

SOME NEW METHODS FOR WAVELET DENSITY ESTIMATION

By D.R.M. HERRICK

G.P. NASON

and

B.W. SILVERMAN

University of Bristol, Bristol

SUMMARY. This article proposes some non-linear, thresholded wavelet density estimators, and investigates the practical problems involved in their implementation. Our proposed thresholding method exploits the non-stationary variance structure of the wavelet coefficients. One proposed method of estimating the variances of the raw coefficients uses the scaling function coefficients. Since these are available as a by-product of the discrete wavelet transform, no extra effort is required to find them. The performance of the methodology is assessed on a real dataset from a forensic application and simulated data from a well known test function.

1. Introduction

In density estimation as in other smoothing problems, wavelet methods can provide a convenient non-linear approach that exploits the localization properties of the wavelets to provide good adaptive estimates. Theoretical properties of such estimators have been addressed in several papers, for example Donoho *et al.* (1996) and Hall and Patil (1995), and a thorough review of the subject can found in Chapter 7 of Vidakovic (1999). The present article takes a practical approach to the method, and in particular takes explicit account of the non-constant variance of the wavelet coefficients when thresholding. Software in **S-Plus** and **C** has been written to implement the methods of this paper, and is available in Herrick (2001).

Paper received July 2001.

AMS (1991) subject classification. .

Key words and phrases.

1.1 *Example data set.* The example data set consists of measurements of the lengths of gene fragments, used as markers for forensic identification. There are measurements corresponding to four loci (the position of the fragment on the gene) for each of two ethnic groups (Afro-Caribbean and Northern-European). These yield eight data sets, of sizes varying from about 420 to 2401. The data values all fall in the range from 1 to 20.

Because of the nature of the sampling mechanism underlying these data, it is not to be expected that their distributions are smooth in a large-scale sense. Instead, the densities may well be very spiky, and any estimation procedure will need to be able to cope with this. The spikyness will give some very steep sections to the density functions, which is why wavelet methods may well be suitable. The sampling density is useful, for example, to calculate a likelihood ratio for comparing whether two sets of gene fragments are independent measurements of the same gene or independent measurements of two independent genes. Kernel methods will underestimate the peaks of the sampling density and so will bias the likelihood ratio. This forensic application is discussed in more detail in Roeder (1994) and Aitken (1995).

Kernel density estimates. Kernel density estimation is discussed in Silverman (1986), Wand and Jones (1995) and Simonof (1996). From the description of the data, and simple summary plots like stem and leaf or histograms, it is obvious that the data are not normal, and in general the distributions are skewed to the left. Three bandwidth selection procedures were used initially: the S-Plus default (page 137 of Venables and Ripley (1994)), the “rule of thumb” method (pages 45–47 of Silverman (1986)) and the Sheather-Jones method, with no sub-sampling (Sheather and Jones (1991)). From there the pictures were smoothed by eye. It can be seen that the two ethnic groups have very similar distributions at locus 1 but quite different distributions at locus 2 and 3. At locus 4, the distributions of both groups peak in roughly the same area, but the Northern European data has an extra peak around $x=4$ which is not apparent in the Afro-Caribbean data.

1.2 *Basic wavelet estimation scheme and notation.* Consider estimating a density function f from i.i.d. data x_1, \dots, x_n , using wavelets. Recall from Donoho *et al.* (1996) that the formal wavelet expansion of the density function is:

$$f(x) \sim \sum_{k \in \mathbb{Z}} c_{Lk} \phi_{Lk}(x) + \sum_{j \in \mathbb{J}_L} \sum_{k \in \mathbb{Z}} d_{jk} \psi_{jk}(x), \quad (1)$$

where $\mathbb{J}_L = \{m \in \mathbb{Z} : m \geq L\}$ and the coefficients c_{Lk} and d_{jk} are given by:

$$c_{Lk} = \int_{-\infty}^{\infty} \phi_{Lk}(x) f(x) dx \quad \text{and} \quad d_{jk} = \int_{-\infty}^{\infty} \psi_{jk}(x) f(x) dx .$$

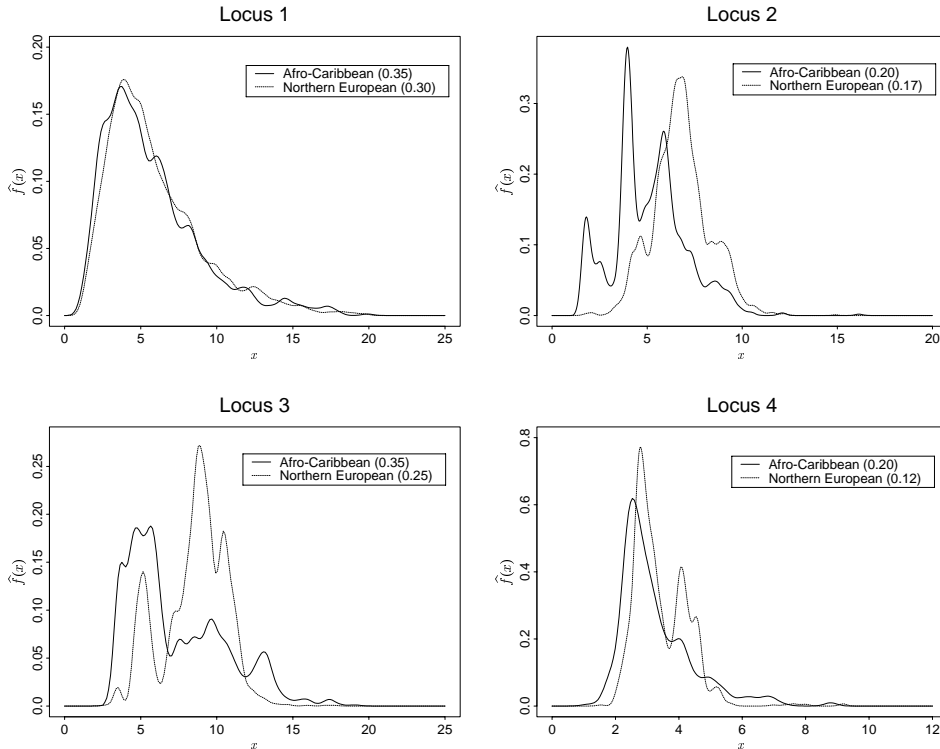


Figure 1. Kernel density estimates of the gene fragment data, using a normal kernel. The smoothing parameters, which were chosen by eye, are given in the legends.

The wavelet basis functions at resolution level j are given by

$$\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k) \quad \text{and} \quad \psi_{jk}(x) = 2^{j/2}\psi(2^j x - k). \quad (2)$$

The primary resolution level L determines the scale of the largest effects that can be affected by the smoothing inherent in the procedure. The methodology of the paper is not sensitive to the choice of L , because very large-scale effects are always retained by the estimation. For the remainder of the paper it is set to a value where a single scaling function at level L has range covering the whole of the observed data.

It is possible to allow for non-dyadic resolutions by rescaling all the data values by a *tuning parameter* $\tau \in (\frac{1}{2}, 1]$; we shall not consider this procedure in detail but it is provided as an option in the software associated with this paper.

Consider a density estimate which has the same form as (1). As for general orthogonal series estimators (Čencov (1962)), the true coefficients can be written as:

$$c_{jk} = E[\phi_{jk}(X)] \quad \text{and} \quad d_{jk} = E[\psi_{jk}(X)] ,$$

where the obvious estimators are:

$$\tilde{c}_{jk} = \frac{1}{n} \sum_{i=1}^n \phi_{jk}(X_i) \tag{3}$$

$$\tilde{d}_{jk} = \frac{1}{n} \sum_{i=1}^n \psi_{jk}(X_i) . \tag{4}$$

These empirical coefficients are calculated for resolution levels L up to some large value J , called the *finest resolution level*. The \tilde{d}_{jk} are then thresholded to give estimated coefficients \hat{d}_{jk} . The resulting density estimate is then:

$$\hat{f}(x) = \sum_k \tilde{c}_{Lk} \phi_{Lk}(x) + \sum_{j=L}^{J-1} \sum_k \hat{d}_{jk} \psi_{jk}(x) .$$

The choice of the parameter J will be discussed later in this article. Note that n is not required to be a power of two and that this theory is in principle also valid for higher dimensional data.

2. Efficient Empirical Coefficient Computation

Direct use of (3) and (4) to compute the empirical coefficients \tilde{c}_{jk} and \tilde{d}_{jk} is slow, and it is more efficient to calculate the \tilde{c}_{Jk} directly and then use the discrete wavelet transform (DWT) to find the \tilde{c}_{Lk} and \tilde{d}_{jk} . Efficient computation of the \tilde{c}_{Jk} will now be considered.

The first step is to determine the values of k for which it is necessary to calculate \tilde{c}_{Jk} . Assume compactly supported wavelets are being used, and that a_ϕ, b_ϕ are integers for which it is known that $\phi(t)$ is zero outside the range (a_ϕ, b_ϕ) . Let $w_\phi = b_\phi - a_\phi$. Let $\lceil x \rceil$ denote the smallest integer $\geq x$ and $\lfloor x \rfloor$ the largest integer $\leq x$. The empirical coefficient \tilde{c}_{Jk} will be zero if none of the data fall in the support of ϕ_{Jk} ; this will certainly be the case if k is outside the range $[k_{\min}, k_{\max}]$, where

$$k_{\min} = \min_{i=1}^n [2^J x_i - b_\phi] = \lceil 2^J x_{\min} - b_\phi \rceil, \quad (5)$$

$$k_{\max} = \max_{i=1}^n [2^J x_i - a_\phi] = \lfloor 2^J x_{\max} - a_\phi \rfloor. \quad (6)$$

Next, we note that because of the bounds on the support of ϕ , there are at most w_ϕ consecutive values of k for which a given data point can fall in the support of ϕ_{Jk} . The total number of pairs (i, k) for which $\phi_{Jk}(x_i)$ is nonzero therefore at most nw_ϕ , whatever the resolution. Considering the extreme case where the relevant k 's are different for each i , no more than nw_ϕ of the coefficients \tilde{c}_{Jk} can be nonzero for any value of the finest resolution level J .

For each data point x_i , the Daubechies-Lagarias algorithm (Daubechies and Lagarias (1992)) is used to approximate the non-zero $\phi_{Jk}(x_i)$. These contributions are then added onto the appropriate empirical coefficients. In other words, rather than calculating the coefficients \tilde{c}_{Jk} one at a time, they are gradually built up by taking each of the x_i in turn and adding on its contributions to all the relevant \tilde{c}_{Jk} .

Having estimated the \tilde{c}_{Jk} , the DWT is used to find the empirical coefficients at coarser levels. In the case of density estimation on the line there is no need to use any boundary conditions, because the full support of all the wavelet functions is considered.

Another consequence of having complete knowledge about the coefficients is that there is no need to have a dyadic number of coefficients at the highest level (J) because we can use zero-padding boundary conditions at every level without introducing any error. The effect is that all the \tilde{c}_{jk} are calculated for $f_c^j \leq k \leq l_c^j$, where f_c^j and l_c^j are the smallest and largest values of k such that $\tilde{c}_{jk} \neq 0$. Similarly, the \tilde{d}_{jk} are calculated for $f_d^j \leq k \leq l_d^j$, where f_d^j and l_d^j are the smallest and largest possible values of k such that $\tilde{d}_{jk} \neq 0$.

Further economies would be possible if a sparse discrete wavelet transform procedure were available. At the higher scales, we would only need to maintain pointers to the positions at which the coefficients were nonzero. Extending to this case is beyond the scope of the present work, but would be an interesting computational topic for future implementation.

3. Choosing the Finest Resolution Level

The finest resolution level J determines the smallest scale on which any effects of interest can be detected or represented. In this section one pos-

sible approach to choosing the parameter J will be investigated. Further possibilities are discussed in section 3.5 of Herrick (2000).

Suppose J is chosen to be the smallest j for which the $\phi_{jk}(x)$ each cover at most one of the distinct data values. If all the data are distinct there may then be nw_ϕ nonzero coefficients. Then the support width of $\phi_{J_0}(x)$ is less than the smallest gap between data points. Assume that data are ordered, and let $\Delta x_i = x_{i+1} - x_i$. In real applications there may well be coincident data. Here, only unique points will be considered, so any of the Δx_i that are zero are ignored. Then to be certain that each $\phi_{jk}(x)$ covers at most one of the data values, J must satisfy

$$J \geq \log_2[w_\phi / \min\{\Delta x_i : \Delta x_i > 0\}]. \quad (7)$$

The smallest J satisfying this condition is called the *minimum inter-point distance* choice of the finest resolution level.

If J is chosen by (7) then the original data points will each correspond to a disjoint block of coefficients in the sequence of high resolution scaling function coefficients, with the blocks separated by zeroes. Thus, for the minimum inter-point distance choice of finest resolution level, the empirical coefficients essentially contain complete information about the data, and no smoothing has been carried out in restricting the wavelet expansion to coefficients below this level.

The number of non-zero coefficients increases with j until the limit of nw_ϕ is reached. Beyond this point, the computational effort required to calculate just the non-zero coefficients should not increase with j . However, the length of the vector needed to store the coefficients will increase, and for large J this storage requirement will be prohibitive. Implementing a sparse DWT as discussed above would of course alleviate this problem.

One useful modification of the minimum inter-point distance idea is to use a *discounted minimum inter-point distance* where we discard the smallest $\nu\%$ of the inter-point distances and choose J so that the support width of ϕ_{J_0} is less than the smallest of the remaining inter-point distances. If a relatively small value of ν leads to an acceptable value of J , we may be assured that very little detail is lost by using this value.

4. Thresholding

Having computed all the empirical wavelet coefficients, the next stage is to threshold (or denoise) them before reconstruction. Most established

methods of threshold selection assume that the noisy wavelet coefficients are independent and normally distributed with equal variance. The empirical coefficients (3) and (4) are, however, not. Although Donoho *et al.* (1996) suggest the use of a level dependent threshold proportional to $\sqrt{j/n}$, the variance structure of the empirical coefficients has not yet been exploited in deriving a thresholding rule.

4.1 *Covariance structure of the empirical coefficients.* Since the observations X_i are assumed to be i.i.d. from f , it is easy to show that

$$\text{Cov} [\tilde{d}_{j_1 k_1}, \tilde{d}_{j_2 k_2}] = \frac{1}{n} \left(\int \psi_{j_1 k_1}(x) \psi_{j_2 k_2}(x) f(x) dx - d_{j_1 k_1} d_{j_2 k_2} \right). \quad (8)$$

Similarly,

$$\text{Cov} [\tilde{c}_{j_1 k_1}, \tilde{c}_{j_2 k_2}] = \frac{1}{n} \left(\int \phi_{j_1 k_1}(x) \phi_{j_2 k_2}(x) f(x) dx - c_{j_1 k_1} c_{j_2 k_2} \right). \quad (9)$$

It can be seen that even coefficients sufficiently separated in the expansion for the integral in (8) to be zero will only be asymptotically uncorrelated. More seriously, it follows from (8) that the variances of the empirical coefficients are not constant even within levels; they are given by

$$\text{Var} [\tilde{d}_{jk}] = \frac{1}{n} \left(\int \psi_{jk}^2(x) f(x) dx - d_{jk}^2 \right). \quad (10)$$

The low resolution coefficients are linear combinations of a relatively large number of independent random variables, so by the central limit theorem, an assumption of normality might well be reasonable. However, these coefficients are still not independent. No normality assumption can reasonably be made about the high resolution coefficients, as these are linear combinations of only a few random variables.

Because of these complications, thresholding empirical coefficients from the Haar basis will be considered first. The simplicity of the Haar basis will enable insights to be gained for the more general case.

4.2 *Haar wavelet basis.* Some results specific to the Haar basis, which will be of use later in this section, will now be derived.

Using the definition of the Haar mother wavelet and scaling function it is easy to see that the true coefficients of the density can be written as

$$d_{jk} = 2^{j/2} \left\{ \int_{2^{-j}k}^{2^{-j}(k+\frac{1}{2})} f(x) dx - \int_{2^{-j}(k+\frac{1}{2})}^{2^{-j}(k+1)} f(x) dx \right\}, \quad (11)$$

$$c_{jk} = 2^{j/2} \int_{2^{-j}k}^{2^{-j}(k+1)} f(x) dx. \quad (12)$$

In addition, notice the important fact that

$$\int \psi_{jk}^2(x) f(x) dx = 2^{j/2} c_{jk} . \quad (13)$$

From (10) and (13) the variance of the empirical Haar wavelet coefficients can be written as:

$$\text{Var} [\tilde{d}_{jk}] = \frac{1}{n} (2^{j/2} c_{jk} - d_{jk}^2) . \quad (14)$$

Now consider the distributions of the coefficients. Let N_{jk} be the number of data points which fall in the interval $[2^{-j}k, 2^{-j}(k+1))$. Then for the Haar basis

$$N_{jk} = n2^{-j/2} \tilde{c}_{jk} . \quad (15)$$

The probability of a random observation falling in the interval $[2^{-j}k, 2^{-j}(k+1))$ is:

$$P\left(X \in [2^{-j}k, 2^{-j}(k+1))\right) = \int_{2^{-j}k}^{2^{-j}(k+1)} f(x) dx = 2^{-j/2} c_{jk} . \quad (16)$$

and so $N_{jk} \sim \text{Bin}(n, 2^{-j/2} c_{jk})$.

Now consider the conditional distribution of \tilde{d}_{jk} given \tilde{c}_{jk} . Let M_{jk} be the number of data points which fall in the interval $[2^{-j}k, 2^{-j}(k+\frac{1}{2}))$. Then,

$$(M_{jk} | N_{jk} = n_{jk}) \sim \text{Bin}(n_{jk}, \frac{1}{2}(1 + d_{jk}/c_{jk})) . \quad (17)$$

From (4)

$$n2^{-j/2} \tilde{d}_{jk} = M_{jk} - (n_{jk} - M_{jk}) \quad (18)$$

and hence the distribution of \tilde{d}_{jk} conditional on \tilde{c}_{jk} can be found exactly, in terms of the binomial distribution.

Thresholding the empirical (noisy) wavelet coefficients is an attempt to find which of the coefficients are really non-zero. This can be thought of as performing multiple hypothesis tests of:

$$H_0 : d_{jk} = 0 \quad v \quad H_1 : d_{jk} \neq 0, \quad (19)$$

which is equivalent to testing on the basis of M_{jk} whether the binomial probability in (17) is equal to $\frac{1}{2}$. It should be noted that the large number of significance tests in this procedure are not independent, as the \tilde{d}_{jk} are not independent. The p -values are in any case conditional on the relevant \tilde{c}_{jk} for each (j, k) . We proceed by fixing a standard p -value and using this for all the tests.

A normal approximation to the binomial distribution makes the calculations simpler. From (14) the conditional variance of \tilde{d}_{jk} under the null hypothesis is $n^{-1}2^{j/2}\tilde{c}_{jk}$, which can be estimated using the empirical scaling function coefficients. Since, for the Haar basis, ϕ is non-negative, the \tilde{c}_{jk} will themselves necessarily be non-negative.

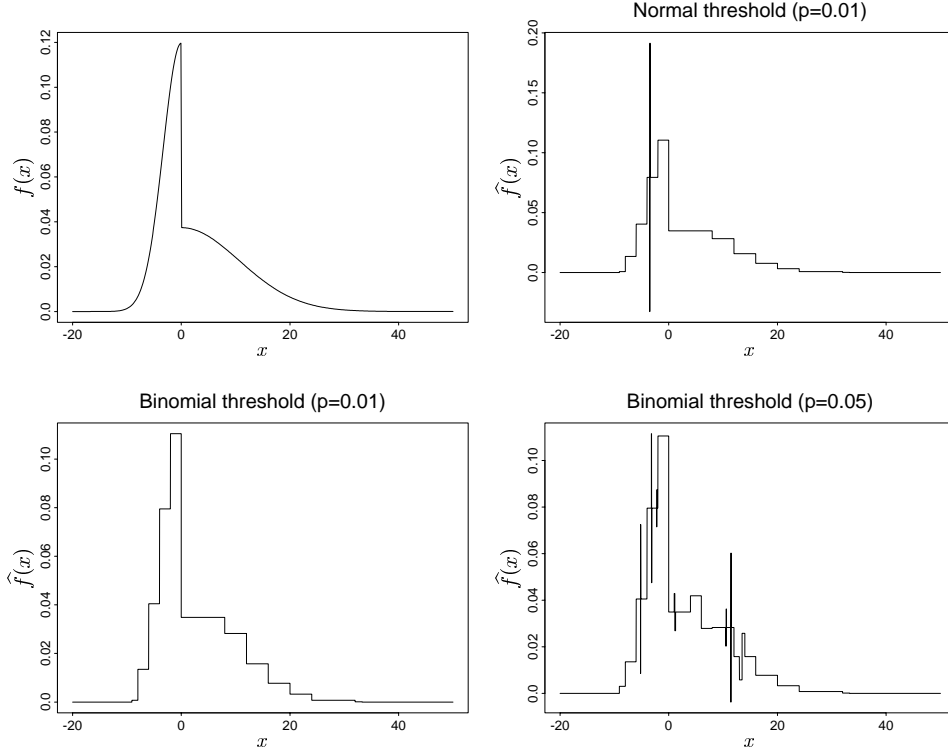


Figure 2. Haar wavelet density estimates of the discontinuous density (the left half is $N(0, 10/3)$ and the right half $N(0, 32/3)$) using a sample of size 1000. The top row shows the true density (left) and the estimate using the normal significance test for thresholding (right). The bottom row shows estimates using the binomial test at different p -values.

Simulated comparison. In a simulation comparison, the exact binomial and normal approximation alternatives give very similar results. They were tested on a sample of 1000 simulated observations from a density containing a discontinuity, which is shown in the top left of figure 2. It consists of two halves of normal distributions with different variances ($10/3$ and $32/3$). The finest resolution level J was set partly subjectively to 10. Support for this choice is provided by the discounted minimum interpoint distance procedure,

as discussed in section 3; the value $\nu=5$ gives $J=11$ and $\nu=10$ gives $J=9$. The p -values at which the significance tests were performed are indicated in the plot titles. The results are shown in figure 2. The binomial thresholding, with a p -value of 0.01 performs well. Note that normal thresholding has let a fine scale coefficient through which the binomial thresholding kills. This sort of problem is to be expected as the normal approximation will be much worse at fine scales, as n_{jk} will be much smaller than at coarser scales.

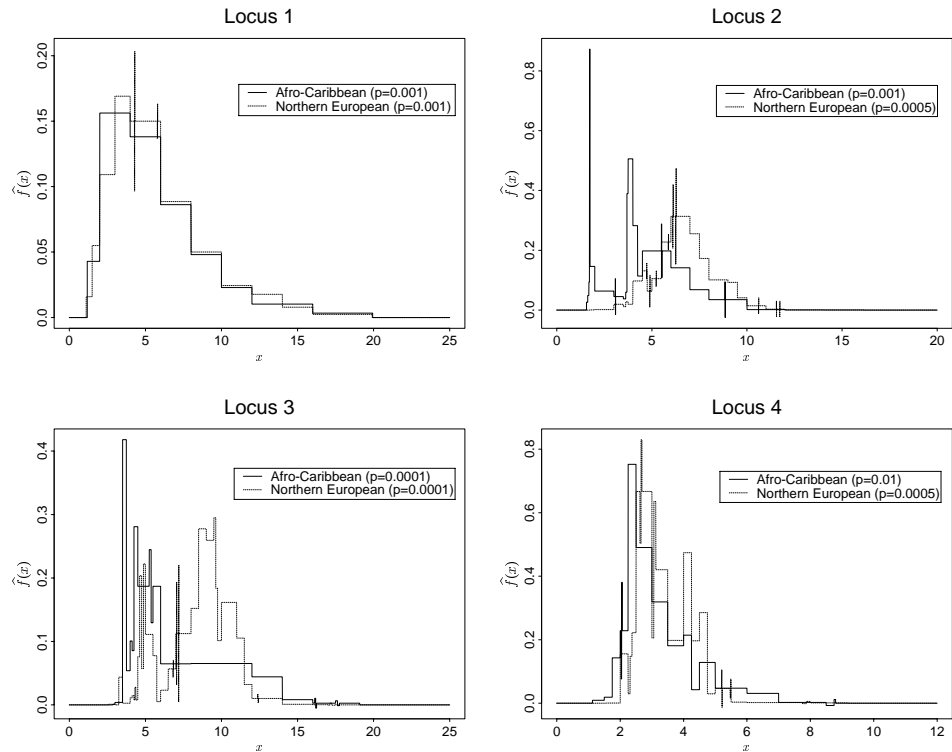


Figure 3. Haar wavelet density estimates of the gene fragment data. The p -values are given in the legends.

Gene fragment data. Figure 3 shows density estimates for the gene fragment data using Haar wavelets with binomial thresholding. The minimum distance approach (without any discounting) suggested the value $J=7$. The general behaviour exhibited is the same as that shown by the kernel density estimates, but the wavelets pick up some large spikes which are very resistant to thresholding (as well, unfortunately, as some smaller blips which would have been better smoothed out). A good example is the huge spike on the

left of locus 2 for the Afro-Caribbean data. Closer examination of the data shows that there is a very tight cluster of points in this position; thus there is good evidence in the data for the spike. This is the type of information that might not be picked out by kernel methods, but which wavelets do pick out. Another such feature is discussed in more detail below.

5. More General Wavelet Bases

5.1 *Using the scaling function coefficients.* In the above discussion, we have used the key property $\psi^2 = \phi$ of the Haar basis. This is obviously not true for any other of the Daubechies wavelets. In order to estimate the variance of the wavelet coefficients in the more general case, it will be necessary to approximate ψ^2 in some way, so that the integral in (10) can be simplified. Following the Haar case, try approximating ψ^2 as a linear combination of scaling functions at some finer level $m \geq 1$:

$$\psi^2(x) \approx \sum_l e_l \phi_{ml}(x) \text{ where } e_l = \int \psi^2(x) \phi_{ml}(x) dx. \quad (20)$$

The coefficients e_l can be found by applying the DWT to a suitably discretized version of ψ^2 . This result can be used to approximate ψ_{jk}^2 :

$$\psi_{jk}^2(x) \approx 2^{j/2} \sum_l e_l \phi_{j+m, l+2^m k}(x),$$

and so the following approximation can now be made:

$$\int \psi_{jk}^2(x) f(x) dx \approx 2^{j/2} 2^{j/2} \sum_l e_l c_{j+m, l+2^m k}.$$

Using this result in (10):

$$\text{Var} [\tilde{d}_{jk}] \approx \frac{1}{n} \left(2^{j/2} \sum_l e_l c_{j+m, l+2^m k} - d_{jk}^2 \right),$$

so that under the null hypothesis (19):

$$\text{Var}_{H_0} [\tilde{d}_{jk}] \approx \frac{2^{j/2}}{n} \sum_l e_l c_{j+m, l+2^m k}.$$

This can be estimated using the empirical coefficients:

$$\widehat{\text{Var}}_{H_0} [\tilde{d}_{jk}] = \frac{2^{j/2}}{n} \sum_l e_l \tilde{c}_{j+m, l+2^m k} . \quad (21)$$

If the empirical coefficients are then assumed to be normally distributed, they can be thresholded by performing multiple significance tests.

A nice feature of this idea is that the \tilde{c}_{jk} are computed by the DWT, so it exploits information that has already been computed. However, it will not be possible to use (21) on the finest $m - 1$ levels of wavelet coefficients, as the necessary scaling function coefficients are not available. These coefficients will simply be set to zero. If m is small and J large then, under mild smoothness assumptions, this will not discard any appreciable signal.

Figure 4 shows the approximation (20) for the Daubechies extremal phase wavelet ($N=2$) and the least asymmetric wavelets ($N=4, 10$) for $m=1, 2, 3$. As can be seen the approximation improves as m and N increase. Visually, the approximations at $m=1$ are clearly not good, but those at $m=2$ fit the major features of ψ^2 well. The approximations at $m=3$ match ψ^2 very closely.

This method was tried out on the data used in figure 2 for a variety of Daubechies wavelet bases for $m=1, 2$ and 3 . Unfortunately the results were very poor in that the resulting estimates contained lots of high frequency noise, unless the thresholding was so severe as to remove most of the signal as well. This technique is investigated further in section 5, using simulated data from a known density and the Haar wavelet. The failure of that experiment indicates that the failure here is not due to the approximation of ψ^2 , but more fundamental problems, such as an overall lack of statistical power.

5.2 *Estimating the variances directly.* Another approach is to estimate the variances of the individual coefficients directly. We can estimate the integral in (10) by using the Monte Carlo integral provided by the data themselves:

$$\int \psi_{jk}^2(x) f(x) dx \approx \frac{1}{n} \sum_{i=1}^n \psi_{jk}^2(X_i) .$$

Under the null hypothesis (19), the estimated variance is:

$$\widehat{\text{Var}}_{H_0} [\tilde{d}_{jk}] = \frac{1}{n^2} \sum_{i=1}^n \psi_{jk}^2(X_i) . \quad (22)$$

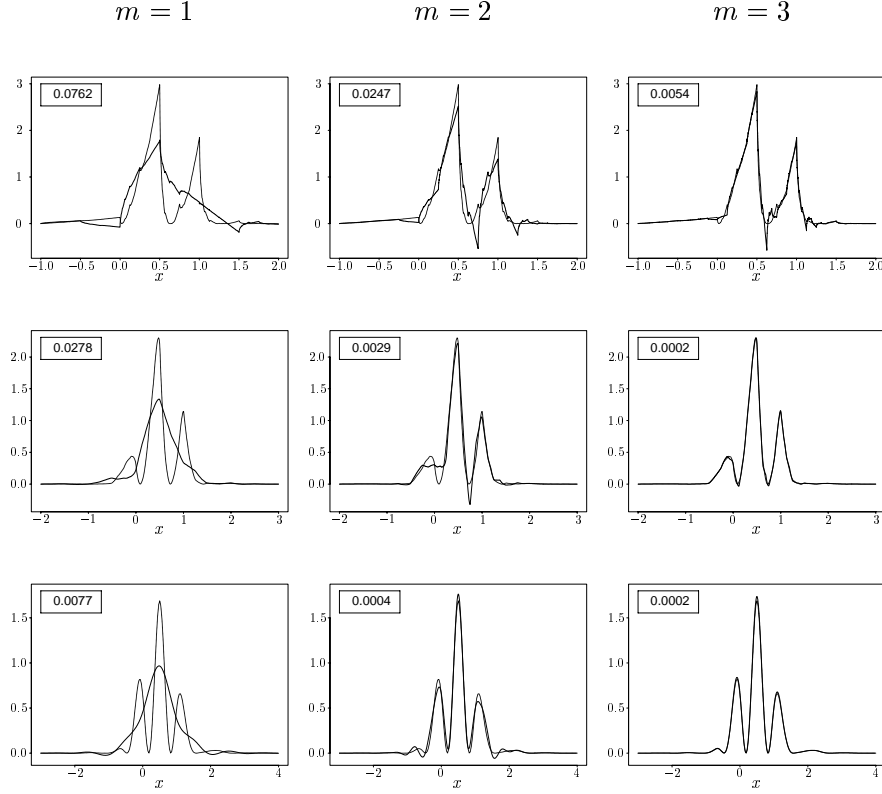


Figure 4. $\psi^2(x)$ (dotted) and the approximation $\sum_l e_l \phi_{m_l}(x)$ (solid) for three different values of m . The top row is for the Daubechies extremal phase wavelet with $N=2$ and the other two rows for the Daubechies least asymmetric wavelets with $N=4$ (middle) and $N=10$ (bottom). The average squared errors of the approximations are displayed in the legends.

An efficient way to calculate these is to estimate the covariance matrix of the \tilde{c}_{Jk} and use the DWT to find the estimated variances (22) of the \tilde{d}_{jk} . From (9)

$$\text{Cov} [\tilde{c}_{Jk_1}, \tilde{c}_{Jk_2}] = \frac{1}{n} \left(\int \phi_{Jk_1}(x) \phi_{Jk_2}(x) f(x) dx - c_{Jk_1} c_{Jk_2} \right). \quad (23)$$

In the calculations of the variance structure of the wavelet coefficients, the wavelet transform with respect to both k_1 and k_2 of this expression is taken. For any particular coefficient, this double transform of the second

term in (23) will be d_{jk}^2 . Under H_0 , $d_{jk}=0$, so it is only necessary to consider the double transform of the estimate of the first term in (23). This estimate is

$$\widehat{E}[\tilde{c}_{Jk_1}\tilde{c}_{Jk_2}] = \frac{1}{n^2} \sum_{i=1}^n \phi_{Jk_1}(X_i)\phi_{Jk_2}(X_i). \quad (24)$$

Notice that (24) defines a band limited matrix, where the bandwidth is $2N-1$. The fast algorithm of ovac and Silverman (2000) can then be applied to find the variances of the wavelet coefficients themselves.

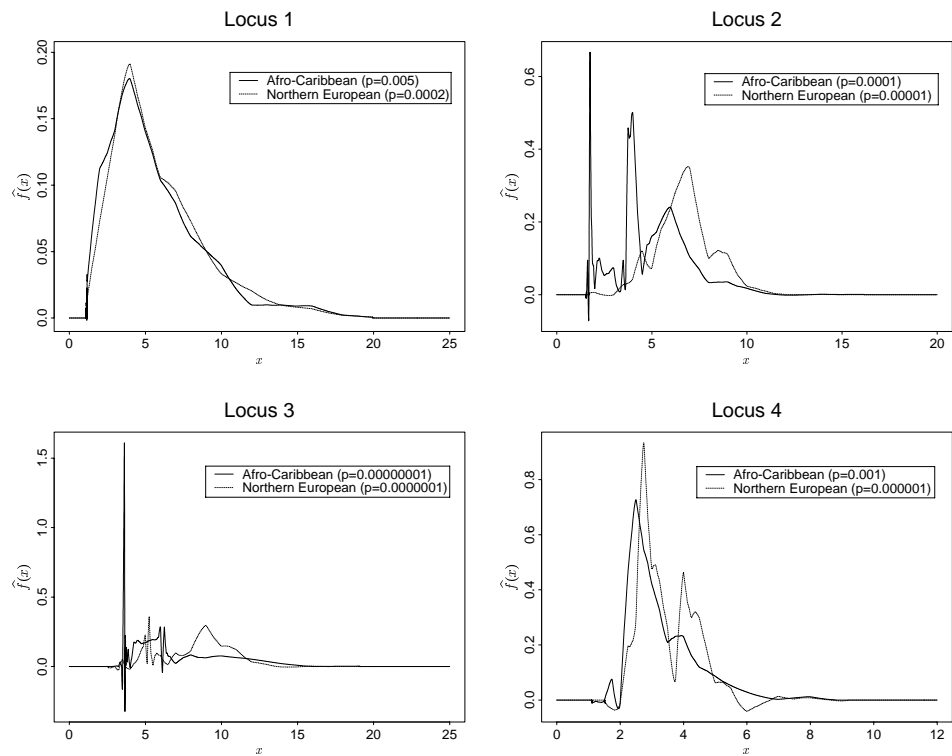


Figure 5. Daubechies least asymmetric wavelet ($N=4$) density estimates of the gene fragment data. The p -values, tuned by eye, are given in the legends.

Gene fragment data. Figure 5 shows density estimates for the gene fragment data using these directly estimated variances in the thresholding. The Daubechies least asymmetric wavelets ($N=4$) were used. As for the Haar estimation, $J=7$ was used. These pictures are quite encouraging, the most striking point being the huge spike on the third locus of the Afro-Caribbean

data, which has survived even the use of a very small p -value. Closer examination reveals that there is a large number of coincidences of observations in the region of this spike, with two values being repeated 23 times each, and several others having multiple repeats. At the spike location about 10% of the entire sample is situated within an interval of length less than 0.1. Thus the large spike is not an artefact but is a good representation of the observed data, and this behaviour is a strength of the wavelet method. See Figure 6 for a close-up of the spike. There is something of an overshoot on each side of the spike with the estimate taking negative values. If the density estimate is to be used for purposes other than diagnostic, the easiest way to deal with this is simply to set the negative values to zero and slightly rescale to maintain the integral of the function being equal to one.

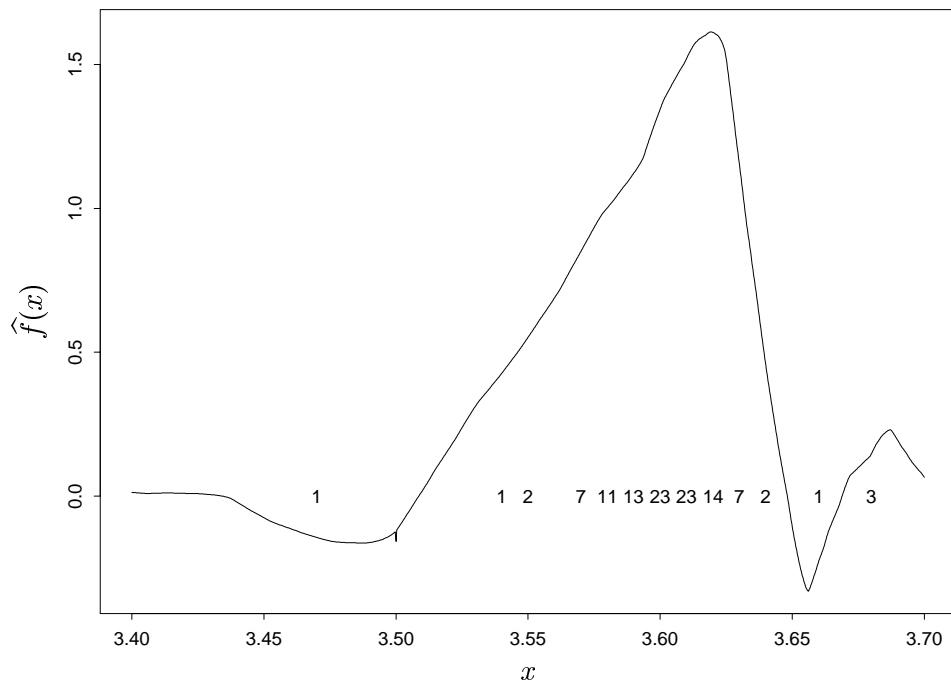


Figure 6. Close up of the spike in the Daubechies least asymmetric wavelet ($N=4$) density estimate of the third locus of the Afro-Caribbean gene fragment data. The numbers indicate the number of coincident observations plotted at their location on the line $y=0$.

6. Investigation using the Claw Density

The various thresholding methods proposed are evaluated using the claw density from Marron and Wand (1992). This is a normal mixture density and is illustrated in figure 7, together with several wavelet density estimates computed from sample sizes of 200.

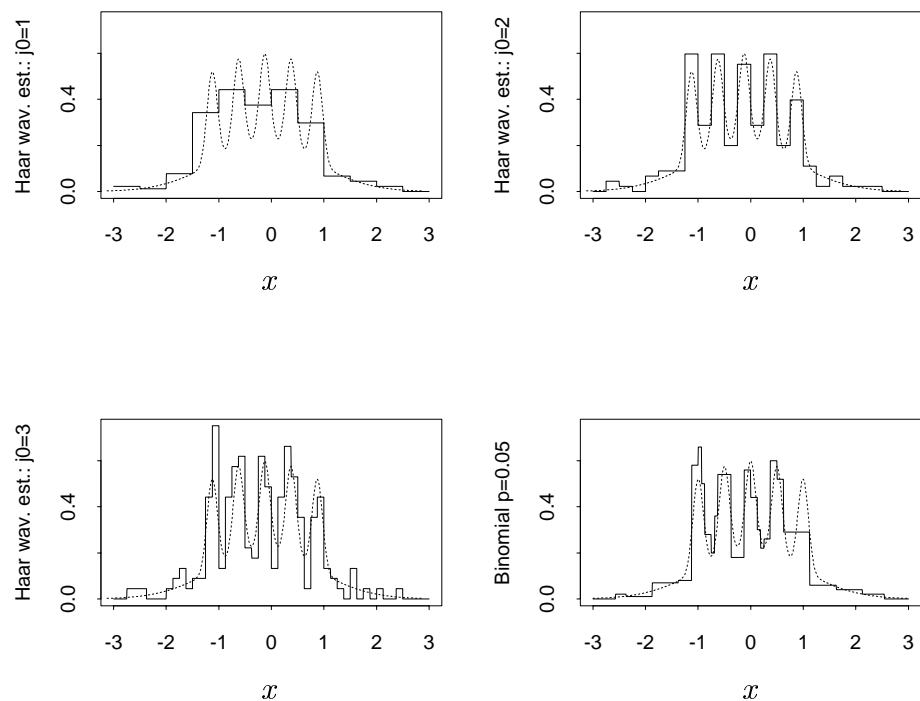


Figure 7. Haar wavelet estimates based on a sample from the claw density of size 200. The top left, top right and bottom left are obtained by truncating the wavelet transform at levels 1, 2 and 3 respectively. The bottom right figure shows the binomial thresholding estimate with $p = 0.1$. True claw density is dotted line, estimates are solid.

The claw density is an extremely demanding example, because the amount of smoothing has to be chosen just right in order to detect the structure; tests such as the Kolmogorov-Smirnov applied to the part of the sample between -1 and 1 are unable to distinguish the middle part of the distribution from uniform.

A simple linear approach is to truncate the Haar wavelet expansion at a particular level j_0 , so that all the \tilde{d}_{jk} for $j > j_0$ are set to zero. As with any linear density estimation method, the choice of the scale of smoothing is critical. Furthermore, because it is now a linear method, the estimates will no longer be adaptive to effects on different scales. The estimate for $j_0 = 1$ is actually quite good at estimating the tails, but fails to give any convincing impression that the underlying density is other than unimodal. For $j_0 = 2$ the claws are very clearly indicated, but to some extent this is because the claw density has lined up perfectly with the wavelet basis. There is spurious variability in the tails, and there is no indication of the shapes of the claws. For $j_0 = 3$ there is considerable spurious variability, especially in the heights of the claws, and in the tails.

An estimate using the binomial method from section 4.2 with $p = 0.1$ is shown. This relatively large p -value was chosen because of the fairly small sample size, in order to give the various hypothesis tests reasonable power. Although one of the claws has been missed, the others are shown more clearly, in that there is an indication of their shape. Furthermore, an indication of the good adaptivity properties of the method is given by the reasonable estimation of the tails of the density.

Acknowledgments. Herrick gratefully acknowledges receipt of an EPSRC Research Studentship. Nason and Silverman were partially supported by EPSRC Research Grant GR/M10229. The authors would like to thank David Balding (University of London & University of Reading) and the Metropolitan Police Forensic Science Laboratory for supplying the data and Brani Vidakovic for suggesting the use of the Daubechies-Lagarias algorithm and making his implementation available to them, which assisted them in writing their own code.

References

- AITKEN, C.G.G. (1995). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Wiley, Chichester.
- ČENCOV, N.N. (1962). Evaluation of an unknown distribution density from observations, *Soviet Mathematics - Doklady*, **3**, 1559–1562.
- DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, Philadelphia.
- DAUBECHIES, I. AND LAGARIAS, J.C. (1992). Two-scale difference equations II. Local regularity, infinite products of matrices and fractals, *SIAM Journal on Mathematical Analysis*, **23**, 1031–1079.
- DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G. AND PICARD, D. (1996). Density Estimation by Wavelet thresholding, *Annals of Statistics*, **24**, 508–539.

- HALL, P. AND PATIL, P. (1995). Formulae for mean integrated squared error of nonlinear Wavelet-based density estimators, *Annals of Statistics*, **23**, 905–928.
- HERRICK, D.R.M. (2000). *Wavelet Methods for Curve and Surface Estimation*. Ph.D. thesis, University of Bristol.
- (2001). Wavelet density estimation software. *Technical Report 01:12*, Department of Mathematics, University of Bristol.
- KOVAC, A. AND SILVERMAN, B.W. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding, *Journal of the American Statistical Association*, **95**, 172–183.
- MARRON, J.S. AND WAND, M.P. (1992). Exact mean integrated squared error, *Annals of Statistics*, **20**, 712–736.
- ROEDER, K. (1994). DNA fingerprinting: A review of the controversy, *Statistical Science*, **9**, 222–278.
- SHEATHER, S.J. AND JONES, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B*, **53**, 683–690.
- SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability, Chapman and Hall, London.
- SIMONOFF, J.S. (1996). *Smoothing Methods in Statistics*. Springer Series in Statistics, Springer-Verlag, New York.
- VANNUCCI, M. AND CORRADI, F. (1999). Covariance structure of wavelet coefficients: Theory and models in a Bayesian perspective, *Journal of the Royal Statistical Society, Series B*, **61**, 971–986.
- VENABLES, W.N. AND RIPLEY, B.D. (1994). *Modern Applied Statistics with S-Plus*, Springer-Verlag, New York.
- VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets*. Wiley, New York.
- WAND, M.P. AND JONES, M.C. (1995). *Kernel Smoothing*. Monographs on Statistics and Applied Probability, Chapman and Hall, London.

D.R.M. HARRICK, G.P. NASON AND B.W. SILVERMAN

DEPARTMENT OF MATHEMATICS

UNIVERSITY OF BRISTOL

UNIVERSITY WALK

BRISTOL

BSS 1TW, U.K.

E-mails: David.Herrick@bristol.ac.uk

G.P.Nason@bristol.ac.uk

B.W.Silverman@bristol.ac.uk