log n)^{2r/(2r+1)}$, which is slower than the $n^{-2r/(2r+1)}$ rate achievable using the estimator $\hat{f}$ with p of size $n^{1/(2r+1)}$. Thus, the convergence rate of the estimator based on pure thresholding is not as good as that obtainable by selecting both the threshold and the primary level of resolution, p , to depend on at least the sample size in an appropriate way. In essence, this phenomenon arises because the purely thresholded estimator has not been smoothed. The term on the right-hand side of (2.12) derives entirely from the integral of squared bias, and is unrelated to variance. Indeed, the operation of thresholding does not involve statistical smoothing, to first order, since it does not adjust variance to first order; that is, to first order, it influences only the bias term. As a result, the purely thresholded density estimator is undersmoothed, with a variance contribution of size n^{-1} and a squared bias of size $(n^{-1} \log n)^{2r/(2r+1)}$. The variance contribution only becomes non-negligible when p is increased to order $n^{1/(2r+1)}$, which produces optimality.

Sometimes thresholding is conducted in a “milder” way than simple inclusion-exclusion, for example by replacing the zero-one indicator function in the estimator defined at (2.5) by a function which takes a continuous range of values over a similar range of its argument. This approach gives rise to alternative estimators, such as

$$\hat{f} = \sum_l \hat{b}_l \phi_l + \sum_{k=0}^{q-1} \sum_l \hat{b}_{kl} w(\hat{b}_{kl}/\delta) \psi_{kl},$$

where for constants $0 < c_1 < c_2 < \infty$ the weight function w satisfies $w(u) = 1$ for $0 < u < c_1$, $0 \leq w(u) \leq 1$ for $c_1 \leq u \leq c_2$, and $w(u) = 1$ for $u > c_2$. However, the arguments above apply without essential change to such “smoothly thresholded” estimators. Smooth

thresholding does not alter the fact that thresholding does not constitute a smoothing operation, since it does not adjust variance to first order. Consequently, it shares with ordinary “hard thresholding” the drawbacks that we have noted.

The arguments in this section explain why thresholding without local choice of p cannot adequately respond to relatively subtle changes in the smoothness of a curve, even if thresholding is accompanied by appropriate global choice of p ; and why purely threshold-based methods cannot quite achieve optimal convergence rates for smoothness classes, even in a global sense.

2.4. Modifications. We begin by describing modifications to the inclusion-exclusion rule that would be practicable if it were feasible to estimate b_{kl} with greater accuracy than order $n^{-1/2}$. Then we suggest a simple way of estimating b_{kl} more accurately for smooth curves, thereby making our procedures practicable. If we could assess the size of b_{kl} accurately then, instead of the rudimentary inclusion-exclusion rule furnished by thresholding, we could attempt something more sophisticated like shrinkage. For example, we might modify the estimator $\hat{f}$, defined at (2.5), by replacing $\hat{b}_{kl}$ by $\hat{b}_{kl} s_{kl}$, for a shrinkage parameter $s_{kl} \in (0, 1)$, yielding the estimator

$$\tilde{f} = \sum_l \hat{b}_l \phi_l + \sum_{k=0}^{q-1} \sum_l s_{kl} \hat{b}_{kl} \psi_{kl}.$$

Of course, mean squared error is minimized by selecting

$$s_{kl} = b_{kl}^2 / (b_{kl}^2 + \text{var } \hat{b}_{kl}) \sim b_{kl}^2 / \{b_{kl}^2 + n^{-1} f(l/p_k)\}.$$

An empirical version of this quantity may be derived by replacing the unknowns $f(l/p_k)$ and b_{kl} by their respective estimators. Alternatively, we could simply replace $\hat{b}_{kl}$ in (2.5) by the more accurate estimator of b_{kl} , without resorting to shrinkage. An estimator of b_{kl} that converges at rate $o(n^{-1/2})$ is given by

$$\tilde{b}_{kl} = (2t + 1)^{-1} \sum_{m=-t}^t \hat{b}_{k, l+m},$$

where $t = t(n)$ diverges to infinity at a suitable rate as n increases.

The asymptotic integrated squared error of the shrinkage-based estimator is given by

$$n^{-1} p + p^{-2r} \kappa^2 \int f f^{(r)^2} \\ \times \sum_{k=0}^{\infty} 2^{-2rk} (f + n p^{-(2r+1)} 2^{-(2r+1)k} \kappa^2 f^{(r)^2})^{-1},$$

which represents an improvement in performance over that obtained using a thresholded estimator. However, note that this improvement has not addressed the size of the variance term, $n^{-1} p$, which derives from the level p of primary resolution. Only by adjusting that quantity in a local sense can we hope to achieve genuine local adaptivity in the context of smooth function estimation. Optimal selection of the rate of increase of t to infinity depends on the modulus of continuity of $f^{(r)}$, and is not investigated here. Instead, in Section 3 we introduce smoothing rules which we regard as more practicable and intuitively appealing for smooth curves.

A version of this method is appropriate when f or its first $r - 1$ derivatives have jump discontinuities. It amounts to using the wavelet coefficients as diagnostics for the discontinuities, and applying the techniques just described between the points of discontinuity.

3. Adaptive bandwidth choice for nonparametric density estimation and regression. We shall develop a locally adaptive bandwidth selection rule for kernel density estimators, and briefly discuss its applications to nonparametric regression. Numerical applications of our method will be described in Section 4.

As in Section 2, let $\{X_1, \dots, X_n\}$ denote a random sample from a distribution with unknown density f , for which we wish to construct an estimator. Denote by K an r 'th order kernel, meaning that

$$\int x^k K(x) dx = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } 1 \leq k \leq r - 1 \\ r! \lambda \neq 0 & \text{if } k = r \end{cases} \quad (3.1)$$

For convenience, we shall take $r \geq 2$, and define the corresponding kernel density estimator by

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^n K\{(x - X_i)/h\}, \quad (3.2)$$

which is designed for densities that possess r derivatives — see for example Silverman (1986, p. 66ff). The pointwise mean squared error of $\hat{f}$ is given by the familiar formula,

$$E(\hat{f}_h - f)^2 \approx (nh)^{-1} \left(\int K^2 \right) f + h^{2r} \lambda^2 f^{(r)^2},$$

which is asymptotically minimized by taking

$$h = h(x) = h_0 g_0(x), \quad (3.3)$$

where $h_0 = \text{const} \cdot n^{-1/(2r+1)}$ and $g_0(x) = \{f(x)/2r f^{(r)}(x)^2\}^{1/(2r+1)}$. Thus, our task is to empirically select a function $g(x)$, which we would then “plug in” to formula (3.3) and use as our bandwidth selector.

If cross-validation is used as the device for choosing g then it is necessary to select a relatively smooth function. For details, the reader is referred to the treatment of the term D_2 in Step (iii) of the proof of the Theorem, given in Section 5, particularly the discussion within square brackets there. We ask that g be chosen from the class $G = G(B_1, \dots, B_4)$ of all functions which vanish outside a compact interval $[-B_1, B_1]$, and satisfy $B_2 < g < B_3$ and $|g^{(r)}| \leq B_4$, where $B_1, \dots, B_4$ are positive constants. Of course, the function g_0 is not necessarily in such a class, but this hardly matters from a practical point of view, since the empirical choice of g is only an approximation to g_0 , which may in turn be approximated arbitrarily closely by functions in classes G for appropriate values of the parameters. The proof of our main result continues to be valid if the parameters B_2^{-1} , B_3 and B_4 are allowed to be functions of n and to diverge to infinity at a rate no faster than n^ϵ , for sufficiently small $\epsilon > 0$; and if the interval $[-B_1, B_1]$ is allowed to increase to the support of f sufficiently slowly. (The latter generalization is not important in the context of the theorem that we shall state below, as there f is assumed to be compactly supported.)

The requirement for $g \in G$ is motivated from both theoretical and practical considerations. Practically, one is interested in estimating a function over an interval $[-B_1, B_1]$. Theoretically, we want to bound g sufficiently far away from zero and infinity so that neither variance or bias becomes too large, leading to the condition $B_2 \leq g \leq B_3$. The condition $|g^{(r)}| \leq B_4$ dictates the smooth transition of local bandwidths so that one does not obtain a “granular” curve. Suffice to say that the class G is rich enough so that the minimum of MISE over all possible choice of g is not significantly smaller than that over $g \in G$.

Take $h = h_0 g$ in the definition (3.2) of $\hat{f}_h$. We write $\hat{f}_g$ to stress the dependence on $g \in G$. The cross-validatory criterion CV is given by

$$\text{CV}(g) = \int \hat{f}_g^2 - 2\{n(n-1)\}^{-1} \sum_{i_1 \neq i_2} \sum h(X_{i_1})^{-1} K\{(X_{i_1} - X_{i_2})h(X_{i_1})^{-1}\},$$

which is an unbiased estimator of

$$\text{MISE}(g) - \int f^2, \quad (3.4)$$

where $\text{MISE}(g) = \int E(\hat{f}_g - f)^2$ denotes mean integrated squared error. Since the subtracted term in (3.4) does not depend on g then it is anticipated that selection of g to minimize $\text{CV}(g)$ will asymptotically minimize $\text{MISE}(g)$, up to terms which either do not depend on g or are negligibly small. The Theorem below will make explicitly clear that this is indeed the case.

We are now in a position to state our main theorem. Let $\text{ISE}(g) = \int (\hat{f}_g - f)^2$ denote integrated squared error, and denote $T = \int f^2 + 2n^{-1} \sum \{f(X_i) - \int f^2\}$, a quantity which does not depend on g . Assume that

K is compactly supported, satisfies (3.1)

and is Lipschitz continuous with exponent $2/r$. (3.5)

The assumption of Lipschitz continuity here is used in the proof to derive bounds to non-linear components of both mean squared error and the cross-validatory criterion over large

function classes. The condition of compact support may be removed at the expense of greater complication in the proofs. In particular, the Standard Normal kernel is permissible. We suppose that the density f has the following property:

$$f \text{ has } r \text{ bounded derivatives and is compactly supported.} \quad (3.6)$$

Again, the condition of compact support is imposed only for the sake of simplicity in proofs, and may be removed at the expense of greater effort.

THEOREM. *Assume conditions (3.5) and (3.6). Then with probability one,*

$$\sup_{g \in G} \left| \text{CV}(g) + T - \int (\hat{f}_g - f)^2 \right| = o\{(nh_0)^{-1}\} \quad (3.7)$$

and

$$\sup_{g \in G} \left| \int (\hat{f}_g - f)^2 - \int E(\hat{f}_g - f)^2 \right| = o\{(nh_0)^{-1}\} \quad (3.8)$$

as $n \rightarrow \infty$.

Thus, under the conditions of the theorem, minimizing CV over g is asymptotically equivalent to minimizing either ISE or MISE over g . It is straightforward to develop versions of the Theorem which apply to cases where the minimization is over a sequence of finite sets G_n , which approximate G arbitrarily closely as $n \rightarrow \infty$. Indeed, empirical methods for implementing the procedure implicit in the theorem must necessarily be conducted on a mesh of functions, rather than in the continuum. We shall give details in Section 4. However, note that the Theorem can fail if the functions within G are not sufficiently smooth; see Step (iii) in Section 5.

An advantage of our approach to local smoothing parameter selection is that it may be very easily adapted to a wide variety of different contexts. These include modifications of the application, such as to nonparametric regression rather than density estimation, and modifications of the method, such as smoothed cross-validation instead of standard cross-validation. In particular, let us consider the version of our technique for nonparametric

regression. Here we suppose that paired data (x_i, Y_i) are generated by the model

$$Y_i = m(x_i) + \epsilon_i,$$

where the function m is specified only in terms of smoothness assumptions, and the independent errors ϵ_i have zero mean. Let $h = h_0 g$, where $g \in G$. The Nadaraya-Watson kernel estimator of m is given by

$$\hat{m}(x) = \left[\sum_{i=1}^n K\{(x - x_i)/h(x)\} Y_i \right] / \left[\sum_{i=1}^n K\{(x - x_i)/h(x)\} \right],$$

in which K may be taken to be a kernel function specified by (3.1). The version of CV in this context is given by

$$\sum_{i=1}^n \{Y_i - \hat{m}_{-i}(x_i)\}^2$$

where

$$\hat{m}_{-i}(x) = \left[\sum_{j \neq i} K\{(x - x_j)/h(x)\} Y_j \right] / \left[\sum_{j \neq i} K\{(x - x_j)/h(x)\} \right].$$

In these formulae, $h = h_0 g$ should be chosen to be an r -times differentiable function. Minimization of CV may be shown to lead to asymptotic minimization of $\int (\hat{m} - m)^2 f$, where f denotes the asymptotic density of the design points x_i and is assumed to be bounded away from zero on its support. The version of our method for smoothed cross-validation, in the context of either density estimation or regression, may be easily constructed and shown to have the properties attributed to ordinary cross-validation in, for example, the Theorem above. Smoothed cross-validation has been discussed by Hall, Marron and Park (1992).

A version of the Theorem may be proved for density functions whose r 'th derivative increases with n , in particular at a rate not exceeding n^ϵ for ϵ sufficiently small. In such cases the optimal local bandwidth is of smaller order than $n^{-1/(2r+1)}$, and it may be shown that for small ϵ , local cross-validation succeeds in capturing local variability in this more general context. For the sake of brevity, we do not give details here.

4. Numerical results.

In order to assess the practical impact of our locally adaptive cross-validation technique, we conducted a simulation study in which we compared our new technique with an ordinary global cross-validation method. Implementation of our technique was straightforward, the only step requiring care involving minimization of the cross-validatory criterion $CV(g)$. In practice, the function g may be obtained as a smooth interpolant across a relatively coarse grid of points, at each of which an appropriate bandwidth was chosen to minimize $CV(g)$. Suppose the grid of points chosen is $(a_1, \dots, a_p)$, and the corresponding bandwidths are $(h_1, \dots, h_p)$. Then g is defined as the cubic spline interpolant of the pairs $\{(a_1, h_1), \dots, (a_p, h_p)\}$ which minimize $CV(g)$. In practical terms, such a requirement amounts to regarding $CV(g)$ as a function of $(h_1, \dots, h_p)$, say $CV(h_1, \dots, h_p)$, and minimizing that function over a suitable set of values of $(h_1, \dots, h_p)$. A minimal requirement on the values of $h_1, \dots, h_p$ is that they each be positive. In practice it is desirable to further restrict the bandwidths so that the individual h_i 's don't fall too far from "sensible" values for bandwidths. An effective way of ensuring that the latter restriction is satisfied is to restrict the minimization process to only consider values of h_i that fall within a reasonable distance of the estimated bandwidth resulting from consideration of a global cross-validatory criterion. Specifically, if $\hat{h}$ minimizes

$$CV(h) = \int \hat{f}_h^2 - \frac{2}{n(n-1)h} \sum_{(i_1, i_2: 1 \leq i_1, i_2 \leq n, i_1 \neq i_2)} K\left(\frac{X_{i_1} - X_{i_2}}{h}\right),$$

where $\hat{f}_h(x)$ is given by (3.2), then the h_i , $1 \leq i \leq n$ are restricted to values satisfying

$$\zeta h \leq h_i \leq \zeta^{-1} h, \quad (4.1)$$

for some $\zeta > 0$. This restriction reformulates the requirement that $|g^{(2)}| \leq B_4$. For practical purposes, the latter condition is mild, and the cross-validatory criterion is usually minimized in the interior of the region defined by (4.1). In our simulations, we have found that $\zeta = 0.1$ is a good choice for ζ . Our calculations were programmed in Fortran on a

Sparc 10 workstation, and the numerical minimization step was carried out using NAG subroutine E04JAF.

In our numerical study, we considered several parent densities that were normal mixtures with varying numbers and sizes of modes. Our aim in this study was to show the efficacy of our method in picking up the various peaks and troughs in the true density. The method based on the global CV criterion was expected to do well in simple unimodal examples, but to falter in cases where several modes of varying size are present. In the latter case, the value of the bandwidth required to accurately estimate the largest mode is typically too small in flatter parts of the density, resulting in undersmoothing in those parts of the curve. By way of contrast, our method adapted the degree of smoothing along the curve to capture both sharp and mild peaks relatively well. Our results are summarized in Figures 1(a) through 4(b), the plots to be read in pairs, the first showing the relative performance of the estimated densities, and the second showing how the bandwidth function changes for different values of x in our local kernel method. In each plot, solid lines denote our locally adaptive method, small-dotted lines represent the ordinary kernel technique, and the large dotted line denotes the true underlying density. Figures 1(a) and 1(b) correspond to the case where the underlying density is unimodal. In this case, our method performs marginally worse than the usual kernel method, although the differences were not significant. Figures 2(a) to 2(d) correspond to the case where the underlying density was a bimodal normal mixture, for samples of size 1000 and 2000. In these cases, the ordinary kernel method performed well in estimating the sharper of the two peaks, but it performed disastrously in estimating the milder peak. By way of contrast, our method captured both peaks very well for each of the two sample sizes considered. Figures 3(a) to 3(d) concern underlying trimodal normal mixture densities, and sample sizes 1000 and 2000. In these cases, the kernel method again performs well at the sharpest peaks but very poorly elsewhere. In this case, the estimate based on our method exhibited some spurious peaks indicating mild undersmoothing in one part of the curve, although these

effects were not serious. Moreover, the quality of our estimator improved as sample size increased from 1000 to 2000. Finally, Figure 4(a) shows the estimated density for a normal mixture with four modes. Again, our method performs well in estimating the four peaks, while the ordinary kernel method struggled badly apart from at the sharpest mode. Again, our method exhibited mild undersmoothing at one of the peaks, but this problem was not severe, and we anticipate that such occurrences become rarer as sample size increases. The accompanying plots relating bandwidth to location show how our method adaptively adjusts the bandwidth. For example, Figures 2(b) and 2(d) show that smaller bandwidths correspond to modes in the density, so that these figures suggest that the true density has peaks between 0 and 0.5 and at 1.5. As the number of modes in the true density increases, the bandwidth plots become harder to interpret because the data arising from one component of the mixture can intermingle with data from other components.

We also conducted simulation studies for samples smaller than 1000. Our methods worked fairly well down to samples of size around 200, although the number of spurious modes indicated by our method naturally increases as sample size decreased. The accuracy of our method dropped off significantly for samples of size 100. Of course, this phenomenon is simply a manifestation of the fact that one can hardly expect to estimate a density with many modes accurately without sufficient data.

5. Proofs. We derive only result (3.7), as (3.8) has a similar but simpler proof. For ease of exposition our proof is divided into three parts.

Step (i). Notation. Define

$$\begin{aligned} U(x_1, x_2) &= h(x_1)^{-1} K\{(x_1 - x_2) h(x_1)^{-1}\}, \\ U_1(x) &= E\{U(x, X)\}, \quad U_2(x) = E\{U(X, x)\}, \\ u &= E\{U_1(X)\} = E\{U_2(X)\}, \\ V(x_1, x_2) &= U(x_1, x_2) - U_1(x_1) - U_2(x_2) + u, \end{aligned}$$

$$\begin{aligned}
V_j(x) &= U_j(x) - f(x), \quad v = E\{V_1(X)\} = E\{V_2(X)\}, \\
D_0 &= \{n(n-1)\}^{-1} \sum_{i_1 \neq i_2} \sum V(X_{i_1}, X_{i_2}), \quad D_j = n^{-1} \sum_{i=1}^n \{V_j(X_i) - v\} \quad (j = 1, 2), \\
D_3 &= \int f \hat{f} - \int f E \hat{f}, \quad D = \int (\hat{f} - f)^2, \\
CV_1 &= CV + \int f^2 - 2n^{-1} \sum_{i=1}^n \left\{ f(X_i) - \int f^2 \right\}.
\end{aligned}$$

In this notation,

$$D - CV_1 = 2(D_0 + D_1 + D_2 - D_3). \quad (5.1)$$

All these quantities depend on the function g through the bandwidth $h = h_0 g$, and when necessary we shall express the dependence by writing g as an argument of D , D_j or CV .

Step (ii). Bounds to D_0 .

Case 1. Coarse grid of g 's. Put $W(x_1, x_2) = V(x_1, x_2) + V(x_2, x_1)$,

$$Y_{i_2} = \sum_{i_1=1}^{i_2-1} W(X_{i_1}, X_{i_2}).$$

Then

$$n(n-1) D_0 = \sum_{1 \leq i_1 < i_2 \leq n} \sum W(X_{i_1}, X_{i_2}) = \sum_{i=2}^n Y_i.$$

The Y_i 's are martingale differences, and so by Rosenthal's inequality (Hall and Heyde 1980, p.23), for all $s \geq 1$,

$$\begin{aligned}
E \left| \sum_{i=2}^n Y_i \right|^{2s} &\leq C(s) \left[E \left\{ \sum_{i=2}^n E(Y_i^2 \mid Y_1, \dots, Y_{i-1}) \right\}^s + \sum_{i=2}^n E|Y_i|^{2s} \right] \\
&\leq 2C(s) \left\{ \sum_{i=2}^n (E|Y_i|^{2s})^{1/s} \right\}^s,
\end{aligned}$$

where $C(s) \geq 1$ denotes a constant depending only on s . Conditional on X_i , Y_i is a sum of independent random variables with zero means. Therefore,

$$E|Y_i|^{2s} \leq C(s) \left(E \left[\sum_{i_1=1}^{i-1} E\{W(X_{i_1}, X_i)^2 \mid X_i\} \right]^s + (i-1) E|W(X_1, X_2)|^{2s} \right).$$

Under the conditions stated in the theorem, $E\{W(X, x)^2\} \leq C_1 h_0^{-1}$ uniformly in x , and $E|W(X_1, X_2)|^{2s} \leq C_2 h_0^{1-2s}$, where $C_1, C_2, \dots$ are constants depending on K, f and the parameters of G (not on g). Therefore,

$$E \left| \sum_{i=2}^n Y_i \right|^{2s} \leq C_3^s C(s)^2 (n^{2s} h_0^{-s} + n^{s+1} h_0^{1-2s}),$$

whence

$$E|D_0|^{2s} \leq C_4^s C(s)^2 (nh_0)^{-2s} h_0^s. \quad (5.2)$$

Results of Burkholder (1973) may be used to show that we may take $C(s) = \{C_0 s(\log s)^\epsilon\}^{2s}$, for any $\epsilon > 0$, where $C_0 = C_0(\epsilon)$ does not depend on s . Hence, by Markov's inequality, for any $\eta > 0$,

$$\sup_{g \in G} P\{|D_0| > \eta(nh_0)^{-1}\} \leq (C_5/\eta)^{2s} \{s(\log s)^\epsilon h_0^{1/4}\}^{4s}. \quad (5.3)$$

Taking $s = h_0^{-\{1-(2r+1)\xi\}/4}$, for arbitrary $0 < \xi < (2r+1)^{-1}$, we see that

$$\sup_{g \in G} P\{|D_0| > \eta(nh_0)^{-1}\} = O\{\exp(-n^{\{4(2r+1)\}^{-1}-\xi})\}$$

for all $\xi > 0$. Therefore, if G_n is a finite subset of G with

$$\#G_n = O\{\exp(n^{\{4(2r+1)\}^{-1}-\xi})\}$$

for some $\xi > 0$, then

$$\sup_{g \in G_n} |D_0| = o\{(nh_0)^{-1}\} \text{ a.s.}$$

as $n \rightarrow \infty$.

Case 2. Fine grid of g 's. Let $\epsilon, \zeta > 0$, and define $H_{n,\zeta}$ to be the class of all sequences $\{(in^{-\zeta}, j_i \epsilon n^{-r\zeta}), -\infty < i < \infty\}$, for integer pairs (i, j_i) , which are such that $(r+1)$ -point, r 'th degree polynomial interpolation among the function values $g^*(in^{-\zeta}) = j_i \epsilon n^{-r\zeta}$ produces a function $g \in G$. Let $G_{n,\zeta}$ denote the class of all functions derived by their polynomial interpolation. Then $\#G_{n,\zeta} \leq \#H_{n,\zeta} = O\{\exp(Cn^\zeta)\}$ for some $C > 0$. (To

appreciate why, observe that since $B_2 \leq g \leq B_3$ and $|g^{(r)}| \leq B_4$ then, given the value of $g^*(in^{-\zeta})$, the number of possible values of $g^*\{(i+1)n^{-\zeta}\}$ such that $g \in G_{n,\zeta}$ equals $O(1)$ uniformly in i .) By choosing ϵ sufficiently small, not depending on ζ , we may ensure that for all $\zeta > 0$,

$$\sup_{g \in G} \inf_{g_1 \in G_{n,\zeta}} \sup_x |g(x) - g_1(x)| = O(n^{-r\zeta}). \quad (5.4)$$

We shall employ such an ϵ on all occasions, and hence do not indicate its value in the notation $G_{n,\zeta}$.

In view of the conclusion of the previous part of the proof, for each $\zeta < \{4(2r+1)\}^{-1}$ we have

$$\sup_{g \in G_{n,\zeta}} |D_0| = o\{(nh_0)^{-1}\} \text{ a.s.} \quad (5.5)$$

Now let $0 < \zeta_1 < \zeta_2 < \infty$ be arbitrary, let $g_2 \in G_{n,\zeta_2}$, and let $g_1 \in G_{n,\zeta_1}$ denote that element of G_{n,ζ_1} which minimizes $\sup_x |g_1(x) - g_2(x)|$. In view of (5.4),

$$\sup_{g_2 \in G_{n,\zeta_2}} \sup_x |g_1(x) - g_2(x)| = O(n^{-r\zeta_1}).$$

Now rework all the steps in the previous part of the proof, with D_0 replaced by $D_0(g_1) - D_0(g_2)$. We obtain instead of (5.2) the result

$$\sup_{g_2 \in G_{n,\zeta_2}} E|D_0(g_1) - D_0(g_2)|^{2s} \leq C_6^s C(s)^2 (nh_0)^{-2s} h_0^s (n^{-r\zeta_1})^{2\delta s},$$

where δ denotes the modulus of Hölder continuity of the kernel K . Therefore, (5.3) holds in the form

$$\begin{aligned} \sup_{g_2 \in G_{n,\zeta_2}} P\{|D_0(g_1) - D_0(g_2)| > \eta(nh_0)^{-1}\} \\ \leq (C_7/\eta)^{2s} \{s(\log s)^\epsilon h_0^{1/4} n^{-(1/2)r\delta\zeta_1}\}^{4s}, \end{aligned}$$

whence it follows that if

$$\#G_{n,\zeta_2} = O\left\{\exp\left(n^{\{4(2r+1)\}^{-1} + (r\delta\zeta_1/2) - \xi}\right)\right\}$$

for some $\xi > 0$, then

$$\sup_{g_2 \in G_{n, \zeta_2}} |D_0(g_1) - D_0(g_2)| = o\{(nh_0)^{-1}\} \text{ a.s.}$$

If in addition

$$\sup_{g \in G_{n, \zeta_1}} |D_0(g)| = o\{(nh_0)^{-1}\} \text{ a.s.}$$

(see (5.5), for example) then

$$\sup_{g \in G_{n, \zeta_2}} |D_0(g)| = o\{(nh_0)^{-1}\} \text{ a.s.} \quad (5.6)$$

Hence, arguing by induction with (5.5) as the first step, we may prove (5.6) with successive values of ζ_2 equal to $\{4(2r+1)\}^{-1} - \xi$, $\{4(2r+1)\}^{-1} (1 + \frac{1}{2} r\delta) - \xi$, $\dots$, $\{4(2r+1)\}^{-1} \{1 + \frac{1}{2} r\delta + \dots + (\frac{1}{2} r\delta)^k\} - \xi$, $\dots$, for arbitrary $\xi > 0$. Since $\delta \geq 2/r$, by hypothesis, then (5.6) holds for all $\zeta_2 > 0$. It is then a trivial matter, noting the Hölder continuity of K and taking ζ_2 sufficiently large, to prove that

$$\sup_{g \in G} |D_0(g)| = o\{(nh_0)^{-1}\} \text{ a.s.} \quad (5.7)$$

Step (iii). Bounds to D_1, D_2 and D_3 . These quantities are “linear” in the data, in the sense that they may be written in the form $\int a d(\hat{F} - F)$ where a is a function, $\hat{F}$ is the empirical distribution function, and $F = E(\hat{F})$ is the distribution function with density f . This means that they are relatively straightforward to bound using empirical process arguments. The result

$$E(D_j) = 0 \quad \text{and} \quad E(D_j^2) = O(n^{-1} h_0^{2r}), \quad (5.8)$$

uniformly in $g \in G$, is very simple to prove for $j = 1$ and 3. When $j = 2$, the first part of (5.8) is trivial and the second part follows from the fact that g has r uniformly bounded derivatives:

$$U_2(x) = \int g(x + h_0 z)^{-1} K\{zg(x + h_0 z)^{-1}\} f(x + h_0 z) dz = f(x) + O(h_0^r) \quad (5.9),$$

uniformly in x , so that $V_2(x) = U_2(x) - f(x) = O(h_0^r)$ and

$$E(D_2^2) = n^{-1} \text{var}\{V_2(X_1)\} = O(n^{-1} h_0^{2r}).$$

Now apply Theorem 37, page 34 of Pollard (1984), to prove that

$$\sup_{g \in G} |D_j(g)| = o\{(nh_0)^{-1}\} \text{ a.s.} \quad (5.9)$$

for $j = 1, 2, 3$. [Note that (5.9) fails if g is not smooth. For example, if g is a step function then $U_2(x) - f(x)$ is of size h_0 for x within $O(h_0)$ of a jump in g , whence it follows that $E(D_2^2)$ is of size $n^{-1} h_0^3$. In particular, (5.8) fails even in the simplest case $r = 2$, and it may be shown that (3.7) fails too. Indeed, $CV(g) + T - \int(\hat{f}_g - f)^2$ is of size $(nh_0)^{-1}(nh_0^5)^{1/2}$ instead of $o\{(nh_0)^{-1}\}$.]

The theorem follows from (5.1), (5.7) and (5.10).

Acknowledgements. The work of Fan was partially supported by an Australian Research Council grant and NSF grant No. DMS-9203135; that of Martin was partially supported by NSF grant No. INT-8913333 under the NSF Cooperative Program; and the work of both Fan and Martin was conducted partly during visits to the Centre for Mathematics and its Applications at the ANU.

REFERENCES

- BROCKMANN, M., GASSER, T. AND HERMANN, E. (1993). Locally adaptive bandwidth choice for kernel regression estimation. *J. Amer. Statist. Assoc.*
- BURKHOLDER, D.L. (1973). Distribution function inequalities for martingales. *Ann. Probab.*, **1**, 19-42.
- DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: SIAM.
- DONOHU, D.L. (1992). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. Technical report No. 403, Department of Statistics, Stanford University.

- DONOHO, D.L. AND JOHNSTONE, I.M. (1992a). Ideal spatial adaptation by wavelet shrinkage. Technical report No. 400, Department of Statistics, Stanford University.
- DONOHO, D.L. AND JOHNSTONE, I.M. (1992b). Minimax risk over ℓ_p -balls for ℓ_q -error. Technical report No. 401, Department of Statistics, Stanford University.
- DONOHO, D.L. AND JOHNSTONE, I.M. (1992c). Minimax estimation via wavelet shrinkage. Technical report No. 402, Department of Statistics, Stanford University.
- FAN, J. AND GIJBELS, I. (1993). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. Institute of Statistics Mimeo Series, Department of Statistics, University of North Carolina, Chapel Hill.
- HALL, P. AND HEYDE, C.C. (1980). *Martingale Limit Theory and its Application*. Academic Press, New York.
- HALL, P., MARRON, J.S. AND PARK, B.U. (1992). Smoothed cross-validation. *Probab. Th. Rel. Fields* **92**, 1–20.
- HALL, P. AND SCHUCANY, W.R. (1989). A local cross-validation algorithm. *Statist. Probab. Lett.* **8**, 109–117.
- HALL, P. AND PATIL, P. (1993a). On Wavelet methods for estimating smooth functions. CMA Research Report SMS-37-93, Australian National University.
- HALL, P. AND PATIL, P. (1993b). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. CMA Research Report SMS-49-93, Australian National University.
- JONES, M.C. (1990). Variable kernel density estimates and variable kernel density estimates. *Aust. J. of Statist.* **32**, 361–372.
- KERKYACYARIAN, G. AND PICARD, D. (1992). Density estimation in Besov spaces. *Statist. Probab. Lett.* **13**, 15–24.
- KERKYACYARIAN, G. AND PICARD, D. (1993a). Density estimation by kernel and wavelet methods, optimality in Besov spaces. Manuscript.
- KERKYACYARIAN, G. AND PICARD, D. (1993b). Linear wavelet methods and other

periodic kernel methods. Manuscript.

KERKYACYARIAN, G. AND PICARD, D. (1993c). *Introduction aux Ondelettes et Estimation de Densité, 1: Introduction aux Ondelettes et a l'Analyse Multiresolution*. Lecture Notes.

MIELNICZUK, J. SARDA, P. AND VIEU, P. (1989). Local data-driven bandwidth choice for density estimation. *J. Statist. Plann. Inf.* **23**, 53–69.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

STRANG, G. (1989). Wavelets and dilation equations: a brief introduction. *SIAM Review* **31**, 614–627.

STRANG, G. (1993). Wavelet transforms versus fourier transforms. *Bulletin (N.S.) Amer. Math. Soc.* **28**, 288–305.

WAND, M.P., MARRON, J.S. AND RUPPERT, D. (1991). Transformations in density estimation (with discussion). *J. Amer. Statist. Assoc.* **86**, 343–361.

Figure 1(a): $N(0,0.04)$, $n=1000$

Figure 1(b): Bandwidth, Normal(0,0.04),n=1000

Figure 2(a): Bimodal normal mixture, $n=1000$
 $N(0,0.25)$, $N(1.5,0.01)$

Figure 2(b): Bandwidth, Bimodal normal mixture, n=1000

Figure 2(c): Bimodal normal mixture, $n=2000$
 $N(0,0.25)$, $N(1.5,0.01)$

Figure 2(d): Bandwidth, Bimodal normal mixture, $n=2000$

Figure 3(a): Trimodal normal mixture, $n=1000$
 $N(-0.75, 0.25)$, $N(0.5, 0.0025)$, $N(1.25, 0.09)$

Figure 3(b): Trimodal normal mixture, $n=1000$

Figure 3(c): Trimodal normal mixture, $n=2000$
 $N(-0.75, 0.25)$, $N(0, 0.0025)$, $N(0.75, 0.09)$

Figure 3(d): Bandwidth, Trimodal normal mixture, $n=2000$

Figure 4(a): 4 modes normal mixture, $n=1000$
 $N(-1.25, 0.04), N(0, 0.0025), N(0.75, 0.09), N(1.5, 0.01)$

Figure 4(b): Bandwidth, 4 modes normal mixture, $n=1000$