Bernoulli **7**(2), 2001, 317–341

Cross-validation for choosing resolution level for nonlinear wavelet curve estimators

PETER HALL[1] and SPIRIDON PENEV[1,2]

[1]*Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia. E-mail: peter.hall@anu.edu.au*
[2]*Department of Statistics, University of New South Wales, NSW 2052, Australia. E-mail: spiro@maths.unsw.edu.au*

We show that unless the target density is particularly smooth, cross-validation applied directly to nonlinear wavelet estimators produces an empirical value of primary resolution which fails, by an order of magnitude, to give asymptotic optimality. We note, too, that in the same setting, but for different reasons, cross-validation of the linear component of a wavelet estimator fails to give asymptotic optimality, if the primary resolution level that it suggests is applied to the nonlinear form of the estimator. We propose an alternative technique, based on multiple cross-validation of the linear component. Our method involves dividing the region of interest into a number of subregions, choosing a resolution level by cross-validation of the linear part of the estimator in each subregion, and taking the final empirically chosen level to be the minimum of the subregion values. This approach exploits the relative resistance of wavelet methods to over-smoothing: the final resolution level is too small in some parts of the main region, but that has a relatively minor effect on performance of the final estimator. The fact that we use the same resolution level throughout the region, rather than a different level in each subregion, means that we do not need to splice together different estimates and remove artificial jumps where the subregions abut.

Keywords: curve estimation; density estimation; generalized kernel methods; kernel estimator; least-squares cross-validation; linear wavelet estimator; nonparametric regression; primary resolution level; thresholding

1. Introduction

1.1. Smoothing-parameter role of the resolution level

Wavelet estimators are well known for their ability to capture irregular features of a curve, and for their performance in distinguishing between stochastic aberrations and deterministic fluctuations; see, for example, Donoho and Johnstone (1994; 1995) and Donoho *et al*. (1995; 1996). However, they generally do less well at estimating smooth parts of a curve, for example in regression problems when the signal to noise ratio is low.

A common cause of this problem is that elementary approaches to wavelet estimation take the primary resolution level equal to 1, which produces over-smoothed estimates in places where the target curve is smooth. This is reflected in the relatively large bias that such estimates exhibit in peaks and troughs. The high bias also means that the estimates

1350–7265 © 2001 ISI/BS

suffer from relatively prominent 'Gibbs phenomenon' wiggles in places where the true curve changes sharply, for example at jump discontinuities.

One way of avoiding these difficulties is to choose the primary resolution level empirically, taking account of the fact that it is a smoothing parameter. From several viewpoints the primary resolution level plays the role of bandwidth – or more concisely, the inverse of bandwidth – in the 'linear' part of a thresholded wavelet estimator, and so correct choice of this quantity can alleviate difficulties caused by over- or under-smoothing in simpler approaches to wavelet-based estimation.

1.2. Cross-validating linear estimators

Substantial progress in choosing the primary resolution level has been made by Tribouley (1995), who suggested a cross-validation algorithm for choosing the level for linear wavelet estimators. When applied to nonlinear estimators, however, Tribouley's approach can produce substantial under-smoothing. It is in the nonlinear case that the majority of interest lies; linear wavelet estimators generally perform no better, and sometimes worse (when rough or asymmetric wavelet functions are used), than their kernel counterparts. Tribouley (1995) uses the nonlinear part of the estimator only as a diagnostic.

To appreciate why problems can arise if one applies an algorithm developed for the linear case to a nonlinear estimator, let us consider wavelet estimation of one of the target densities considered by Tribouley (1995), the double exponential density, as a case in point. The derivative of this density has a discontinuity at the origin, and we argue in Section 5 that this forces an estimator of primary resolution level based on a linear wavelet estimator to be of size $n^{1/4}$, where n denotes sample size. By way of contrast, the optimal resolution level, in terms of minimizing integrated risk, is $n^{1/(2r+1)}$, where $r \geqslant 2$ denotes the order of the wavelet. Since $n^{1/4}$ is an order of magnitude larger than $n^{1/(2r+1)}$, the cross-validation algorithm under-smooths by an order of magnitude. (Recall that resolution level is like the inverse of bandwidth.) This can result in an erratic curve estimate.

Wavelet estimators are forgiving of over-smoothing. That is why the simple estimator employed in the WaveThresh program (Nason and Silverman 1994), which takes the primary resolution level equal to 1, performs relatively well; in asymptotic terms, the penalty for even this very large degree of over-smoothing is only a logarithmic function of sample size. However, wavelet estimators are no more forgiving of under-smoothing than are conventional kernel estimators, and so choosing the resolution level too large, by an order of magnitude, can have serious consequences. Therefore, the severe under-smoothing noted in the previous paragraph can be a significant problem.

1.3. Direct cross-validation of nonlinear estimators

On the other hand, applying cross-validation directly to the nonlinear form of a wavelet estimator is unsatisfactory, for at least two reasons. First, the hard-thresholded version of a nonlinear estimator is a discontinuous function of the primary resolution level. That difficulty, combined with the usual stochastic fluctuations of the cross-validation criterion, makes it very

difficult to select the resolution level. In principle these problems can be alleviated by passing to a form of soft thresholding, but in practice the erratic fluctuations persist.

More serious, at least for large samples, is the fact that in the nonlinear case the cross-validation criterion does not accurately approximate the integrated squared error (ISE) of the wavelet estimator. The rationale behind cross-validation is that the criterion should equal ISE up to terms that either do not depend on the resolution level or are negligibly small relative to ISE. However, we show in Section 5.6 (see particularly Theorem 5.4) that in the nonlinear case, and in the case of target densities that are only piecewise smooth, the criterion includes stochastic terms that are of larger order than ISE yet depend non-negligibly on the primary resolution level. Therefore, choosing the resolution level so as to minimize the cross-validation criterion does not produce asymptotic minimization of ISE. Instead, it minimizes a measure of stochastic error that has a different order from, and has no close connection to, ISE.

1.4. Multiple linear cross-validation for smoothing nonlinear estimators

In the present paper we suggest a means of overcoming the problems discussed in Sections 1.2 and 1.3. We introduce a 'multiple cross-validation' algorithm for linear wavelet estimators, enabling us to estimate the resolution level that is appropriate for nonlinear estimators. Our technique involves dividing the region of estimation into subregions, applying cross-validation of the linear estimator to estimate the resolution level for each subregion, and taking the minimum of these values as the resolution level for the entire curve. (On some occasions a larger value than the minimum might be used, but typically the minimum is an appropriate choice.) We argue that this is more satisfactory than employing a different resolution level over each of the subregions, since that approach requires curve estimates to be spliced together where the subregions join, in order to remove artificial jump discontinuities. We show, both numerically and theoretically, that the high-degree resistance of wavelet methods to over-smoothing confers good performance in places where a larger resolution level might otherwise have been used.

We also point out, in the context of adaptive choice of primary resolution level, that it makes a lot of sense to eschew the usual dyadic definition of resolution level, and instead choose it in the continuum. The issue of choice of smoothing parameter is really only important, at least for wavelet estimators, when signal to noise ratio is low; and in such cases a large body of statistical experience with kernel methods argues strongly against restricting oneself to dyadic bandwidth choice. So it does too for choice of resolution level for wavelet estimators.

Nason (1996) suggested using cross-validation to select the threshold parameter for wavelet shrinkage, and Hurvich and Tsai (1998) proposed a threshold selector based on cross-validation and Akaike's information criterion. Both contributions focused on a mean-squared-error view of fidelity. It should be noted that in conventional asymptotic terms, where the data distribution is kept fixed as sample size increases, and the target function is piecewise smooth, the threshold has only a high-order effect on mean squared error; see, for example, Hall and Patil (1996). This property is also readily seen numerically. Thus, the

threshold would not normally be regarded as a smoothing parameter. By way of contrast, varying the resolution level can have a substantial, first-order impact on both mean squared error and qualitative numerical properties of the curve estimator.

The main features of our algorithm, and its relationship to other methods, are given in the next section. Details of the algorithm, and theoretical properties, are presented for density estimation in Sections 3 and 5, respectively. In Section 5.6 we show that cross-validation for choosing the primary resolution level, when applied directly to nonlinear wavelet estimators, does not produce asymptotic optimality in squared-error terms. Numerical properties, for both density estimation and regression, are summarized in Section 4.

2. Summary of methodology

Our algorithm is as follows.

1. Construct a wavelet pilot estimator of the true curve. The approaches suggested by Donoho and Johnstone (1994; 1995) and Donoho *et al*. (1995), essentially the same as that incorporated into the WaveThresh package, are possible choices for this step since they do not require explicit selection of threshold or resolution level. However, the nonlinear estimator whose resolution level is computed using the linear method of Tribouley (1995) is often a better choice for the pilot. The fact that it tends to be under-smoothed means that it gives a good indication of places where the target density is rough.
2. Visually divide the region $\mathscr{R}$, where the final estimator is required, into disjoint subregions $\mathscr{S}_1, \ldots, \mathscr{S}_m$, such that $\mathscr{R} = \bigcup_\ell \mathscr{S}_\ell$. Each subregion is an interval or a union of intervals, where the pilot estimator has relatively homogeneous roughness.
3. Using a relatively standard cross-validation algorithm, but employing only the linear part of the wavelet curve estimator, make an empirical choice of resolution level for each of the subregions. Take the minimum of these levels as the 'final' estimator of the resolution level.
4. Substitute this resolution level estimate into the nonlinear, thresholded wavelet estimator, and apply it across the full region $\mathscr{R}$.

To elaborate on the kind of subdivision that we have in mind in step 2, let us suppose the pilot estimator suggests that the target curve exhibits several clearly defined 'bumps', as well as an interval where the curve behaves particularly erratically. Then we could put the latter interval into one of the subregions $\mathscr{S}_\ell$, put each of the high-curvature 'tops' of the bumps into a separate subregion, and put the relatively low-curvature 'sides' of the bumps into further subregions. Alternatively, we could put the union of the high-curvature parts of bumps into one subregion, and the union of the low-curvature sides into another, in which case there would be just three subregions in all (including the interval where behaviour is particularly erratic). Note particularly that we do not necessarily put the entire bump into the same subregion; this avoids problems with under-smoothing, caused by curves where the tops of bumps are relatively sharp.

The algorithm is surprisingly robust against misspecification of different subregions $\mathscr{S}_\ell$. We shall explain why in Section 3, noting there that in most cases it is necessary only to specify a subregion where f is smooth; subregions where it is rough take care of themselves.

When determining resolution level in step 3 of the algorithm we have a choice of following 'standard' practice, where the level is taken to be an integer power of 2, or allowing resolution level to vary in the continuum. The former approach allows use of Mallat's pyramid algorithm for computation, but can produce significant deterioration in performance in mean-squared-error terms; see Hall and Nason (1997). It is analogous to using a kernel smoother with its bandwidth restricted to being an integer power of $\frac{1}{2}$. We adopt the continuum approach, and in Section 4 give a numerical illustration of its superior performance (although at the expense of greater computational labour).

The good performance of estimators computed using our algorithm is underpinned by the following three properties, each of which will be verified theoretically in Section 5.1.

(A) In places where the true curve is smooth, the linear form of the wavelet estimator is virtually equivalent to its nonlinear, thresholded form, in the sense that the thresholding operation is essentially degenerate. Therefore, since the focus of our resolution-level choice method is effectively on places where the true curve is smooth (see (C) below), then almost nothing is lost through confining attention to the linear part of the estimator in step 3 of the algorithm.

(B) The 'optimal' resolution level, which acts like the inverse of the bandwidth of a conventional kernel estimator (see, for example, Hall and Patil 1995a; 1995b; 1996), is smaller for smoother parts of the curve. Therefore, by using the smallest of the resolution levels determined by cross-validation, as suggested in step 3 of the algorithm, we are in effect focusing on the smoothest parts of the curve.

(C) In places where the true curve is rough, choice of primary resolution level for the nonlinear wavelet estimator has only a minor impact on performance. This is because nonlinear terms capture relatively erratic fluctuations of the true curve, and their ability to do this is largely unaffected by the primary resolution level. This provides support for our decision to focus on relatively smooth parts of the curve.

Property (B) also indicates a potential shortcoming of our algorithm: if the smoothest part of the curve is virtually flat then we might wish to exclude that part from calculation of the minimum resolution level, since a level that represents a compromise between 'smoothest' and 'roughest' places is generally preferable in such extreme settings. Of course, our technique of identifying smooth and rough parts of the curve through a pilot estimator provides the opportunity to identify such cases.

A fourth property argues that we should not, in place of step 3, apply cross-validation to the whole region where we wish to estimate the curve, unless the curve is particularly smooth and homogeneous there (in which case we may not be interested in using wavelet methods at all). For simplicity we discuss this property below in the case where roughness comes about through a jump discontinuity; but, more generally, other forms of roughness cause the same problem. Theoretical justification for the property is given in Section 5.1.

(D) If a low-order derivative of the true curve has a jump discontinuity at a point in the interior of an interval $\mathscr{I}$, and if we choose the resolution level by applying cross-validation to the linear form of the wavelet estimator on a region that includes $\mathscr{I}$, then the empirical level that we obtain will be too large by an order of magnitude, relative to that which gives good mean-square performance.

It is for this reason that our approach differs from the main method suggested by Tribouley (1995), which, we argue in Sections 4 and 5, does not necessarily produce nonlinear estimators that have good mean-squared-error performance or give qualitatively pleasing results. It should be stressed, however, that Tribouley developed her methods for the linear case.

Tribouley (1995) also proposed a second method for linear estimators. This involves: (a) dividing $\mathscr{R}$ into subregions $\mathscr{S}_\ell$ that are each fairly homogeneous in terms of their smoothness; (b) using separate cross-validations on those subregions; (c) constructing the respective curve estimates over the sets $\mathscr{S}_\ell$; and finally, (d) putting these estimates together, side by side, to form the final linear estimate. While this method is related to our own, we argue that ours has advantages even in the linear case. In particular, the estimate produced by steps (a)–(d) has jump discontinuities in places where subregions join, which need to be removed using a separate 'splicing' algorithm.

Secondly, putting the 'joins' at valleys between separate modes or bumps, as suggested by Tribouley (1995) for her second method, can be inadequate if there is a discontinuity in the derivative of the curve at the mode. In this case the final wavelet estimator will again under-smooth on the sides of bumps, since it focuses too hard on getting the tops of bumps right. In our experience the pilot wavelet estimator is often not capable of distinguishing between a sharp peak on the one hand, and a smooth peak with stochastic fluctuations near the mode on the other; hence our suggestion earlier that the 'tops' and 'sides' of bumps be treated as different subregions $\mathscr{S}_\ell$. However, treating the tops and sides separately, and using Tribouley's second method, will lead to a large number of jump discontinuities at joins in the final estimate; all of these need splicing.

Thirdly, our experience with local smoothing in the kernel case suggests that attempting local smoothing of wavelet estimators, for example by using a different primary resolution level for each of two different bumps in a smooth, bimodal density, is often fruitless if the resolution level is constrained to be a power of 2. For example, if the bumps are of approximately the same height then the width of one has to be about twice that of the other before different bandwidths on a dyadic scale (as employed by Tribouley, 1995) are going to be capable of applying effective adaptive smoothing. We shall give an example in Section 4, employing the same bimodal density as Tribouley (1995), and showing the marked improvement that can be achieved by choosing resolution levels in the continuum.

3. Details of methodology

For the sake of brevity we give details only in the case of density estimation; that of regression is similar. Let the wavelet basis be the sequence of functions represented by

$\phi_k(x) = p^{1/2}\phi(px - k)$ and $\psi_{ik}(x) = p_i^{1/2}\psi(p_i x - k)$, where $p > 0$ denotes the primary resolution level, $p_i = p\,2^i$, and the scaling function ϕ and wavelet ψ are assumed to be compactly supported. We assume, too, that the wavelet is of order r, in the sense that for a largest integer $r \geqslant 1$ and a constant $\kappa \neq 0$,

$$\int x^s \psi(x)\,\mathrm{d}x = \begin{cases} 0 & \text{if } 0 \leqslant s \leqslant r-1, \\ r!\kappa & \text{if } s = r. \end{cases} \tag{3.1}$$

Then the wavelet expansion of a density f is given by

$$f = \sum_{-\infty < k < \infty} b_k\,\phi_k + \sum_{i=0}^{\infty} \sum_{-\infty < k < \infty} b_{ik}\,\psi_{ik},$$

where $b_k = \int \phi_k f$ and $b_{ik} = \int \psi_{ik} f$. Given data $X_1, \ldots, X_n$ from the distribution with density f, estimators of b_k and b_{ik} are respectively $\hat{b}_k = n^{-1}\sum_j \phi_k(X_j)$ and $\hat{b}_{ik} = n^{-1}\sum_j \psi_{ik}(X_j)$, giving rise to the thresholded wavelet estimator,

$$\hat{f}(x|p) = \sum_{-\infty < k < \infty} \hat{b}_k\,\phi_k(x) + \sum_{i=0}^{q-1} \sum_{-\infty < k < \infty} \hat{b}_{ik}\,I(|\hat{b}_{ik}| > \delta_i)\,\psi_{ik}(x), \tag{3.2}$$

where δ_i denotes the threshold at resolution level i.

There are various ways of selecting δ_i, of which the two most popular are the 'constant' threshold $\delta_i = C\,(n^{-1}\log n)^{1/2}$ (not depending on i) and the 'level-dependent' threshold $\delta_i = C(i/n)^{1/2}$, where in each case C is a constant depending only on f (or empirically, on a pilot estimator of f). The constant threshold case is widely used, and there we may take $C \geqslant (2 \sup f)^{1/2}$, where the supremum is over the region $\mathscr{R}$ where we wish to estimate f. See, for example, Donoho *et al.* (1995). Level-dependent thresholding has been treated in the context of density estimation by Delyon and Juditsky (1996) and Donoho *et al.* (1996), for example.

In our numerical work in the case of density estimation we shall explore both the constant and level-dependent thresholding approaches. In the case of regression, and in our theoretical arguments, we shall confine attention to constant thresholding. Using that approach, but taking $p \equiv 1$ in the definition of $\hat{f}(x|p)$ in (3.2), we obtain the pilot estimator mentioned in step 1 of the algorithm in Section 2.

The linear part of $\hat{f}(\cdot|p)$ is the estimator

$$\hat{f}_{\mathrm{lin}}(x|p) = \sum_{-\infty < k < \infty} \hat{b}_k\,\phi_k(x) = (nh)^{-1}\sum_{j=1}^{n} K(x/h,\, X_j/h),$$

where

$$K(x,\, y) = \sum_{-\infty < k < \infty} \phi(x + k)\phi(y + k) \tag{3.3}$$

and $h = p^{-1}$. Thus, the linear estimator is a generalized kernel estimator with bandwidth equal to the inverse of the resolution level; see, for example, Hall and Patil (1995a). Let

$\hat{f}_{\mathrm{lin},-j}(\cdot|p)$ denote the version of $\hat{f}_{\mathrm{lin}}(\cdot|p)$ in which the sample size is reduced to $n-1$ by deleting observation X_j.

Let the subregions to which we apply cross-validation be $\mathscr{S}_1, \ldots, \mathscr{S}_m$. The cross-validation criterion on $\mathscr{S}_\ell$, as a function of the resolution level p and computed for the linear estimator $\hat{f}_{\mathrm{lin}}(\cdot|p)$, is

$$\mathrm{CV}_\ell(p) = \int_{\mathscr{S}_\ell} \hat{f}_{\mathrm{lin}}(x|p)^2 \, \mathrm{d}x - 2n^{-1} \sum_{j=1}^{n} \hat{f}_{\mathrm{lin},-j}(X_j|p) I(X_j \in \mathscr{S}_\ell). \tag{3.4}$$

This is borrowed from Hall and Schucany (1989), where it was used for local bandwidth selection in the context of conventional curve estimators. See Mielniczuk *et al*. (1989) for a closely related criterion.

Let $\hat{p}_\ell$ denote the value of p that minimizes $\mathrm{CV}_\ell(p)$. Following the algorithm given in Section 2 we do the minimization in the continuum, not just for the dyadic sequence 2^k for $k \geqslant 1$. Put

$$\hat{p} = \min_{1 \leqslant \ell \leqslant m} \hat{p}_\ell. \tag{3.5}$$

We take $\hat{p}$ as our estimator of primary resolution level for constructing the final estimator, which is $\tilde{f} = \hat{f}(\cdot|\hat{p})$.

Tribouley (1995) suggests calculating CV_ℓ, and in particular the series on the right-hand side of (3.4), using Parseval's identity. However, it may be shown that, since only a subset of the support of f is involved, this leads to edge effects which can reduce performance of the final estimator. In particular, the cross-validation function produced by Parseval's identity is different from that in (3.4).

The fact that the estimator of primary resolution level is taken as the minimum of values for different subregions makes our algorithm inherently robust against misspecification of the regions $\mathscr{S}_\ell$ in step 2. Indeed, provided that only one of the subregions is an interval over which the target function is relatively smooth, the resolution level defined in (3.5) will be appropriate for estimating f in places where it is erratic or rough. The relative resistance of wavelet methods to over-smoothing comes to our aid here: the final resolution level may be too small in some parts of $\mathscr{R}$, but that nevertheless has only a minor effect on performance of the final estimator.

We shall show in Section 5 that if f is piecewise smooth on $\mathscr{R}$, and if the $\mathscr{S}_\ell$ are such that none of their endpoints is a point of jump discontinuity of f or of one of its first $r-1$ derivatives, then our algorithm will produce an estimator of primary resolution level of the correct order. Of course, it is possible to consistently estimate all the jump discontinuities of f and all its derivatives, for example using wavelet methods; see, for example, Wang (1998). Therefore, an automatic procedure for selecting $\mathscr{S}_1, \ldots, \mathscr{S}_m$, giving minimal order of ISE, is readily constructed in this case. Of course, f would not be as simple as a function whose only irregularities were jump discontinuities, but nevertheless the robustness properties noted in the previous paragraph ensure that selection of appropriate subregions $\mathscr{S}_\ell$ is not a critical task; in most cases it is necessary only to find one subregion where f is relatively smooth, and then the subregions where f is rough will be accommodated.

4. Numerical results

4.1. First problem: bimodal densities

Our first example is of the case where f is the density of a normal mixture,

$$\tfrac{1}{2}N(0,\ 1) + \tfrac{1}{2}N(3,\ 0.3^2). \tag{4.1}$$

The same density was treated by Tribouley (1995). Figure 4.1 shows typical graphs of the two abutting linear estimates produced by Tribouley's approach (dotted line), and the nonlinear estimate obtained using our method (long-dashed line). The ISE values are respectively 1.73×10^{-2} and 1.24×10^{-2}. To enable close comparison with Tribouley's method, we used only two subregions to construct both estimates; they were the subregions employed by Tribouley (1995), i.e. $[-3, 2]$ and $[2, 5]$. Primary resolution level was chosen in the continuum for our method, and dyadically for the method of Tribouley. (However, for her approach, here and below, we used the cross-validation formula (3.4) directly instead of the

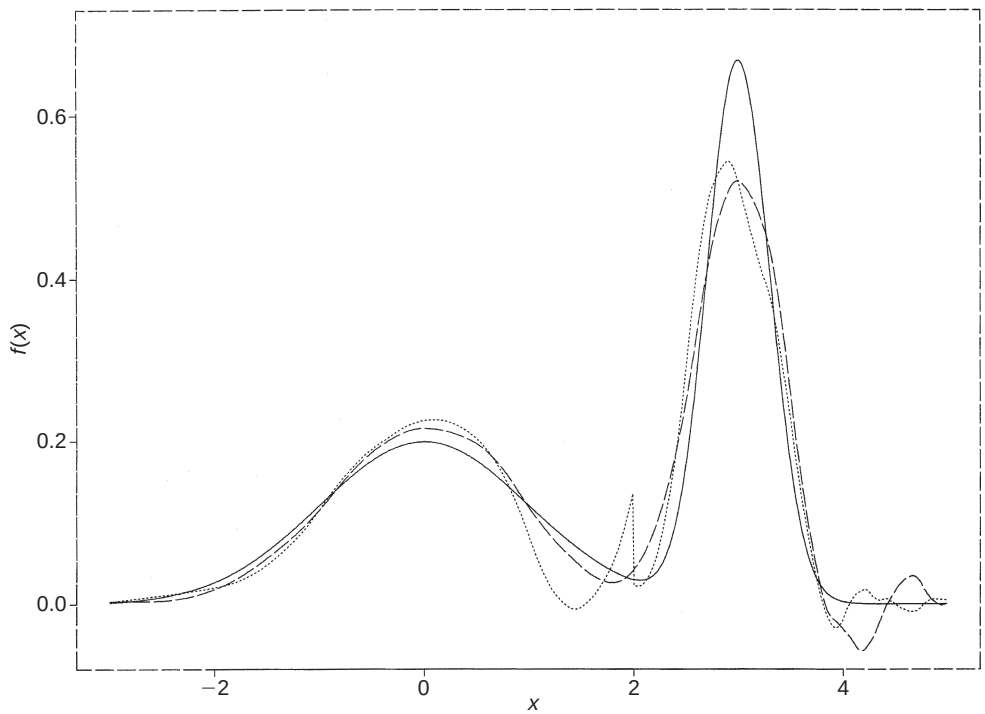

Figure 4.1. First bimodal example. The solid line depicts the true density of the normal mixture in (4.1), the dotted line is the estimate suggested by Tribouley's (1995) second method, and the long-dashed line is the estimate proposed in section 2, although using only two subregions.

approximation based on Parseval's identity; this removed problems arising from inexactness at edges of subregions.) Sample size was $n = 200$. We employed constant thresholding for both estimators. The wavelet used in this example was from the Daubechies family with $r = 5$. Filter coefficients for this wavelet are given by Daubechies (1992, p. 195). Given its support width, the wavelet has extremal phase and the greatest number of vanishing moments. Similar results are obtained with different wavelets, but we rely on our theoretical account in Section 5, rather than repeated simulation, to affirm the validity of this generalization.

The main disadvantage of Tribouley's method in this setting is the visually displeasing 'join' between estimates of the separate bumps. That problem can be overcome by constructing the linear or nonlinear estimate over the whole real line, using the primary resolution level derived for the linear estimate. However, that can lead to serious under-smoothing problems, particularly if the second bump of the density is sharper than that for the distribution at (4.1).

For example, consider the normal mixture

$$\tfrac{1}{2}N(0,\, 0.5^2) + \tfrac{1}{2}N(1.5,\, 0.1^2). \tag{4.2}$$

To construct a density estimate using multiple cross-validation we employed four subregions, as suggested in Section 2; these were $(-\infty, -0.4]$, $[-0.4, 0.6]$, $[0.6, 1.2]$ and $[1.2, \infty)$. Typical results are illustrated in Figure 4.2. Figure 4.2(a) shows the nonlinear wavelet estimate based on the multiple cross-validation estimate of primary resolution level; and, for comparison, Figure 4.2(b) shows the linear and nonlinear wavelet estimates when resolution level is chosen by standard cross-validation on the whole real line. For all three estimates depicted in Figure 4.2 we chose resolution level in the continuum, and took $n = 800$.

The influence of the narrower second bump is clearly evident from Figure 4.2; it leads to significant under-smoothing of the first bump, and hence to excessive stochastic fluctuations, unless the multiple cross-validation approach is used. Depending on sample size, as well as sample, these fluctuations can be either more or less pronounced if resolution level is chosen dyadically rather than in the continuum. Figure 4.3 shows the effect of using dyadically chosen primary resolution level in the nonlinear estimate in Figure 4.2(a); random fluctuations have now become even more of a problem. In the realization illustrated in Figure 4.3, continuum choice of primary resolution level suggests $p = 2.43$ (giving rise to the estimator depicted by the long-dashed line), whereas dyadic choice requires $p = 4$ (indicated by the dotted line). Using the latter results in ISE being inflated by a factor of 1.55.

4.2. Second problem: double-exponential density

This example, where the derivative has a jump discontinuity at the origin, was investigated by Tribouley (1995) using a linear wavelet estimator and conventional cross-validation. Theoretical arguments given in Section 5 show that better ISE performance can be achieved using a nonlinear estimator and multiple cross-validation. Figure 4.4 illustrates this point numerically. There, the values of ISE, averaged over 500 samples of size $n = 800$, are

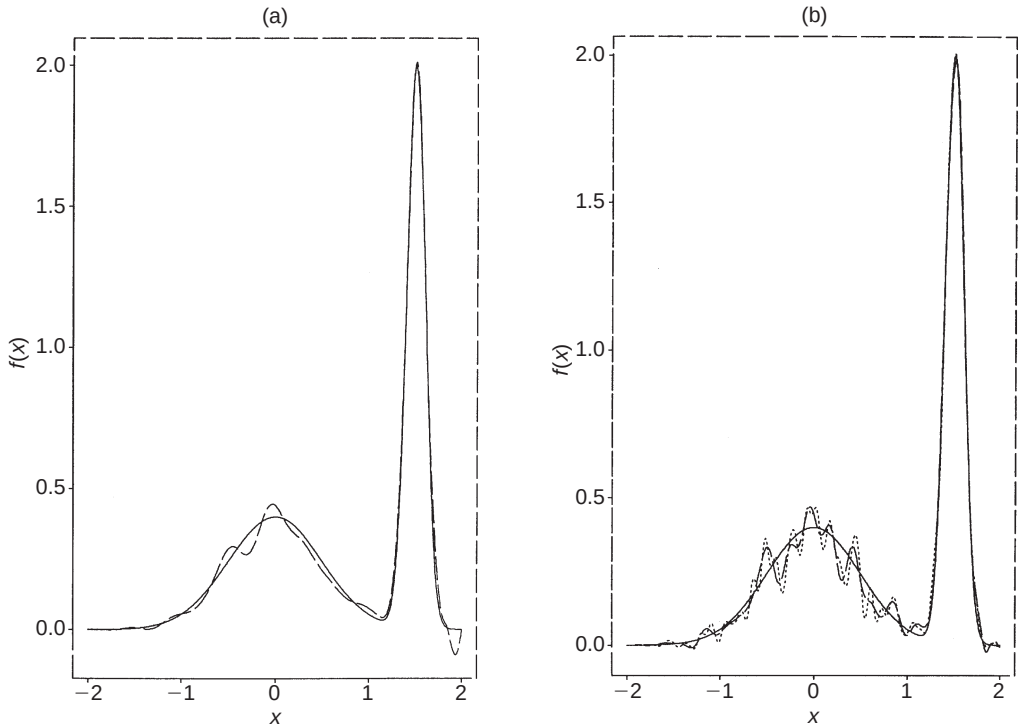

Figure 4.2. Second bimodal example. (a) The true density of the normal mixture at (4.2) (solid line) and the nonlinear estimate (long-dashed line), with resolution level computed using multiple cross-validation. (b) The true density, the linear wavelet estimate with primary resolution level computed using standard cross-validation (dotted line), and the nonlinear wavelet estimate using the same resolution level (short-dashed line).

graphed as a function of the choice of the subregions used for multiple cross-validation. In this section alone, the wavelet used was from the Daubechies family with $r = 3$.

We employed two subregions, $[-\xi, \xi]$ and $(-\infty, -\xi] \cup [\xi, \infty)$, and so the figure plots the average value of the ISE as a function of ξ. Virtually identical results are obtained for three subregions, $[-\xi, \xi]$, $(-\infty, -\xi]$ and $[\xi, \infty)$. The sampled density was $f(x) \equiv \mathrm{e}^{-2|x|}$, and resolution level was chosen in the continuum. The optimal value of ξ is about 0.7. As $\xi \to \infty$, the value of the average ISE is asymptotic to that which would be obtained using Tribouley's method, although for continuous choice of resolution level. (The average ISE for dyadic choice of resolution level is greater.) Graphs of nonlinear estimates obtained using our multiple cross-validation method with $\xi \approx 0.5$ generally have fewer spurious bumps than linear estimates constructed using standard cross-validation.

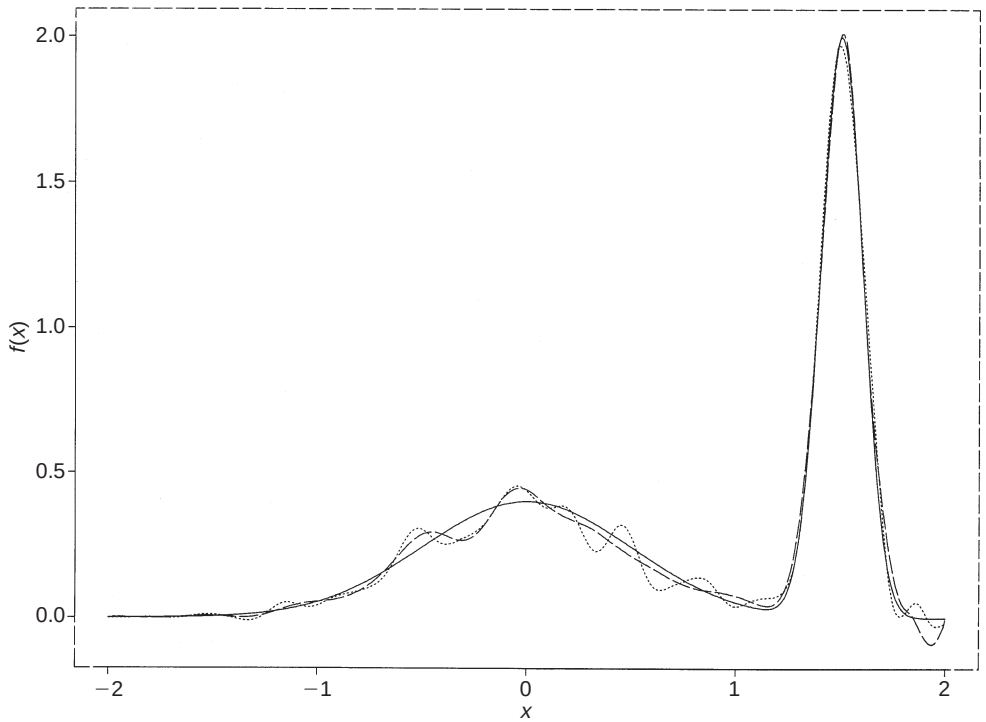

Figure 4.3. Effect of using dyadic resolution level. The solid line depicts the true density of the normal mixture in (4.2), the long-dashed line shows the nonlinear estimate with resolution level chosen in the continuum, using multiple cross-validation, and the dotted line depicts the nonlinear estimate with dyadic resolution level chosen by multiple cross-validation. The sample was the same as that for Figure 4.2.

4.3. Third problem: comb density

Here we consider the 'comb density' of Marron and Wand (1992), being the density of the five-component normal mixture,

$$\sum_{i=0}^{5}(2^{5-i}/63)N\{(65-96\times 0.5^{i})/21,\,(32/63)^{2}/2^{2i}\}.$$

It is depicted by the solid line in Figure 4.5.

The figure also shows typical results of first applying standard cross-validation to the linear wavelet estimator, and then using the resolution level derived in that way to construct either the linear estimate (dotted line in Figure 4.5) or the nonlinear estimate (long-dashed line). Both estimates are highly susceptible to stochastic fluctuations among data in the support of the lowest-frequency bump of the comb density, and for this reason neither is

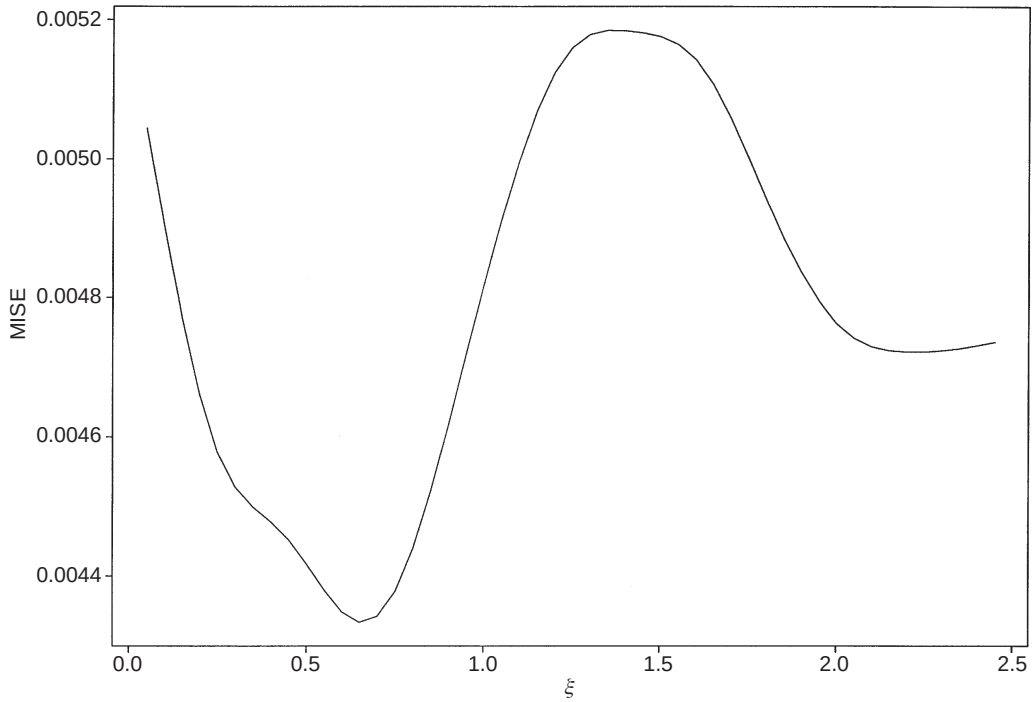

Figure 4.4. Effect of choice of subregions: plot of average value of ISE as a function of ξ. The true density was $e^{-2|x|}$, and the estimates were computed using multiple cross-validation for two subregions, of which one was $[-\xi, \xi]$.

satisfactory. However, by virtue of having been constructed using a primary resolution level that is heavily influenced by high-frequently parts of the curve, the linear estimate performs about as well as its nonlinear counterpart here. (For Figures 4.5 and 4.6, sample size was $n = 800$ and constant thresholding was used. Resolution level was chosen in the continuum.)

Next we divided the support of the density into three subregions, representing the supports of the lowest-frequency bump, the second lowest-frequency bump, and the rest of the density, respectively. We performed multiple cross-validation, taking the estimate of primary resolution level to be the minimum of the three cross-validated levels for the subregions, and used this level to construct first the linear estimate (dotted line in Figure 4.6) and then the nonlinear estimate (long-dashed line in Figure 4.6). The sample was the same as for Figure 4.5.

The linear estimator now performs much better in the first two low-frequency bumps of the comb density, since the primary resolution level has (in effect) been optimized for at least the first of these places; but it performs poorly in the high-frequency bumps. On the other hand, the nonlinear estimator now performs well across the entire range of

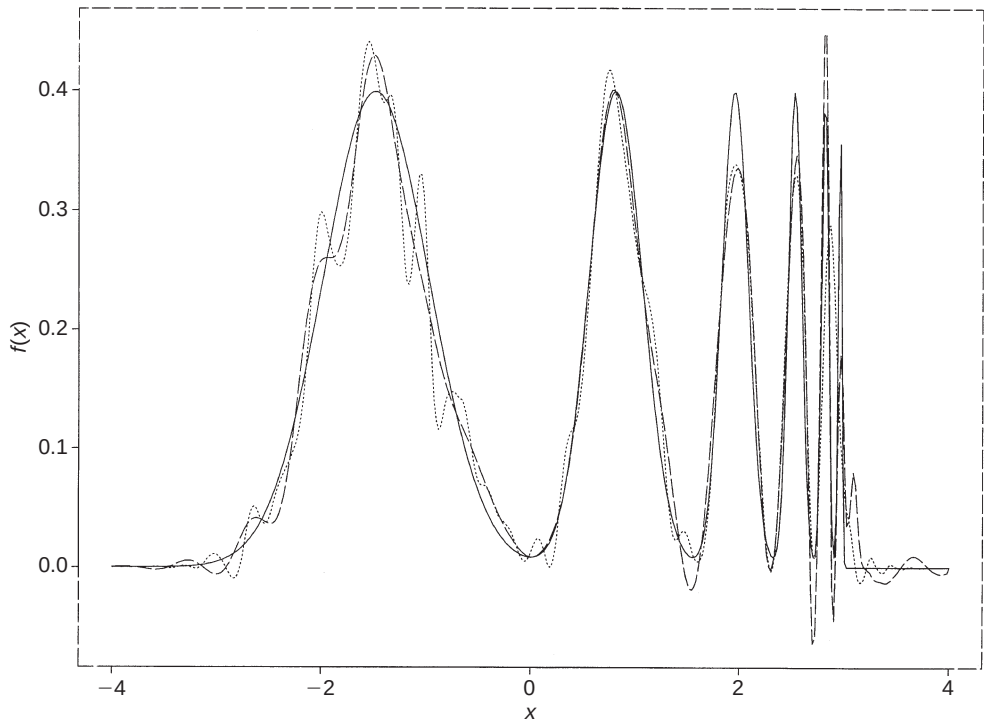

Figure 4.5. Comb density estimates using one-subregion cross-validation. Primary resolution level was calculated using standard cross-validation for the linear estimator on the whole real line. This level was then used to compute the linear estimate (dotted line) and nonlinear estimate (long-dashed line).

frequencies. Using a level-dependent threshold for the nonlinear estimate improves performance of the estimate of the third lowest-frequency bump, but introduces three small, spurious wiggles to the estimate of the lowest-frequency bump.

4.4. Fourth problem: nonparametric regression

We made observations of the pair (x_i, Y_i), for $1 \leqslant i \leqslant 1000$, where the x_i were equally spaced on $\mathscr{T} \equiv (-17\pi/30, 17\pi/30]$, $Y_i = f(x_i) + \varepsilon_i$, the ε_i were independent and identically distributed standard normal random variables, and the regression mean f was defined by

$$f(x) = \begin{cases} -\cos^2\{x + (\pi/15)\} & \text{if } x \in (-17\pi/30, 0], \\ \cos^2\{x - (\pi/15)\} & \text{if } x \in (0, 17\pi/30]. \end{cases}$$

Our aim was to estimate f over $\mathscr{T}$. The wavelet used was from the Daubechies family with $r = 5$.

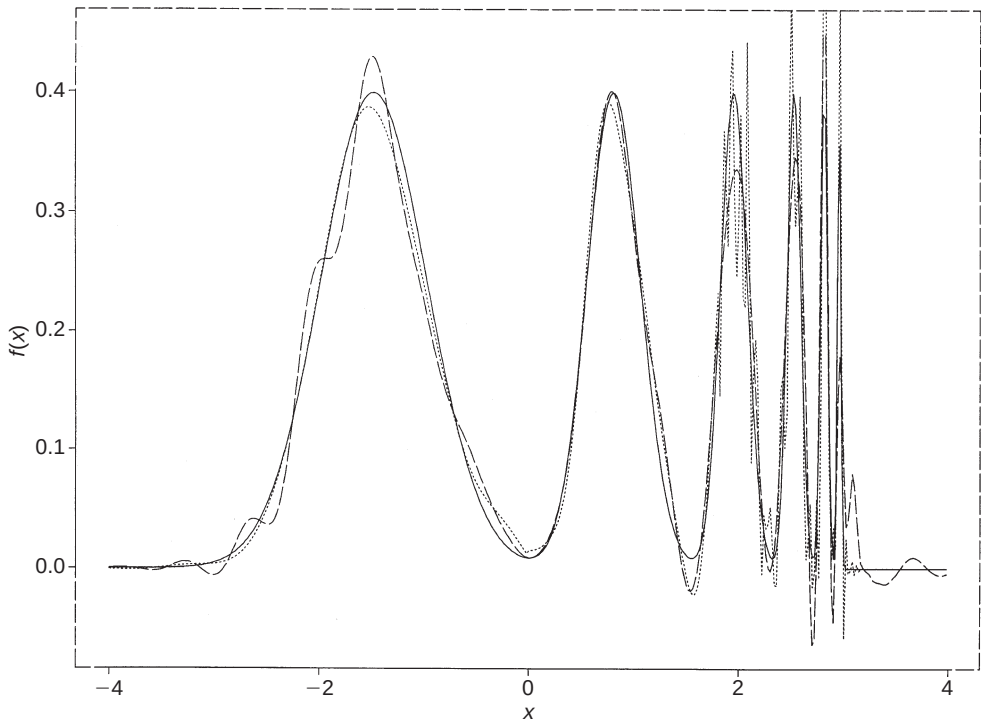

Figure 4.6. Comb density estimates using three-subregion cross-validation. Primary resolution level was calculated using multiple cross-validation of the linear estimator on three subregions. This level was applied to the linear estimate (dotted line) and nonlinear estimate (long-dashed line).

First we chose the primary resolution level, p, using conventional crossvalidation for a linear estimator. The results were satisfactory, but particularly in its right-hand half the estimated regression curve showed many spurious wiggles that were due to using too small a value of p.

Next we divided $\mathscr{T}$ into three subregions, $(-17\pi/30, -0.1]$, $(-0.1, 0.1]$ and $(0.1, 17\pi/30]$, and employed multiple cross-validation there, taking the overall estimate of primary resolution level to be the smallest of the three cross-validated values. A typical realization of the resulting nonlinear estimate is depicted in Figure 4.7. The long-dashed and dotted lines there show estimates for continuous and dyadic choice, respectively, of resolution level; they correspond to ISEs of 1.15×10^{-2} and 1.85×10^{-2}, respectively. This extent of reduction of mean squared error is typical of the quantitative advantages of choice of resolution level in the continuum.

The wiggles at the left-hand end of the curve estimate, and at the jump discontinuity, are caused by effects of Gibbs phenomenon type. They are less serious than those which arise when taking the primary resolution level equal to 1, as in the WaveThresh package, since

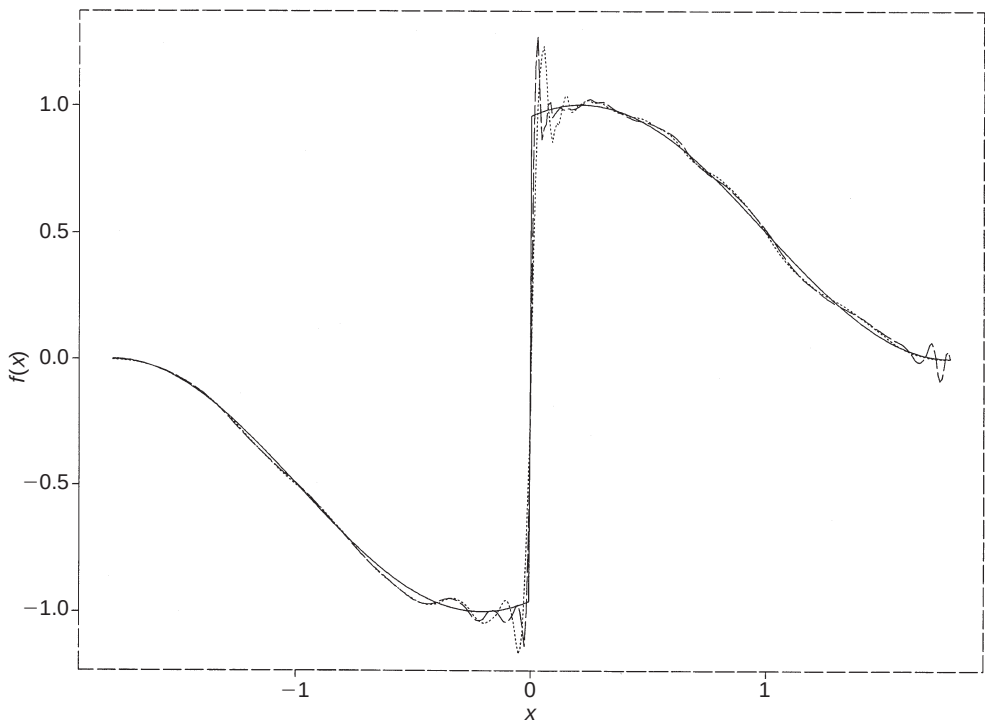

Figure 4.7. Regression mean with discontinuity. Both the graphed wavelet-based estimates are nonlinear. Primary resolution level was calculated using multiple cross-validation of the linear estimator on three subregions, using either dyadic choice (giving the estimate shown by the dotted line) or choice in the continuum (long-dashed line) of resolution level.

they are produced by excessive bias at discontinuities, and bias is less pronounced for our cross-validated estimates.

5. Theoretical properties

5.1. Properties (A)–(D) in Section 2

Our aim here is to provide theoretical support for the four main properties asserted in Section 2, which underpin our cross-validation algorithm. As in Section 3, we confine attention to the case of density estimation, although regression may be treated similarly. For the sake of simplicity we quantify 'roughness' of the target function in terms of the order of the derivative where the first discontinuities occur. Alternative approaches will be discussed in Section 5.5.

We begin by addressing the basic behaviour of the estimators $\hat{f}(\cdot|p)$ and $\hat{f}_{\text{lin}}(\cdot|p)$. Assume that $\mathscr{R}$ is a compact interval, and ϕ and ψ are compactly supported, Hölder continuous on the real line, and satisfy the rth-order condition (3.1) with $r \geqslant 1$ and $\kappa \neq 0$. Suppose, too, that the threshold in the definition of $\hat{f}(\cdot|p)$ is taken as $\delta = \delta_i = C(n^{-1}\log n)^{1/2}$, where $C \geqslant (2\sup f)^{1/2}$. Call these conditions (C_1). Note that they imply that the series at (3.3) has only a finite number of non-zero terms, that number being bounded uniformly in x and y, and that K is Hölder continuous in both variables. These properties make limit theory for the linear estimator $\hat{f}_{\text{lin}}(\cdot|p)$ relatively straightforward.

In our first result below we suppose that f has r continuous derivatives in an open interval $\mathscr{R}'$ containing $\mathscr{R}$; call this condition (C_2). In the second result we address the case of a relatively irregular function, for example where f has r continuous derivatives in $\mathscr{R}'$ in a piecewise sense, where for each $0 \leqslant s \leqslant r$, $f^{(s)}$ has only a finite number of points of discontinuity at each of which it has left- and right-hand limits. Call this condition (C_3). Define

$$A_1 = \int_{\mathscr{R}} f \quad \text{and} \quad A_2 = \kappa^2(1 - 2^{-2r})^{-1} \int_{\mathscr{R}} (f^{(r)})^2,$$

where κ is as at (3.1).

For smooth functions f the estimators $\hat{f}_{\text{lin}}(\cdot|p)$ and $\hat{f}(\cdot|p)$ enjoy essentially the same ISE expansions on $\mathscr{R}$, as our first theorem shows.

Theorem 5.1. *If* (C_1) *and* (C_2) *hold, and* $p \to \infty$ *and* $n/p \to \infty$, *then*

$$\int_{\mathscr{R}} \{\hat{f}_{\text{lin}}(x|p) - f(x)\}^2 \, dx = A_1 n^{-1} p + A_2 p^{-2r} + o_p(n^{-1}p + p^{-2r}). \tag{5.1}$$

If (C_1) *and* (C_3) *hold, and* $p \to \infty$, $q \to \infty$, $p_q \delta^2 \to 0$ *and* $p^{2r+1}\delta^2 \to \infty$, *where* $p_q = p2^q$, *then*

$$\int_{\mathscr{R}} \{\hat{f}(x|p) - f(x)\}^2 \, dx = A_1 n^{-1} p + A_2 p^{-2r} + o_p(n^{-1}p + p^{-2r}). \tag{5.2}$$

Results (5.1) and (5.2) are derived by Hall and Patil (1995a; 1995b), respectively.

The terms in $n^{-1}p$ and p^{-2r} on the right-hand sides of (5.1) and (5.2) represent the dominant contributions to error about the mean and to squared bias, respectively, in the ISE formulae. Of course, (5.1) and (5.2) are very similar to their counterparts in ISE and mean ISE expansions for standard kernel estimators; see, for example, formula (2.12) of Wand and Jones (1995, p. 21). It follows from (5.1) and (5.2) that in both cases the asymptotically optimal primary resolution level is given by

$$p_{\text{opt}} = B_1 n^{1/(2r+1)}, \tag{5.3}$$

where $B_1 = (2rA_2/A_1)^{1/(2r+1)}$.

Importantly, (5.1) fails if f or one of its first $r-1$ derivatives has a jump discontinuity in $\mathscr{R}$. To appreciate this point, suppose f has $t+1 \leqslant r$ piecewise continuous derivatives on an open interval $\mathscr{R}'$ containing $\mathscr{R}$, with left- and right-hand derivatives existing at each

point, and equal to one another at all but a finite number of points; $\mathscr{R}$ contains a point of non-degenerate jump discontinuity in $f^{(t)}$; and f has $t-1$ continuous derivatives in $\mathscr{R}$. Call this condition (C_4).

Theorem 5.2. *If* (C_1) *and* (C_4) *hold, and* $p \to \infty$ *and* $n/p \to \infty$, *then*

$$\int_{\mathscr{R}} \{\hat{f}_{\text{lin}}(x|p) - f(x)\}^2 \, dx = A_1 n^{-1} p + A_3 p^{-(2t+1)} + o_p(n^{-1}p + p^{-(2t+1)}), \tag{5.4}$$

where $A_3 > 0$ *depends on the locations, sizes and number of jump discontinuities of* $f^{(t)}$ *in* $\mathscr{R}$.

To derive (5.4), note that there is a term of size p^{-t} in the bias expansion of $\hat{f}_{\text{lin}}(\cdot|p)$ in a neighbourhood of radius $O(p^{-1})$ of each jump discontinuity of $f^{(t)}$ in $\mathscr{R}$. This gives rise to a term that is asymptotic to a constant multiple of $(p^{-t})^2 p^{-1} = p^{-(2t+1)}$ in the bias contribution to ISE. The total of these contributions, over all jump discontinuities of $f^{(t)}$ in $\mathscr{R}$, equals the contribution $A_3 p^{-(2t+1)} + o(p^{-(2t+1)})$ to the right-hand side of (5.4). It should be added to the right-hand side of (5.1), giving (5.4).

Result (5.4) shows that instead of (5.3), the asymptotically optimal primary resolution level is now given by

$$p_{\text{opt}} = B_2 n^{1/\{2(t+1)\}}, \tag{5.5}$$

where $B_2 = \{(2t+1)A_3/A_1\}^{1/\{2(t+1)\}}$.

By way of contrast, the nonlinear estimator $\hat{f}(\cdot|p)$ is able to adapt well to aberrations in the target function, and in consequence condition (C_2), under which (5.2) holds, permits jump discontinuities in any of the first r derivatives of f. These issues of comparative performance are addressed in more detail in Remark 2.6 of Hall and Patil (1995b), albeit in the context of a comparison of standard kernel estimators and $\hat{f}(\cdot|p)$. The case of generalized kernel estimators, such as $\hat{f}_{\text{lin}}(\cdot|p)$, is similar to that of standard kernel estimators.

Each of the key properties (A)–(D) stated in Section 2 is given theoretical support by the results noted above. In particular, property (A) is reflected in the fact that the expansions of $\hat{f}(\cdot|p)$ and $\hat{f}_{\text{lin}}(\cdot|p)$ are identical when f has r continuous derivatives; see (5.1) and (5.2). Result (5.5), and a comparison of (5.3) and (5.5), show that as the roughness of f decreases (i.e. as the value of t increases), the order of the optimal resolution level decreases, as asserted by property (B).

Such a comparison also shows that if f is rougher than the order of the chosen wavelet would suggest (i.e. if the order r of the wavelet is strictly larger than the number of continuous derivatives that f possesses on $\mathscr{R}$), and if, when constructing the nonlinear estimator $\hat{f}(\cdot|p)$, we use a resolution level that is dictated by optimal performance of the linear estimator $\hat{f}_{\text{lin}}(\cdot|p)$, then we can obtain a value of p that is too large by an order of magnitude – specifically, it will be of size $n^{1/\{2(t+1)\}}$, for $t \leqslant r-1$, whereas it should be of size only $n^{1/(2r+1)}$. This is a theoretical interpretation of property (D).

The fact that the ISE expansion at (5.2) does not depend on jump discontinuities of the first r derivatives of f (should those discontinuities exist), regardless of choice the

resolution level p, reflects the fact that choice of resolution level is relatively unimportant to the ability of the nonlinear wavelet estimator in capturing relatively rough features of the true curve. This provides theoretical justification for property (C). See also Section 5.4.

5.2. Theoretical performance of cross-validation

Let $\mathscr{R} = \bigcup_{1 \leqslant \ell \leqslant m} \mathscr{S}_\ell$ denote a subdivision of $\mathscr{R}$ into subregions $\mathscr{S}_\ell$, representing the subdivision arising in step 2 of the algorithm in Section 2, and assume $\mathscr{S}_\ell$ is a non-degenerate interval, that no points of discontinuity of f, $f^{(1)}, \ldots, f^{(r-1)}$ lie on the boundary of $\mathscr{S}_\ell$, and that $f^{(r)}$ does not vanish identically on $\mathscr{S}_\ell$, for $1 \leqslant \ell \leqslant m$. Call this condition ($C_5$). Let t_ℓ denote the smallest integer $t \in [0, r-1]$ such that $f^{(t)}$ has a jump discontinuity at a point in $\mathscr{S}_\ell$, and put $t_\ell = r$ if no such integer exists. If $t_\ell \leqslant r-1$, define

$$p_{\text{opt},\ell} = B_{2\ell} n^{1/\{2(t_\ell+1)\}},$$

where $B_{2\ell} = \{(2t_\ell + 1)A_{3\ell}/A_{1\ell}\}^{1/\{2(t_\ell+1)\}}$ and $A_{1\ell}$, $A_{3\ell}$ are the versions of A_1, A_3 in (5.4) that arise when $\mathscr{R}$ on the left-hand side of (5.4) is replaced by $\mathscr{S}_\ell$. (Thus, $p_{\text{opt},\ell}$ is the version of p_{opt}, at (5.5), when $\mathscr{S}_\ell$ replaces $\mathscr{R}$.) If $t_\ell = r$, put

$$p_{\text{opt},\ell} = B_{1\ell} n^{1/(2r+1)},$$

where $B_{1\ell} = (2rA_{2\ell}/A_{1\ell})^{1/(2r+1)}$ and $A_{1\ell}$, $A_{2\ell}$ are the versions of A_1, A_2 in (5.1) that arise when $\mathscr{R}$ on the left-hand side of (5.1) is replaced by $\mathscr{S}_\ell$. (Thus, $p_{\text{opt},\ell}$ is the version of p_{opt} at (5.3), when $\mathscr{S}_\ell$ replaces $\mathscr{R}$.) Let $\hat{p}_\ell$ denote the empirical resolution level that minimizes the cross-validation criterion CV_ℓ defined in (3.4).

Theorem 5.3. *If* (C_1), (C_2) *and* (C_5) *hold then*

$$\frac{\int_{\mathscr{S}_\ell} \{\hat{f}_{\text{lin}}(x|\hat{p}_\ell) - f(x)\}^2 \, dx}{\inf_{p>0} \int_{\mathscr{S}_\ell} \{\hat{f}_{\text{lin}}(x|p) - f(x)\}^2 \, dx} \to 1, \tag{5.6}$$

$$\hat{p}_\ell = \{1 + o_p(1)\} p_{\text{opt},\ell} \tag{5.7}$$

for $1 \leqslant \ell \leqslant m$.

Results (5.6) and (5.7) may be derived by modifying, in relatively minor ways, arguments of Stone (1984) and Hall (1983), respectively. In particular, note that, except for the fact that the linear estimator $\hat{f}_{\text{lin}}(\cdot|p)$ is of the generalized kernel type, rather than a standard kernel estimator, and $\mathscr{R}$ is properly contained in the support of f, the results of Stone are directly applicable to it.

Properties (5.6) and (5.7) together confirm that $\hat{p}_\ell$ provides asymptotic minimization of ISE on $\mathscr{S}_\ell$, and that it is asymptotic to the optimal primary resolution level there.

5.3. Comparison of our method with those of Tribouley (1995)

Assume that the decomposition $\mathscr{R} = \bigcup_\ell \mathscr{S}_\ell$ produces at least one subregion $\mathscr{S}_\ell$ where there are no points of discontinuity of $f^{(r-1)}$, and that $f^{(r)}$ does not vanish identically on any of the regions $\mathscr{S}_\ell$. (These are the sorts of subregion that are suggested by the discussion following the algorithm in Section 2.) Then it follows from the versions of (5.1) or (5.4) (depending on whether $t_\ell = r$ or $t_\ell < r$, respectively) with $\mathscr{R}$ replaced by $\mathscr{S}_\ell$, for $1 \leqslant \ell \leqslant m$, and from (5.6) and (5.7), that our empirical estimator $\hat{p} \equiv \min \hat{p}_\ell$ of the primary resolution level (see step 3 of the algorithm) is asymptotic to $\min p_{\mathrm{opt},\ell}$, where $p_{\mathrm{opt},\ell}$ was defined in Section 5.2; that $p_{\mathrm{opt},\ell} \sim B_3 n^{1/(2r+1)}$, where $0 < B_3 < \infty$; and that this is the optimal order for the primary resolution level to take.

On the other hand, if we use cross-validation for the full region $\mathscr{R}$, instead of dividing the region into subregions $\mathscr{S}_\ell$, and if $f^{(t)}$ has a finite number of non-degenerate jump discontinuities in $\mathscr{R}$, then it follows from (5.4) and from the version of (5.6) with $\mathscr{S}_\ell$ replaced by $\mathscr{R}$ that the empirical estimator of primary resolution level will be asymptotic to $n^{1/\{2(t_0+1)\}}$, where t_0 (assumed to satisfy $0 \leqslant t_0 \leqslant r-1$) denotes the minimum of values of t such that $f^{(t)}$ has a jump discontinuity in the interior of $\mathscr{R}$. This is the main approach suggested by Tribouley (1995), although it should be stressed that her techniques were developed for linear, rather than nonlinear, wavelet estimators. The method is identical to Tribouley's second approach in the case of estimating unimodal densities, such as the double exponential, where (when applied to nonlinear wavelet estimators of non-smooth functions) it suffers from problems similar to those noted immediately below.

A resolution level of size $n^{1/\{2(t_0+1)\}}$ is an order of magnitude larger than the 'optimal' order $n^{1/(2r+1)}$, and as a result the corresponding wavelet estimator will show more stochastic variability than is optimal. Indeed, using the analogy that the effective bandwidth of a wavelet estimator is the inverse of its primary resolution level, a 'global' cross-validation approach which addresses the whole region $\mathscr{R}$ simultaneously, rather than treating its decomposition into subregions, produces a curve estimate that in many respects is like that obtained using too small a bandwidth for a standard kernel estimator. We provided numerical evidence for this property in section 4.

5.4. Robustness of nonlinear estimator against over-smoothing

The nonlinear wavelet estimator $\hat{f}(\cdot|p)$ is more resistant against choice of too small a value of p than result (5.2) might indicate. Indeed, the regularity condition $p^{2r+1}\delta^2 \to \infty$ imposed there does not allow p to be of smaller order than $(n^{-1}\log n)^{1/(2r+1)}$, if (5.2) is to hold. In the contrary case, where $p = O\{(n\log n)^{1/(2r+1)}\}$, the mean squared error (and mean ISE) is generally of order $(n^{-1}\log n)^{2r/(2r+1)}$; see Remark 2.3 of Hall and Patil (1995b). This robustness against over-smoothing is not available to the linear estimator $\hat{f}_{\mathrm{lin}}(\cdot|p)$, for example. That it exists for nonlinear estimators provides further support for property (C) in Section 2, which underpins our method.

However, the problems suffered by choosing too *large* a value of p are of quite a

different nature, and the nonlinear estimator is no more resistant against them than is the linear estimator $\hat{f}_{\text{lin}}(\cdot|p)$.

5.5. Other measures of 'roughness'

In the discussion above we have quantified roughness in terms of discontinuities: a density is rougher if jump discontinuities occur in derivatives of lower order. This leads to a particularly simple and transparent theoretical description of performance, but of course it is not the only approach that could be taken. An alternative is to allow the density f to depend on n and equal $f + \eta_n$, where η_n denotes a function that integrates to 0 and is supported on a decreasingly small interval, as $n \to \infty$. For example, we might take $\eta_n(x) = \psi(\lambda_n x)$, where ψ denotes a fixed, smooth, compactly supported function satisfying $\int \psi = 0$, and $\{\lambda_n\}$ is a sequence of positive numbers diverging to ∞. In this model for roughness, rougher densities correspond to those for which λ_n is larger. It is possible to construct a theoretical account of this model that parallels that given above, and so provides further support for the properties (A)–(D) in Section 2 that motivate our algorithm for empirically choosing the primary resolution level.

5.6. Failure of direct nonlinear cross-validation

The cross-validation criterion CV_{ℓ}, defined in (3.4), is a special case of the general criterion,

$$\text{CV}(s) = \int \hat{f}(\cdot|s)^2 - 2n^{-1} \sum_{j=1}^{n} \hat{f}_{-j}(X_j|s), \tag{5.8}$$

where $\hat{f}(\cdot|s)$ is a density estimator computed from a sample $\mathcal{X} = \{X_1, \ldots, X_n\}$; $\hat{f}_{-j}(\cdot|s)$ is the version of $\hat{f}(\cdot|s)$ calculated from $\mathcal{X} \backslash \{X_j\}$; CV, $\hat{f}$ and $\hat{f}_{-j}$ depend on the smoothing parameter, s (e.g. bandwidth for a kernel estimator, resolution level for a wavelet estimator); and, for notational simplicity, we have taken the region of interest to be the whole real line.

Of course, the type of thresholding (hard or soft) and the smoothness of the wavelet function both affect performance, along with choice of primary resolution level. The distinction between hard and soft thresholding is present only in second-order terms in an expansion of ISE, however, and smoothness of ψ affects ISE in more nebulous ways. It is sometimes suggested that any known fractal properties of the target function could be reflected in the estimator by choosing a wavelet whose graph had a similar fractal dimension. This is beyond the scope of our work, not least because it is not really possible to choose the threshold or the smoothness of ψ by focusing on ISE properties. On the other hand, the results in Section 5.1 make it clear that there is a connection between optimal choice of primary resolution level, p, and wavelet order, r. Cross-validation allows us to assess empirically the effect that this linkage has on L^2 performance, and to minimize (at least asymptotically) its impact on estimator error.

The rationale behind cross-validation is that $\text{CV}(s)$ is an almost unbiased approximation to $Q(s) \equiv \int \hat{f}(\cdot|s)^2 - 2\int \hat{f}(\cdot|s) f$; and Q differs from ISE, $I(s) \equiv \int \{\hat{f}(\cdot|s) - f\}^2$, only

through the quantity $\int f^2$, which does not depend on s. Therefore, minimizing CV with respect to s *should* produce at least asymptotic minimization of ISE.

The difficulty with this argument is that most density estimators of practical interest converge to the true f at a faster rate than $n^{-1/4}$, and so ISE converges to zero a faster rate than $n^{-1/2}$. However, standard information-theoretic arguments show that we cannot approximate the 'diagonal' term in the ISE expansion, $D(s) \equiv \int \hat{f}(\cdot|s)f$, at a rate better than $n^{-1/2}$. Nevertheless, this does not often present a serious problem, since the error of size $n^{-1/2}$ in the approximation to $D(s)$ by the second term,

$$\hat{D}(s) \equiv n^{-1} \sum_{j=1}^{n} \hat{f}_{-j}(X_j|s), \tag{5.9}$$

in the formula for CV(s) in (5.8) does not depend on the smoothing parameter. Indeed, in many cases

$$\hat{D}(s) = D(s) - 2n^{-1} \sum_{j=1}^{n} \left\{ f(X_j) - \int f^2 \right\} + \text{terms of smaller order than } I(s). \tag{5.10}$$

The series on the right-hand side is of order $n^{-1/2}$, but does not depend on s. In the context of kernel methods this property is made explicit by Hall (1983), and is implicit in work of Stone (1984).

Unfortunately, however, the property often fails to hold when $\hat{f}$ is a nonlinear wavelet estimator. If the underlying density is smooth then the nonlinear term in a wavelet estimator vanishes with high probability, and so the estimator is virtually identical to its linear component, in which case the wavelet form of (5.10) holds true. But if the nonlinear component is not negligible then (5.10) fails, and with it fails the validity of the cross-validation algorithm.

To make this explicitly clear, let us assume conditions (C_1), that $\mathscr{R} = [a, b]$, that $a < 0 < b$, and that f has r bounded, continuous derivatives on $[a, 0)$ and on $(0, b]$, with left- and right-hand limits at 0, but $f(0-) \neq f(0+)$. Assume that ϕ and ψ are continuous, with support contained in the interval $[-v, v]$. Take the threshold δ_i to equal $C(n^{-1} \log n)^{1/2}$, for arbitrary $C > 0$, put $\nu = n^{1/(2r+1)}$ and let p equal the integer part of $\gamma\nu$, where $\gamma > 0$ is fixed. Take $s = p$ in CV(s) and $\hat{D}(s)$, as in (5.8) and (5.9).

Theorem 5.4. *Under the above conditions, there exists a sequence of random variables U_n, not depending on γ, and a Gaussian process $V(\gamma)$, with zero mean and marginal distribution depending non-degenerately on γ, such that*

$$(n\nu)^{1/2}\{\hat{D}(p) - U_n\} \to V(\gamma)$$

in distribution.

Note that $(n\nu)^{-1/2} = n^{-(r+1)/(2r+1)}$ is of strictly larger order than $n^{-2r/(2r+1)}$, which is the order of the ISE of $\hat{f}(\cdot|p)$; see (5.2). Therefore, CV(p) does not accurately approximate even the order of $\hat{f}(\cdot|p)$, up to terms that do not depend on p. Therefore, minimizing the cross-validation criterion for a nonlinear wavelet estimator does not asymptotically minimize

ISE. The same argument, with the same conclusion, may be used in the context of regression.

Proof of Theorem 5.4. Let $\mathscr{K}$ denote the set of integers lying in $[-v, v]$. Note that $\#\mathscr{K} \leqslant 2v + 1$. Write $\tilde{\mathscr{K}}$ for the complement of $\mathscr{K}$ in the set $\mathbb{Z}$ of all integers, and let $\hat{b}_{k,-j}$ and $\hat{b}_{ik,-j}$ denote the versions of $\hat{b}_k$ and $\hat{b}_{ik}$ computed from $\mathscr{X}\backslash\{X_j\}$ rather than $\mathscr{X}$. Given subsets $\mathscr{J}$, $\mathscr{J}_0$, $\mathscr{J}_1$, ... of $\mathbb{Z}$, put

$$S_j(\mathscr{J}, \mathscr{J}_0, \mathscr{J}_1, \ldots) \equiv \sum_{k\in\mathscr{J}} \hat{b}_{k,-j}\phi_k(X_j) + \sum_{i=0}^{q-1}\sum_{k\in\mathscr{J}_i} \hat{b}_{ik,-j} I(|\hat{b}_{ik,-j}| > \delta_i)\psi_{ik}(X_j).$$

Observe that

$$\hat{D}(s) = n^{-1}\sum_{j=1}^{n} S_j(\tilde{\mathscr{K}}, \tilde{\mathscr{K}}, \ldots) + n^{-1}\sum_{j=1}^{n} S_j(\mathscr{K}, \mathscr{K}, \ldots). \tag{5.11}$$

The term $S_j(\tilde{\mathscr{K}}, \tilde{\mathscr{K}}, \ldots)$ is not influenced at all by the discontinuity in f at 0; all the influence has been incorporated into $S_j(\mathscr{K}, \mathscr{K}, \ldots)$. In fact, the first series on the right-hand side of (5.11) may be treated as in the case of linear estimators, and thereby shown to equal $o_p\{(nv)^{-1/2}\}$ plus terms that do not depend on γ. The second series is quite different, however.

To appreciate the differences, note that $0 \in \mathscr{K}$ and that the contribution of the pair $(i, k) = (i, 0)$ to the second series in (5.11) equals

$$n^{-1}\sum_{j=1}^{n} \hat{b}_{i0,-j}\psi_{i0}(X_j). \tag{5.12}$$

(We have dropped the indicator function since, for the present choice of (i, k) and our choice of δ_i, the indicator equals 1 for all sufficiently large n, with probability 1.) The quantity in (5.12) equals

$$\begin{aligned} n^{-1}\sum_{j=1}^{n} \hat{b}_{i0}\psi_{i0}(X_j) + O_p(n^{-1}) &= \hat{b}_{i0}^2 + O_p(n^{-1}) \\ &= b_{i0}^2 + 2(\hat{b}_{i0} - b_{i0})\, b_{i0} + O_p(n^{-1}). \end{aligned} \tag{5.13}$$

Modulo the deleted indicator function, the term b_{i0}^2 on the right-hand side of (5.13) represents that part of $D(s)$ that this component of $\hat{D}(s)$ is estimating. Furthermore,

$$\begin{aligned} b_{i0} &= p_i^{1/2}\int \psi(p_i u) f(u)\mathrm{d}u = p_i^{-1/2}\int \psi(u) f(p_i^{-1}u)\mathrm{d}u \\ &= p_i^{-1/2}\int \psi(u)\{f(0-)I(u<0) + f(0+)I(u>0)\}\mathrm{d}u + o(p_i^{-1/2}) \\ &= p_i^{-1/2} b + o(p_i^{-1/2}), \end{aligned}$$

where $b \equiv f(0-) \int_{u<0} \psi(u) \mathrm{d}u + f(0+) \int_{u>0} \psi(u) \mathrm{d}u$.

Given $t > 0$, define

$$T_1(t) = (t\nu/n)^{1/2} \sum_{j=1}^{n} \{\psi(t\nu X_j) - \mu(t)\},$$

where $\mu(t) = \mathrm{E}\{\psi(t\nu X_1)\}$. Put $T_2(t) = t^{-1/2} T(t)$. Let Z be a Gaussian process defined on the real line, with zero mean and covariance function

$$\sigma(t_1, t_2) = \int \psi(t_1 u) \psi(t_2 u) J(u) \mathrm{d}u,$$

where $J(u) = f(0+)$ if $u > 0$ and $J(u) = f(0-)$ if $u \leqslant 0$. Using standard methods for deriving invariance principles, it may be proved that $T_2 \to Z$ weakly in the space $C[t_1, t_2]$ of continuous functions on $[t_1, t_2]$, for any $0 < t_1 < t_2 < \infty$.

Hence, for any fixed $i_0 \geqslant 0$,

$$(n\nu)^{1/2} \sum_{i=0}^{i_0} (\hat{b}_{i0} - b_{i0}) b_{i0} \to b \sum_{i=0}^{i_0} 2^{-i/2} Z(2^i \gamma)$$

as $n \to \infty$, where the convergence is in distribution. Moreover, it may be proved by Markov's inequality that, for each $\varepsilon > 0$,

$$\lim_{i_0 \to \infty} \limsup_{n \to \infty} P\left\{ \left| (n\nu)^{1/2} \sum_{i=i_0}^{q-1} (\hat{b}_{i0} - b_{i0}) b_{i0} \right| > \varepsilon \right\} = 0,$$

$$\lim_{i_0 \to \infty} P\left\{ \left| \sum_{i=i_0}^{\infty} 2^{-i/2} Z(2^i \gamma) \right| > \varepsilon \right\} = 0.$$

Therefore,

$$(n\nu)^{1/2} \sum_{i=0}^{q-1} (\hat{b}_{i0} - b_{i0}) b_{i0} \to b \sum_{i=0}^{\infty} 2^{-i/2} Z(2^i \gamma)$$

as $n \to \infty$, where again the convergence is in distribution. We may deduce from this result, the fact that $q = O(\log n)$, and the discussion between (5.12) and (5.13) about indicator functions, that

$$(n\nu)^{1/2} \left\{ n^{-1} \sum_{j=1}^{n} \sum_{i=0}^{q-1} \hat{b}_{i0,-j} I(|\hat{b}_{i0,-j}| > \delta_i) \psi_{i0}(X_j) - \sum_{i=0}^{q-1} b_{i0}^2 \right\} \to 2b \sum_{i=0}^{\infty} 2^{-i/2} Z(2^i \gamma)$$

in distribution. The fact that the asymptotic distribution depends non-degenerately on γ means that the effect of p on the value of the term in (5.13) is not negligible at the level $(n\nu)^{-1/2}$, which is of strictly larger order than that of ISE.

Similar arguments apply to the other terms (i.e. $k \neq 0$) that contribute to the second series at (5.11), and lead to the theorem.

Acknowledgement

The reviewers' constructive comments have proved particularly helpful.

References

Daubechies, I. (1992) *Ten Lectures on Wavelets.* Philadelphia: Society of Industrial and Applied Mathematics.

Delyon, B. and Juditsky, A. (1996) On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.*, **3**, 215–228.

Donoho, D.L. and Johnstone, I. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho, D.L. and Johnstone, I. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224.

Donoho, D.L., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B*, **57**, 301–369.

Donoho, D.L., Johnstone, I., Kerkyacharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Ann. Statist.*, **24**, 508–539.

Hall, P. (1983) Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.*, **11**, 1156–1174.

Hall, P. and Nason, G.P. (1997) On choosing a non-integer resolution level when using wavelet methods. *Statist. Probab. Lett.*, **34**, 5–11.

Hall, P. and Patil, P. (1995a) On wavelet methods for estimating smooth functions. *Bernoulli*, **1**, 41–58.

Hall, P. and Patil, P. (1995b) Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.*, **23**, 905–928.

Hall, P. and Patil, P. (1996) Effect of threshold rules on the performance of wavelet-based curve estimators. *Statist. Sinica*, **6**, 331–345.

Hall, P. and Schucany, W.R. (1989) A local cross-validation algorithm. *Statist. Probab. Lett.*, **8**, 109–117.

Hurvich, C.M. and Tsai, C.-L. (1998) A crossvalidatory AIC for hard wavelet thresholding in spatially adaptive function estimation. *Biometrika*, **85**, 701–710.

Marron, J.S. and Wand, M.P. (1992) Exact mean integrated squared error. *Ann. Statist.*, **20**, 712–736.

Mielniczuk, J., Sarda, P. and Vieu, P. (1989) Local data-driven bandwidth choice for density estimation. *J. Statist. Plann. Inference*, **23**, 53–69.

Nason, G.P. (1996) Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc. Ser. B*, **58**, 463–479.

Nason, G.P. and Silverman, B.W. (1994) The discrete wavelet transform in S. *J. Comput. Graph. Statist.*, **3**, 163–191.

Stone, C.J. (1984) An asymptotically optimal window selection rule for kernel density estimators. *Ann. Statist.*, **12**, 1285–1297.

Tribouley, K. (1995) Practical estimation of multivariate densities using wavelet methods. *Statist. Neerlandica*, **49**, 41–62.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman & Hall.

Wang, Y. (1998) Change curve estimation via wavelets. *J. Amer. Statist. Assoc.*, **53**, 163–172.

Received June 1999 and revised November 2000