Choice of wavelet smoothness, primary resolution and threshold in wavelet shrinkage

Guy P. Nason*

2nd January 2001

Abstract

This article introduces a fast cross-validation algorithm that performs wavelet shrinkage on data sets of arbitrary size and irregular design and also simultaneously selects good values of the primary resolution and number of vanishing moments.

We demonstrate the utility of our method by suggesting alternative estimates of the conditional mean of the well-known Ethanol data set. Our alternative estimates outperform the Kovac-Silverman method with a global variance estimate by 25% because of the careful selection of number of vanishing moments and primary resolution. Our alternative estimates are simpler than, and competitive with, results based on the Kovac-Silverman algorithm equipped with a local variance estimate.

We include a detailed simulation study that illustrates how our cross-validation method successfully picks good values of the primary resolution and number of vanishing moments for unknown functions based on Walsh functions (to test the response to changing primary resolution) and piecewise polynomials with zero or one derivative (to test the response to function smoothness).

Keywords: cross-validation; Ethanol data; fast wavelet shrinkage updates; Kovac-Silverman algorithm.

1 Introduction

Wavelet shrinkage is a technique for estimating curves in the presence of noise which is appealing because it is nearly minimax for a wide range of functions, computationally practical and spatially adaptive (see the seminal work of Donoho *et al.* (1995)). This paper assumes a familiarity with wavelets and wavelet shrinkage up to the level of Nason and

*Department of Mathematics, University of Bristol, University Walk, BRISTOL, BS8 1TW, UK

Silverman (1994). We also rely heavily on developments in Kovac and Silverman (2000). For a recent survey of research in the area see Vidakovic (1999) or Abramovich, Bailey and Sapatinas (2000).

Most early wavelet shrinkage techniques relied on Mallat's (1989) pyramid algorithm for computing the discrete wavelet transform (DWT) which in its standard form requires data to be equally spaced and contain 2^J values. For this limited data situation wavelet shrinkage works by taking the DWT, then thresholding or shrinking the coefficients, and then taking the inverse transformation. A great deal of research effort has been expended on methods to choose the threshold value of wavelet shrinkage, again see Vidakovic (1999) for an excellent overview. However, the threshold, albeit important, is but one parameter involved in wavelet shrinkage. For successful shrinkage the following criteria need also to be chosen well:

the primary resolution Thresholding of coefficients is applied to coefficients whose resolution level is equal to or finer than the primary resolution. The primary resolution parameter is similar to the usual bandwidth parameter in linear smoothing methods. For wavelet shrinkage choice of the primary resolution was first investigated by Hall and Patil (1995). Hall and Nason (1997) suggest that actually choosing the primary resolution on a continuous scale may be advantageous. However, even if the primary resolution is to be chosen on a discrete scale, as in standard wavelet shrinkage, it is critical to use good values (just as the bandwidth is critical in linear smoothing).

the analysing wavelet. Very little detailed attention has been paid to the problem of which wavelet should be used in wavelet shrinkage. The Daubechies' (1988) series of compactly supported wavelets provide a family of mother wavelets of varying smoothness, ψ_V, where V is the number of vanishing moments and $\psi_V \in C^{\mu V}(\mathbb{R})$ where $\mu \approx 0.2$ and $C^\alpha(\mathbb{R})$ is the space of α times continuously differentiable functions on $\mathbb{R}$. We only consider the Daubechies' extremal phase wavelets in this article ranging from the discontinuous $V = 1$ Haar wavelet to the smooth $V = 10$ wavelet. Let h_V denote the quadrature mirror filter associated with Daubechies' extremal phase wavelet of order V and N_{h_V} be the length of this filter.

The type of wavelet transform is also important. For example, the translation-invariant transform of Coifman and Donoho (1995) often gives better results than using the DWT.

how the threshold is applied Early work described in Donoho *et al.* (1995) considered two methods for applying the threshold to wavelet coefficients, the "keep-or-kill"

hard thresholding (similar to model selection) and "shrink-or-kill" soft thresholding. Other techniques have been suggested such as the firm thresholding of Gao and Bruce (1997) and more recently Bayesian wavelet shrinkage which also has a thresholding interpretation, see Chipman *et al.* (1997) or Abramovich *et al.* (1998).

This article proposes a fast-update cross-validation method which aims to find good combinations of the threshold, number of vanishing moments (wavelet smoothness), V, and primary resolution parameters. Often there is no unique "optimal" combination as the parameters interact and sometimes quite different combinations give similar results. The cross-validation method described below could be extended to incorporate different choices of threshold application and type of wavelet transform but it is not clear that such choices could be implemented in a fast-update algorithm. Cross-validation for threshold selection in wavelet shrinkage was proposed, especially for functions sparsely represented by wavelets, by Nason (1996).

Generalized cross-validation for wavelet shrinkage was proposed by Jansen, Malfait and Vial (1997) and other cross-validation techniques were proposed by Wang (1996) and Weyrich and Warhola (1998) but of course cross-validation as a general technique has been around for a very long time; see Stone (1974) for further details. Recently, papers such as Hall and Turlach (1997), Sardy *et al.* (1999) and Kovac and Silverman (2000) have adapted the wavelet shrinkage methodology to data sets with arbitrary size and irregular design. See also Antoniadis, Grégoire and Vial (1997) and Antoniadis and Pham (1998) for work on fast *linear* wavelet methods for random designs. We should also mention that recent algorithms based on the lifting transform show great promise for curve and surface estimation for irregular data (see for example, Daubechies *et al.* (1999) for an excellent review).

Our cross-validatory method is a development of Kovac and Silverman (2000) and as such works with data sets of irregular design and arbitrary size but ours additionally gives useful information on which wavelet and primary resolution to use as well as choosing the threshold value.

1.1 Wavelet shrinkage and the Kovac-Silverman algorithm

First we establish some notation and describe the data model specified by Kovac and Silverman (2000). We suppose that $f(x)$ for $x \in (0,1)$ is the function that we wish to estimate and that we observe n data points $g(x_i)$ according to the model

$$g(x_i) = f(x_i) + \epsilon_i \tag{1}$$

where $\{\epsilon_i\}_{i=0}^{n-1}$ is i.i.d. noise with mean zero and variance σ^2 and $\{x_i \in (0,1)\}_{i=0}^{n-1}$ are not necessarily equally spaced.

Kovac and Silverman (2000) propose choosing a new equally spaced grid $t_0, \ldots, t_{N-1}$ on $(0,1)$ where $N = 2^J$ for some $J \in \mathbb{N}$ and interpolate the observed data onto the new grid. They propose choosing $t_k = (k+0.5)/N$ for $k = 0, \ldots, N-1$ and choose $N(= 2^J)$ such that $J = \min\{j \in \mathbb{Z} : 2^j > n\}$. Throughout their article they linearly interpolate the original data to new values, y_k, on the grid by

$$y_k = \begin{cases} g_0 & \text{if } t_k \leq x_0 \\ g_i + (t_k - x_i)\frac{g_{i+1}-g_i}{x_{i+1}-x_i} & \text{if } x_i \leq t_k \leq x_{i+1} \\ g_{n-1} & \text{if } t_k \geq x_{n-1}, \end{cases} \tag{2}$$

where $g_i = g(x_i)$ and $k = 0, \ldots, N-1$, although they admit that higher order interpolants or other reweighting schemes might also be of some use. Writing the original and interpolated data as vectors $\mathbf{y} = (y_0, \ldots, y_{N-1})$ and $\mathbf{g} = (g_0, \ldots, g_{n-1})$ the linear transform described by (2) can be written in matrix form by

$$\mathbf{y} = R\mathbf{g},$$

where the interpolation matrix R depends on $\mathbf{t}$ and $\mathbf{x}$. Each row of R always contains either one or two non-zero entries which always sum to one. Interpolation to a grid is a useful technique but certainly not new see, for example, Jones and Lotwick (1983) or Silverman (1986). Kovac and Silverman (2000) then apply wavelet shrinkage to the interpolated data, $\mathbf{y}$, which first involves taking the DWT by

$$\mathbf{w} = W_V\mathbf{y},$$

where W_V is the $N \times N$ orthogonal matrix associated with Daubechies' extremal phase wavelet with V vanishing moments (in practice Mallat's fast algorithm is used but the matrix multiplication representation is mathematically convenient).

Given model (1) above and in particular the i.i.d. assumption on the noise the covariance matrix, Σ_Y of the interpolated data is given by

$$\Sigma_Y = \sigma^2 RR^T.$$

Kovac and Silverman (2000) exploit the fact that for the linear interpolation scheme described above Σ_Y is actually a band matrix. After applying the DWT to $\mathbf{y}$ Kovac and Silverman (2000) show that the variances of the individual wavelet coefficients can be

computed exactly, up to knowledge of σ^2 which has to be estimated from the data, using a fast algorithm of computational order $\mathcal{O}(b2^J)$ where $b = \max(b_Y, N_{h_V})$ and b_Y is the bandwidth of the matrix Σ_Y. A useful consequence of Kovac and Silverman's work is that the variance of all the wavelet coefficients can be computed with no more effort than computing the wavelet coefficients themselves, which is $\mathcal{O}(N_{h_V} 2^J)$ and also fast. Vannucci and Corradi (2000) also present a fast algorithm to compute the variance-covariance matrix of the wavelet coefficients and link it to the two-dimensional DWT.

Kovac and Silverman (2000) show how knowledge of the wavelet coefficient variances permits extension of the universal and SURE thresholds of Donoho and Johnstone (1994; 1995) to their interpolated data situation. We are also interested in choosing the threshold, but by cross-validation, and also simultaneously selecting good values for the number of vanishing moments, V, and primary resolution, p. The next section shows that it is possible to apply full leave-one-out cross-validation to the Kovac-Silverman set-up and still retain a fast algorithm.

2 Leave-one-out cross-validation

Cross-validation is a well-established technique for assessing model prediction error and, in our situation, selecting good choices of the threshold, number of vanishing moments, and primary resolution. In the following sections we describe how to obtain a leave-one-out estimate of the prediction error. That is, our wavelet shrinkage estimator, $\hat{f}_{t,V,p}(x)$ with threshold t, number of vanishing moments, V, and primary resolution p aims to minimize the mean integrated squared error (MISE)

$$M(t,V,p) = \mathbb{E} \int_0^1 \left\{ \hat{f}_{t,V,p}(x) - f(x) \right\}^2 dx.$$

With the methodology below we could easily choose another form of loss function. In particular, with wavelet shrinkage, we might be interested in doing better near known discontinuities or inhomogeneities for example. However, using MISE for now, we can estimate M by

$$\hat{M}(t,V,p) = n^{-1} \sum_{i=0}^{n-1} \left\{ \hat{f}^{-i}_{t,V,p}(x_i) - g_i \right\}^2,$$

where $\hat{f}^{-i}_{t,V,p}$ is an estimate of f constructed from all the data except the ith point. To find good values of (t,V,p) we minimize $\hat{M}$. However, we do not believe that the first minimizer we come across is in any sense "optimal". Unlike, say, the cross-validation score developed in Nason (1996) the score $\hat{M}$ has multiple minima and as many as possible

should be investigated further. The next few sections describe the construction and efficient computation of the leave-one-out predictor $\hat{f}^{-i}_{t,V,p}(x_i)$ which estimates $f(x_i)$ using all the original data points apart from (x_i, g_i). The key to the efficiency is that removing the ith original data point only changes grid points and thus only wavelet coefficients local to x_i.

2.1 Leave-one-out and interpolation

Removing the ith original data point only has a very local effect on the interpolated data points because with linear interpolation only those grid points that lie in an interval with the ith point as one of the end points are affected. For these points the interpolated point t_k is either to the left or the right of x_i. If $i = 0$ and x_0 is removed (or $i = n - 1$ and x_{n-1} is removed) then new grid points to the left of x_1 are updated to take the value g_1 (or those to the right of x_{n-2} take the value g_{n-2}).

Assume that $x_{i-1} \leq t_k \leq x_i$. We can compute the value of the updated interpolated points $\bar{y}_k$ from the old ones y_k by the simple formula

$$\bar{y}_k = y_k + (t_k - x_{i-1})L_i,$$

where

$$L_i = \frac{g_{i+1} - g_{i-1}}{x_{i+1} - x_{i-1}} - \frac{g_i - g_{i-1}}{x_i - x_{i-1}}.$$

For each removed point, L_i only has to be computed once and only $\bar{y}_k$ local to the ith original point have to be recomputed. If N is chosen well then only a few y_k need to be updated. We record the indices of those t_k whose y_k value has been updated and pass them onto the next stage.

2.2 Updating the wavelet transform

The previous section tells which of the $y_k, k = 0, \ldots, N-1$ have changed. The Mallat (1989) DWT algorithm is a recursive algorithm which takes the $\{y_k\}_{k=0}^{N-1}$ as input and computes coarser versions (called father wavelet coefficients) and detail coefficients between successive levels of coarse coefficients. For the father wavelet coefficients the formula for computing a coarser approximation to the data c^{j-1} from a finer approximation c^j is given by

$$c^{j-1}_\ell = \sum_k h_V(k - 2\ell)c^j_k. \tag{3}$$

The finest scale approximation of $c^J_k = y_k$ initializes the algorithm and then coarser approximations, $c^{J-1}, c^{J-2}, \ldots, c^1, c^0$ are generated using successive applications of (3).

The mother wavelet coefficients, d^{j-1} represent the detail lost when moving from a finer scale c^j to a coarser scale c^{j-1} and are computed by a similar formula to (3) except that the smoothing filter h_V is replaced by a "detail extraction" filter g_V, we refer to Nason and Silverman (1994) or Vidakovic (1999) for further details.

The key point for efficiency is that changing a single y_k only affects those wavelet coefficients which are derived from the y_k and only those coefficients local to t_k will be changed. In summary, changing a y_k changes very few wavelet coefficients.

More specifically, if the single father wavelet coefficient c_k^j is changed then only the coarser father wavelet coefficients c_ℓ^{j-1} where

$$\left\lceil \frac{k - N_{h_V} + 1}{2} \right\rceil \leq \ell \leq \left\lfloor \frac{k}{2} \right\rfloor \tag{4}$$

need to be recomputed (here $\lceil x \rceil$ is the smallest integer greater than or equal to x, and $\lfloor x \rfloor$ is the largest integer less than or equal to x). Recall that the DWT is recursive starting with the $\{y_k\}_{k=0}^{N-1}$ as the input so formula (4) shows which coefficients need to be recomputed at a coarser resolution level and then supplies the indices of those changed recursively to the same routine for the next coarsest level.

Further efficiency gains can be achieved by noting the range of the changed c_k^j coefficients and only recomputing those coarser c^{j-1} that are involved. For example, if c_k^j has changed for $k_{\text{min}} \leq k \leq k_{\text{max}}$ then we only need to recompute c_ℓ^{j-1} for

$$\left\lceil \frac{k_{\text{min}} - N_{h_V} + 1}{2} \right\rceil \leq \ell \leq \left\lfloor \frac{k_{\text{max}}}{2} \right\rfloor$$

in other words $\frac{1}{2}(k_{\text{max}} - k_{\text{min}} + N_{h_V} + 1)$ coarser coefficients at resolution level $j-1$ need to be computed from $k_{\text{max}} - k_{\text{min}} + 1$ coefficients at level j. Since each coefficient computation takes $\mathcal{O}(N_{h_V})$ the recursive update of wavelet coefficients is effectively $\mathcal{O}(N_{h_V} J)$ and hence extremely fast (i.e. effectively $\mathcal{O}(1)$ with respect to N and n). Note that a similar algorithm can be developed for the DWT inversion.

2.3 Updating the wavelet coefficient variance factors

Given the covariance matrix Σ_Y of $\mathbf{y}$ the wavelet coefficients' covariance matrix is given by $W_V \Sigma_Y W_V^T$. Removing the ith point alters the covariance matrix $\Sigma_Y = RR^T$ because the $N \times n$ matrix R changes to a $N \times (n-1)$ matrix $\bar{R}$. Let $\mathbf{r}_k$ denote the kth row of R. Define $\mathcal{R}_i = \{k : \mathbf{r}_k \text{ has a non-zero entry at position } i\}$. Then rows not in $\mathcal{R}_i$ in both R and $\bar{R}$ will be the same except that those in $\bar{R}$ will be one entry missing where there was a zero in R. However, rows $k \in \mathcal{R}_i$ will be different in R and $\bar{R}$. Therefore, the difference

D_i between Σ_Y and $\bar{\Sigma}_Y = \bar{R}\bar{R}^T$ will be zero apart from a cross-shaped region where any row or column of D_i in $\mathcal{R}_i$ can, in general, be non-zero. However, since both Σ_Y and $\bar{\Sigma}_Y$ are band matrices most of the entries in the cross-shaped region in D_i, except those close to the main diagonal (less than b away from the main diagonal) will be zero.

Summarizing we can compute the covariance matrix $\bar{\Sigma}_Y$ for the interpolated data from Σ_Y using

$$\bar{\Sigma}_Y = \Sigma_Y + D_i,$$

where D_i is almost all zero apart from some of the rows/columns near the main diagonal and in $\mathcal{R}_i$. To compute the wavelet coefficient variance for the new interpolated data we only need to consider $W_V D_i W_V^T$ since we already know $W_V \Sigma_Y W_V^T$ from the Kovac-Silverman algorithm. Computation of $W_V D_i W_V^T$ can be easily performed using the updating wavelet transform described in the previous section since multiplication by W or W^T is simply an application of the DWT. Since most of the entries in rows or columns of D_i are zero the updating algorithm can be executed first with a zero transform, and then with those non-zero entries in each row/column of D_i. Again application of such a transform is extremely fast: $\mathcal{O}(N_{h_V} J)$.

2.4 Thresholding, inversion and optimisation

Using the above information about which wavelet coefficients' variance change one can identify those coefficient positions where the quantity $\tau_{jk} = \hat{d}_{jk}/\hat{\sigma}_{jk}$ has changed (where $\hat{\sigma}_{jk}^2 = \hat{\sigma}^2 \gamma_{jk}$ are the updated variances of the wavelet coefficients, γ_{jk} are the variance factors as in Kovac and Silverman (2000), $\hat{\sigma}^2$ is some estimate of σ^2 and $\hat{d}_{jk}$ are the updated empirical coefficients.) In wavelet shrinkage an estimate of σ is typically computed by using a robust estimator such as the median absolute deviation (MAD) of the wavelet coefficients at the finest level, divided by 0.6745. The estimate $\hat{\sigma}$ can itself be updated quickly by keeping track of which coefficients in the finest resolution level change.

For thresholding after point removal we only need note which τ_{jk} have changed. If all of those τ_{jk} that changed were previously thresholded, and if after point i removal they are subsequently all smaller in absolute size than the threshold then the estimate does not change. This means that inversion does not have to be performed and the prediction error simply taken from the non-removed point estimate. However, if any τ_{jk} have changed their status since last time the estimate $\hat{f}_{t,V,p}(x_i)$ has to be recomputed using the efficient inverse algorithm described in section 2.2.

For optimisation we have found that a grid search algorithm works extremely effectively for finding minimal values of $\hat{M}$. We also have used a golden section search method but

Table 1: Table showing values of $\hat{M}$ ($\times 1000$ to 3 s.f.) with $t \approx 3.11$ fixed for various values of the primary resolution, p, and number of vanishing moments, V.

	Number of vanishing moments, V									
p	1	2	3	4	5	6	7	8	9	10
0	328	204	228	192	260	167	195	117	133	217
1	320	160	138	196	165	167	143	127	133	160
2	221	165	112	151	130	132	152	98	133	126
3	148	102	122	133	115	131	129	107	118	116
4	154	102	127	131	119	119	134	124	109	103
5	142	121	121	126	129	124	117	115	111	102
6	135	138	138	127	130	139	143	142	135	121

this tends to get stuck in one of the multiple minima. Another strategy that we adopt is to condition on the universal threshold $t^* = \sqrt{2 \log N}$ and then optimise $\hat{M}(t^*, V, p)$ to find good values of (V, p). Then using the good values of (V, p) we optimise over the threshold t. This strategy is effective in practice because the universal threshold makes a useful starting value for the optimiser as its value is independent of (V, p).

3 Example: the ethanol data

Before we describe a simulation study we present an applied example in detail. The well-studied `ethanol` data from Brinkman (1981) has been analysis by Cleveland *et al.* (1993) and Hastie (1993) but more importantly for our purposes by Kovac and Silverman (2000). The data consist of $n = 88$ measurements from an experiment where ethanol was burned in a single cylinder engine. The concentration of the total amount of nitric oxide and nitrogen dioxide in engine exhaust, normalized by work done by the engine is related to the "equivalence ratio", a measure of the richness of the air ethanol mixture. Note that the range of the x-axis or "Equivalence ratio" variable is $(0.535, 1.232)$ so this was linearly shifted to $(0, 1)$. The `ethanol` data are plotted in the top left-hand corner of Figure 1.

For the purposes of this example we fixed the threshold value t to be equal to the universal threshold value of Donoho and Johnstone (1994). The default Kovac-Silverman method chooses $N = 128$ equally-spaced grid points to interpolate the original data. Thus the universal threshold computed was $t = \sqrt{2 \log N} \approx 3.11$. With this threshold fixed Table 1 shows our computed value of $\hat{M}$ for all values of the primary resolution ranging from 0 to 6 and for numbers of vanishing moments (smoothness) ranging from 1 to 10 from the Daubechies "extremal phase" series. From the table it is clear that the lowest value of $\hat{M}$ occurs for $V = 8, p = 2$ where $\hat{M} = 98$. However, $V = 10, p = 5$ and $V = 2$, $p = 3$ or 4 might also be of interest (one could continue to higher values of $\hat{M}$).

Next, we conditioned on the three different pairs of (V, p) and optimized over the

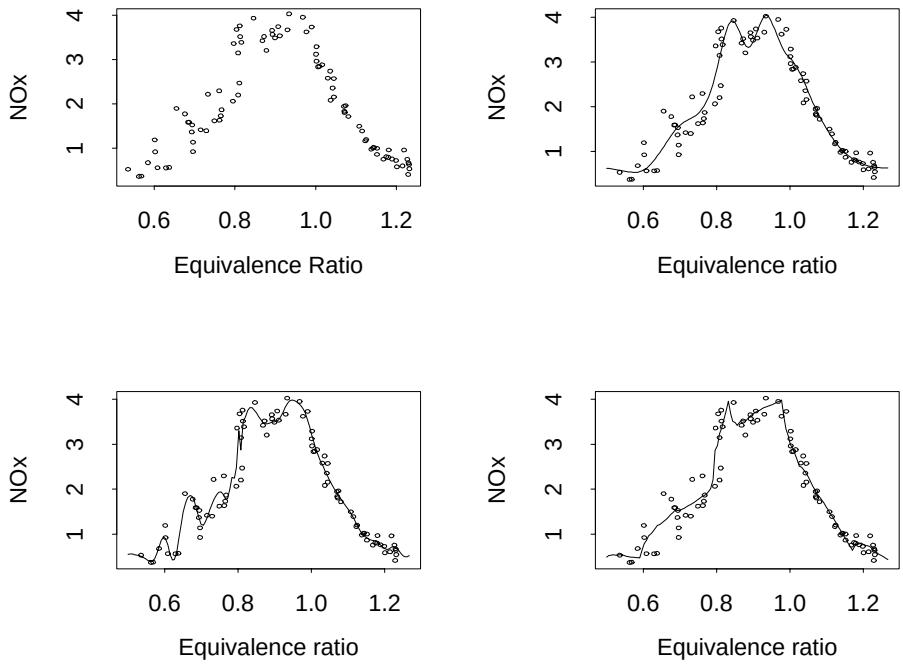

Figure 1: Top left: ethanol data with `NOx` emission versus `Equivalence ratio`. Clockwise from top right: estimates $\hat{f}_{t,V,p}(x)$ for (t, V, p) equal to $(3.11, 8, 2)$, $(3.13, 10, 5)$ and $(2.88, 2, 3)$.

threshold value to minimize $\hat{M}$. In the cases $(8, 2)$, $(10, 5)$ and $(2, 3)$ the minimizing thresholds were 3.11, 3.13 and 2.88 all of which are not actually too far from the universal threshold. The estimated curves $\hat{f}_{t,V,p}(x)$ are shown in Figure 1. The top-right plot in Figure 1 shows our "best" estimate for the underlying curve.

Kovac and Silverman (2000) use Daubechies' "extremal phase" wavelet with $V = 5$ vanishing moments and a primary resolution of 3. Referring to Table 1 again one can see that in terms of minimizing M this combination is only ranked 31st out of all the 60 combinations tried. Therefore we obtain an approximate 25% improvement on Kovac and Silverman by using our best combination. The improvement using our search method is also demonstrated by comparing the plot for our best estimate and their two best in their bottom row in Figure 1 of Kovac and Silverman (2000). Both of their plots use a (V, p) combination of $(5, 3)$ with the universal threshold. Their bottom-left uses a *global* estimate of variance, their bottom-right uses a *local* estimate of variance by noticing that the variance of the `NOx` variable decreases with `Equivalence ratio`. Their use of the more complex local variance estimate is motivated by the fact that their estimate with the global variance contains a small spike at about 0.8 (like the one in our bottom-left plot of

Figure 1). However, after examining our top-right plot of Figure 1 which tracks the double bump and does not have a spike at 0.8 we claim that actually Kovac and Silverman (2000) need not go to the trouble of forming a local variance estimate but merely need change the number of vanishing moments V to 8 and their primary resolution to 2. We repeated the Kovac-Silverman (2000) analysis with the new number of vanishing moments and primary resolution and the resulting estimate is significantly better than theirs and looks more like our top-right plot in Figure 1 (note it is not exactly the same since they use soft thresholding and choose t to be exactly the universal threshold, whereas we use hard thresholding and optimize the value of t).

In no way are we trying to denigrate Kovac and Silverman (2000). Indeed, this paper is based on their extremely useful methodology. However, we have used the above example to stress that choice of number of vanishing moments and primary resolution is extremely important, probably as important as choice of threshold but considerably neglected by much of the literature and available software.

4 Simulation study

We performed several simulations that show that once the number of vanishing moments and primary resolution are correctly selected that our cross-validation method produces broadly similar results to `GlobalSURE` type thresholding (and also universal thresholding, although the goal of universal thresholding is not MISE minimisation). The `GlobalSURE` thresholding method was introduced by Nason (1996) and is a single threshold version of the level-dependent technique based on Stein's unbiased risk estimation for wavelet shrinkage introduced by Donoho and Johnstone (1995).

However, as the `ethanol` example demonstrates in the previous section it is important to show that methodology can adapt to features such as the smoothness of the underlying function or the scale of the features (which require particular choices of numbers of vanishing moments or primary resolution). The first simulation ("adapting to primary scale") concentrates on choice of primary resolution, the next three simulations ("adapting to smoothness") concentrate on choice of number of vanishing moments.

The "adapting to smoothness" simulations fix the threshold to be the universal threshold using sample sizes of $n = 100, 200$ and 500 and for each sample size perform 10 simulations to find the best combinations of number of vanishing moments and primary resolution. In each case we use Gaussian noise with zero mean and the variance, σ^2, is specified in each section below. The "adapting to smoothness" simulations demonstrate how the primary resolution is most influenced by discontinuities in lowest-order deriva-

tives and that the number of vanishing moments chosen by the cross-validation algorithm is influenced by the underlying smoothness of the true function. However, these results are not hard and fast and occasionally the cross-validation technique gets it wrong.

4.1 Adapting to primary scale

The underlying true function for this simulation is the Walsh function $W(p, x)$ defined on $x \in (0, 1)$ which is a piecewise constant function taking only the values 0 or 1 starting at 0 and then switching to 1 and then back to 0 and so on. The number of switches in the interval is parametrised by p and the distance between each switch is $1/p$. A convenient way to think about Walsh functions is as a sine wave that has been "blocked"! The parameter p is more formally known as the *sequency* number of the Walsh function and it is akin to the frequency parameter of a sine wave, but not exactly the same as the Walsh function is not always periodic on $(0, 1)$. See Stoffer (1991) for further information on Walsh functions and their applications in statistics.

Table 2 demonstrates that the selected primary resolution increases with the fineness of the true Walsh function although there appears to be quite some variability in the selected primary resolution values at $p = 2\frac{2}{3}$. However, notice that for (e.g.) $p = 8 \cdot 2\frac{2}{3}$ the width of the Walsh peaks is $3/64$ and so a primary resolution matching this of 4 or 5 (corresponding to nearest widths of 8/256 or 4/256 of the Haar wavelet at this resolution level) might have been expected. However, our algorithm chooses primary resolution of 7, and indeed the other primary resolutions also "over-estimate" in this way. This effect is presumably because of the addition of noise which causes the procedure to be conservative and use finer scale wavelets. Conceptually, the best wavelet basis for representing this set of Walsh functions should be the Haar basis. Table 3 shows that our cross-validation method nearly always selects Haar to be the best basis in this situation.

The ten simulations in Tables 2 and 3 are based on sample sizes of $n = 200$ with noise variance of $\sigma^2 = 0.001^2$. The deliberately large signal to noise ratio in this simulation is to verify that in low-noise situations the cross-validation procedure chooses reasonable values of the primary resolution and number of vanishing moments. Clearly as the signal to noise ratio decreases our procedure will choose the "best" values far less often. We leave to further work the interesting questions: for which values of the signal to noise ratio does it become very hard to select good parameters and whether competitors such as SURE can do a better job.

Table 2: Best primary resolution for Walsh function $W(p, x)$ at sequency number $p = q2^{\frac{2}{3}}$.

	Simulation Number									
q	1	2	3	4	5	6	7	8	9	10
1	1	0	0	1	2	6	0	3	3	2
2	5	4	6	4	4	4	4	4	4	7
4	6	6	6	6	6	6	6	7	6	6
8	7	7	7	7	7	7	7	7	7	7

Table 3: Best number of vanishing moments for Walsh function $W(p, x)$ at sequency number $p = q2^{\frac{2}{3}}$.

	Simulation Number									
q	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	4	1	1
8	1	1	1	1	1	1	8	1	1	1

4.2 Adapting to smoothness: no derivatives

The underlying true function for this example is

$$f_0(x) = \begin{cases} x & \text{for } x \in [0, \frac{1}{2}), \\ 1 - x & \text{for } x \in [\frac{1}{2}, 1] \\ 0 & \text{elsewhere.} \end{cases}$$

This function is continuous on $[0, 1]$ but is not differentiable everywhere (the first derivative has a discontinuity at $\frac{1}{2}$). The variance of the noise for these simulations $\sigma^2 = 0.01^2$. The best primary resolutions and numbers of vanishing moments for each simulation/sample size combination are shown in Tables 4 and 5. Over all sample sizes the modal number of vanishing moments is 3 and the associated approximate wavelet smoothness is $3\mu = 3 \times 0.2 = 0.6$. However, at smaller sample sizes the wavelet with 7 vanishing moments is selected 4 times out of 10, as many times as the $V = 3$ wavelet. The modal primary resolution appears to be 4 but $p = 3$ is also often chosen.

Table 4: Best primary resolution for function with no derivatives at different sample sizes.

	Simulation Number									
Sample Size	1	2	3	4	5	6	7	8	9	10
100	3	4	3	3	3	4	4	4	2	4
200	3	2	3	4	4	3	4	4	5	3
500	4	4	4	4	4	4	4	5	3	4

Table 5: Best number of vanishing moments for function with no derivatives at different sample sizes.

	Simulation Number									
Sample Size	1	2	3	4	5	6	7	8	9	10
100	7	3	3	2	3	8	7	7	7	3
200	7	3	3	2	6	3	3	3	9	3
500	2	3	3	7	2	3	3	9	2	7

Table 6: Best primary resolution for function with one derivative at different sample sizes.

	Simulation Number									
Sample Size	1	2	3	4	5	6	7	8	9	10
100	3	2	2	2	2	2	2	0	2	2
200	2	2	2	2	2	3	2	2	2	2
500	2	2	2	2	4	2	2	4	2	2

4.3 Adapting to smoothness: one derivative

The underlying true function for this example is

$$f_1(x) = \begin{cases} \frac{x^2}{2} & \text{for } x \in [0, \frac{1}{4}), \\ -\frac{x^2}{2} + \frac{x}{2} - \frac{1}{16} & \text{for } x \in [\frac{1}{4}, \frac{3}{4}), \\ \frac{x^2}{2} - x + \frac{1}{2} & \text{for } x \in [\frac{3}{4}, 1] \\ 0 & \text{elsewhere.} \end{cases}$$

This function is continuous on $[0, 1]$, its first derivative is continuous on $[0, 1]$ but the first derivative is not differentiable everywhere (the second derivative has discontinuities at $\frac{1}{4}$ and $\frac{3}{4}$). The variance of the added noise for these simulations was $\sigma^2 = 0.001^2$. The best primary resolutions and numbers of vanishing moments for each simulation/sample size combination are shown in Tables 6 and 7. Over all sample sizes the modal number of vanishing moments is 9 and the associated approximate wavelet smoothness is approximately $9\mu = 9 \times 0.2 = 1.8$. However, at smaller sample sizes the wavelet with 10 vanishing moments is selected 5 times out of 10. It is interesting to note that the smoothness of the wavelets selected for the example with one derivative is approximately twice that in the example with no derivatives, which is what one would expect. The modal primary resolution

Table 7: Best number of vanishing moments for function with one derivative at different sample sizes.

	Simulation Number									
Sample Size	1	2	3	4	5	6	7	8	9	10
100	8	10	9	10	10	9	10	8	10	9
200	9	9	9	9	9	9	10	9	9	9
500	9	9	9	9	9	9	9	10	9	9

Table 8: Best primary resolution for function with mixed derivatives at different sample sizes.

	Simulation Number									
Sample Size	1	2	3	4	5	6	7	8	9	10
100	4	4	4	3	4	4	4	4	4	5
200	4	4	4	4	5	4	4	4	4	4
500	5	4	5	5	5	5	5	5	5	5

Table 9: Best number of vanishing moments for function with mixed derivatives at different sample sizes.

	Simulation Number									
Sample Size	1	2	3	4	5	6	7	8	9	10
100	8	8	8	8	7	7	3	8	4	8
200	7	7	10	8	3	8	7	7	7	4
500	6	3	3	7	3	7	6	10	3	8

is 2.

4.4 Adapting to smoothness: mixed derivatives

The underlying true function for this example mixes the two functions, $f_0(x)$ and $f_1(x)$ from the previous two examples.

$$f_{\text{mixed}}(x) = f_0(2x) + f_1(2x-1).$$

This function is continuous on $[0,1]$ but the first derivative has a discontinuity at $\frac{1}{4}$ and the second derivative has a discontinuities at $\frac{5}{8}$ and $\frac{7}{8}$. The variance of the added noise for these simulations was $\sigma^2 = 0.01^2$. The best primary resolutions and numbers of vanishing moments for each simulation/sample size combination are shown in Tables 8 and 9. The modal primary resolution is 4 for the smaller sample sizes (agreeing with the primary resolution for the "no" derivative case in section 4.2) but 5 for the $n = 500$ sample size. The primary resolution is most greatly influenced by the lowest-order derivative, as one might expect from the work of Hall and Patil (1995). The number of vanishing moments at the $n = 500$ sample size is 3 for four simulations and around 6/7 for most of the others. At lower sample sizes the number of vanishing moments is generally larger than the ones selected for the $f_0(x)$ example, but not as high as for the $f_1(x)$ example which indicates that maybe some sort of compromise is being made.

5 Conclusions and further work

In this article we have introduced a fast cross-validation method that performs wavelet shrinkage on data sets of irregular design and arbitrary size and also selects good values of the number of vanishing moments and primary resolution.

Our cross-validation method has been shown to work well on the `Ethanol` data set and on simulated data where the scale (primary resolution) and smoothness (vanishing moments) of the underlying true function can be controlled. Further work could be performed to investigate the conditions under which our method would break down both in terms of diminishing signal to noise ratio and in non-Gaussian and correlated noise situations. Our method could easily be extended to use level-dependent thresholds which would be of use with correlated data. It would also be interesting to see how well a MISE estimator such as `GlobalSURE` would perform in place of the cross-validation estimate.

Herrick (1999) uses cross-validation with the Kovac-Silverman (2000) algorithm in the two-dimensional case, although not using the fast version described above and as such the implementation is slow. It remains to be seen whether a fast version is plausible: the fast point insertion/deletion techniques of Green and Sibson (1978) for the Voronoi diagram/Delaunay triangulation used for data interpolation would certainly be of value.

Acknowledgments

The author would like to particularly thank Arne Kovac for helpful discussions and access to his code. He would like to thank Bernard Silverman, David Herrick, others in the Bristol Statistics Group and the audience of the wavelets contributed papers session at the 1999 ISI Helsinki meeting for helpful conversations and suggestions. The author would also like to thank the referees and Editors for supplying extremely helpful comments and suggestions.

References

Abramovich, F., Bailey, T.C., & Sapatinas, T. 2000. Wavelet analysis and its statistical applications. *J. R. Statist. Soc. D*, **49**, 1–29.

Abramovich, F., Sapatinas, T., & B.W., Silverman. 1998. Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society Series B*, **60**, 725–749.

Antoniadis, A., Grégoire, G., & Vial, P. 1997. Random design wavelet curve smoothing. *Statistics and Probability Letters*, **35**, 225–232.

Antoniadis, A., & Pham, D.T. 1998. Wavelet regression for random or irregular design. *Computational Statistics and Data Analysis*, **28**, 353–369.

Brinkman, N.D. 1981. Ethanol – a single-cylinder engine study of efficiency and exhaust emissions. *SAE Transactions*, **90**, 1410–1424.

Chipman, H.A., Kolacyzk, E., & McCulloch, R. 1997. Adaptive Bayesian Wavelet Shrinkage. *Journal of the American Statistical Association*, **92**, 1413–1421.

Cleveland, W.S., Grosse, E., & Shyu, W.M. 1993. Local Regression Models. *Pages 309–376 of:* Chambers, J.M., & Hastie, T.J. (eds), *Statistical Models in S.* Pacific Grove, California: Chapman and Hall.

Coifman, R.R., & Donoho, D.L. 1995. Translation-invariant de-noising. *Pages 125–150 of:* Antoniadis, A., & Oppenheim, G. (eds), *Lecture Notes in Statistics*, vol. 103. Springer-Verlag.

Daubechies, I. 1988. Orthonormal bases of compactly supported wavelets. *Communications in Pure and Applied Mathematics*, **41**, 909–996.

Daubechies, I., Guskov, I., Schröder, P., & Sweldens, W. 1999. Wavelets on irregular point sets. *Philosophical Transactions of the Royal Society, A*, **357**, 2397–2413.

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

Donoho, D.L., & Johnstone, I.M. 1995. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, **90**, 1200–1224.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., & Picard, D. 1995. Wavelet shrinkage: asymptopia? (with discussion). *Journal of the Royal Statistical Society Series B*, **57**, 301–337.

Gao, H.Y., & Bruce, A.G. 1997. WaveShrink with firm shrinkage. *Statistica Sinica*, **7**, 855–874.

Green, P.J., & Sibson, R. 1978. Computing Dirichlet tesselations in the plane. *The Computer Journal*, **21**, 168–173.

Hall, P., & Nason, G.P. 1997. On choosing a non-integer resolution level when using wavelet methods. *Statistics and Probability Letters*, **34**, 5–11.

Hall, P., & Patil, P. 1995. Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Annals of Statistics*, **23**, 905–928.

Hall, P., & Turlach, B.A. 1997. Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *Annals of Statistics*, **25**, 1912–1925.

Hastie, T.J. 1993. Generalized Additive Models. *Pages 309–376 of:* Chambers, J.M., & Hastie, T.J. (eds), *Statistical Models in S*. Pacific Grove, California: Chapman and Hall.

Herrick, D.R.M. 1999. *Wavelet methods for curve and surface estimation*. Ph.D. thesis, Department of Mathematics, University of Bristol.

Jansen, M., Malfait, M., & Bultheel, A. 1997. Generalised cross validation for wavelet thresholding. *Signal Processing*, **56**, 353–369.

Jones, M.C., & Lotwick, H.W. 1983. On the errors involved in computing the empirical characteristic function. *Journal of Statistical Computation and Simulation*, **17**, 133–149.

Kovac, A., & Silverman, B.W. 2000. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association*. **95**, 172–183.

Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674–693.

Nason, G.P. 1996. Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society Series B*, **58**, 463–479.

Nason, G.P., & Silverman, B.W. 1994. The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics*, **3**, 163–191.

Sardy, S., Percival, D.B., Bruce, A.G., Gao, H.Y., & Stuetzle, W. 1999. Wavelet shrinkage for unequally spaced data. *Statistics and Computing*, **9**, 65–75.

Silverman, B.W. 1986. *Density estimation*. London: Chapman and Hall.

Stoffer, D.S. 1991. Walsh-Fourier analysis and its statistical applications. *Journal of the American Statistical Association*, **86**, 461–485.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society Series B*, **36**, 111–47.

Vannucci, M., & Corradi, F. 2000. Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *Journal of the Royal Statistical Society Series B*, **62**, 971–986.

Vidakovic, B. 1999. *Statistical modeling by Wavelets.* New York: Wiley.

Wang, Y. 1996. Function estimation via wavelet shrinkage for long-memory data. *Annals of Statistics*, **24**, 466–484.

Weyrich, N., & Warhola, G.T. 1998. Wavelet shrinkage and generalized cross validation for image denoising. *IEEE Trans. Im. Process.*, **7**, 82–90.