EXTENDING THE SCOPE OF WAVELET REGRESSION METHODS BY COEFFICIENT-DEPENDENT THRESHOLDING

ARNE KOVAC, BERNARD W. SILVERMAN

ABSTRACT. Various aspects of the wavelet approach to nonparametric regression are considered, with the overall aim of extending the scope of wavelet techniques, to irregularly-spaced data, to regularly-spaced data sets of arbitrary size, to heteroscedastic and correlated data, and to data some of which may be downweighted or omitted as outliers.

At the core of the methodology discussed is the following problem: if a sequence has a given covariance structure, what is the variance and covariance structure of its discrete wavelet transform? For sequences whose length is a power of 2, an algorithm for finding all the variances and within-level covariances in the wavelet table is developed and investigated in detail. In particular, it is shown that if the original sequence has band-limited covariance matrix, then the time required by the algorithm is linear in the length of the sequence.

Up to now, most statistical work on wavelet methods presumes that the number of observations is a power of 2 and that the independent variable takes values on a regular grid. The variance-calculation algorithm allows data on any set of independent variable values to be treated, by first interpolating to a fine regular grid of suitable length, and then constructing a wavelet expansion of the gridded data. The gridded data will, in general, have a band-limited covariance matrix, and the algorithm therefore allows the elements of the wavelet transform to be thresholded individually using thresholds proportional to their standard deviation.

Various thresholding methods are discussed and investigated. Exact risk formulae for the mean square error of the methodology for given design are derived and used, to avoid, as far as possible, the need for simulation in assessing performance. Both for regular and irregular data, good performance is obtained by noise-proportional thresholding, with thresholds somewhat smaller than the classical universal threshold.

The general approach allows outliers in the data to be removed or downweighted, and aspects of such robust techniques are developed and demonstrated in an example. Another natural application is to data that are themselves correlated, where the covariance of the wavelet coefficients is not due to an initial grid transform but is an intrinsic feature of the data. The use of the method in these circumstances is demonstrated by an application to data synthesized in the study of ion channel gating. The basic approach of the paper has many other potential applications, and some of these are discussed briefly.

Arne Kovac is wissenschaftlicher Mitarbeiter, Fachbereich 6, Mathematik und Informatik, Universität Gesamthochschule Essen, 45117 Essen, Germany (E-mail: Arne.Kovac@uni-essen.de); and Bernard W. Silverman is Professor of Statistics, School of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK. (E-mail: B.W.Silverman@bristol.ac.uk) This work was started when AK was a research student at the University of Bristol, supported by the German Academic Exchange Service and was continued while BWS was a Fellow at the Center for Advanced Study in the Behavioral Sciences, Stanford, supported by NSF grant SBR-9601236. The authors gratefully thank Guy Nason and Martin Löwendick for their helpful comments.

1. INTRODUCTION

Wavelet methods are the topic of much current interest in statistics. They have been most widely studied in the context of non-parametric regression, where it is of interest to estimate a function f on the basis of observations $y_1, \ldots, y_n$ at time points $t_1, \ldots, t_n$, modelled as

$$y_i = f(t_i) + \varepsilon_i$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are noise.

With some notable exceptions, the current literature mainly deals with the case where the ε_i are independent and identically distributed, the points t_i are equally spaced, and n is a power of 2. The methodology we shall develop will allow all these assumptions to be relaxed, though we shall only develop particular departures in detail. We shall especially be concerned with non-equally spaced points t_i and with general sample size, with robust methods that allow outliers to be downweighted in the fitting process, and with correlated and heteroscedastic errors ε_i.

Most wavelet-based methods make use of the discrete wavelet transform (DWT) described, for example, by Mallat (1989a). In its standard form, this provides a multiresolution analysis of a vector c_J of 2^J values. In the 'classical' wavelet regression setting, these values are just the data points y_i, but more generally they may be obtained from the original data in a number of different ways depending on the precise context.

In classical wavelet regression, the variance matrix of c_J is a multiple of the identity matrix. Because the DWT is an orthogonal transform, this implies that the wavelet coefficients are also uncorrelated with equal variances. Among other authors, Johnstone and Silverman (1997) have considered the case of wavelet thresholding where the noise is correlated but stationary. In this situation the variances of the wavelet coefficients at each level are identical, and therefore Johnstone and Silverman consider thresholding the coefficients level by level. But what if c_J has more general variance matrix? The individual coefficients in the DWT will then be heteroscedastic and correlated, in general. In this paper we set up an algorithm which will yield all the variances and all the within-level covariances of the DWT for a wide range of covariance matrices Σ. Provided the covariance matrix Σ is band-limited, the running time of the algorithm will be linear in the number of wavelet coefficients 2^J. There is no requirement that Σ be a stationary variance matrix.

This algorithm has broader potential uses than the ones we develop in this paper, and since it is in a sense the core of our methodology, we present it and discuss its complexity properties in Section 2, before applying it in specific regression contexts.

Our regression algorithm for generally positioned t_i falls into three main parts. The first phase, if necessary, is to map the original data to a grid of 2^J equally-spaced points to produce a vector $\tilde{y}$. Even if the original data are independent and identically distributed, the gridded values will have covariance matrix Σ that will have, in general, a nonstationary band-limited structure. A general covariance matrix may also arise from correlated or heteroscedastic data. The second phase is to carry out a discrete wavelet transform of the vector $\tilde{y}$, and to use the algorithm set out in Section 2 to find its within-level covariance structure, possibly up to a multiplicative constant. The third phase is to threshold the DWT using a thresholding method that may take into account the heteroscedasticity of the coefficients, and to invert to find the estimate itself. Our investigation of simulated examples indicates that a good approach is to use coefficient-wise thresholding with thresholds proportional to their standard deviations, and to use the SURE (Stein unbiased risk estimate) approach to determine the constant of proportionality. Theoretical support for coordinatewise thresholding of this kind is provided by Johnstone and Silverman (1997).

The method adapts easily to handle robust estimation, where outlying observations are downweighted in the fitting process. Various aspects of this approach are set out and discussed in Section 6. In Section 7, we go on to consider and demonstrate the application of our approach to heteroscedastic and correlated data.

Finally, some suggestions of possible avenues for future research are given in Section 8. Software for the implementation of the methodology discussed in the paper is available from the authors.

2. Calculation of the variances of wavelet coefficients

2.1. **Linear filters and the DWT algorithm.**

As a preliminary to setting out our algorithm for finding all the within-level covariances, we first review some elementary aspects of wavelets, partly in order to fix notation that will be useful later.

Assume that ψ is a mother wavelet of order m, and that ϕ is the corresponding scaling function or father wavelet. The wavelet functions ψ_{jk} which are derived from the mother wavelet by the relationship

$$\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k) \tag{1}$$

form an orthonormal basis of $L^2(\mathbb{R})$, and $\psi(t)$ is orthogonal to polynomials of degree up to m. See Meyer (1992) for an exact definition.

In order to discuss the discrete wavelet transform (DWT), it will be useful to set up some linear filter notation. A linear filter $\mathcal{F}$ is a mapping whose action on a doubly infinite sequence (x_i) is defined by

$$(\mathcal{F}x)_k = \sum_i f_{i-k} x_i \tag{2}$$

where f_j is a doubly-infinite sequence of coefficients. In this paper, we only consider filters for which only finitely many f_j are nonzero, so that (2) is a finite sum and there are no issues of convergence to consider. If x is a vector of finite length, then the definition of $\mathcal{F}x$ depends on a choice of treatment at the boundaries. Common are periodic boundary conditions, where x is extended periodically to give an infinite sequence, and symmetric boundary conditions, where x is reflected at the boundaries.

The other operator that will be useful to define is a "binary decimation" operator $\mathcal{D}_0$ that chooses every even member of a sequence:

$$(\mathcal{D}_0 x)_j = x_{2j}.$$

The scaling function ϕ satisfies a self-similarity equation

$$\phi(x) = \sum_{k\in\mathbb{Z}} h_k \phi(2x - k),$$

for some sequence (h_k). We shall assume throughout that the family of wavelets being used is such that ϕ has bounded support, so that h_k is zero outside the range $0 \le k < N$ for

$$\begin{tikzcd}
 & \boxed{d^{J-1}} & \boxed{d^{J-2}} & \cdots & \boxed{d^1} & \boxed{d^0} \\
y = c^J \arrow[ur, "\mathcal{D}_0\mathcal{G}"] \arrow[r, "\mathcal{D}_0\mathcal{H}"] & c^{J-1} \arrow[ur, "\mathcal{D}_0\mathcal{G}"] \arrow[r, "\mathcal{D}_0\mathcal{H}"] & c^{J-2} & \cdots & c^1 \arrow[ur, "\mathcal{D}_0\mathcal{G}"] \arrow[r, "\mathcal{D}_0\mathcal{H}"] & \boxed{c^0}
\end{tikzcd}$$

FIGURE 1. Scheme of the Discrete Wavelet Transformation. The data vector y is decomposed to vectors of wavelet coefficients $d^{J-1}, d^{J-2}, \ldots, d^0, c^0$. Two linear filters $\mathcal{G}$ and $\mathcal{H}$ which arise from a multiresolution analysis and a binary decimation operator $\mathcal{D}_0$ are recursively applied to the data. The vectors c^j and d^j are called the *smooth* and the *detail* at level j.

some integer N. If we now define

$$g_k = (-1)^k h_{1-k}$$

then the mother wavelet satisfies

$$\psi(x) = \sum_{k \in \mathbb{Z}} g_k \phi(2x - k).$$

See, for example, Daubechies (1992); Chui (1992); Meyer (1992). Denote by $\mathcal{G}$ and $\mathcal{H}$ the linear filters defined by the coefficients (g_k) and (h_k) respectively.

To carry out the DWT, these operators are applied to the data y in the following manner (see figure 2.1): Let $c^J = y$. We now define, recursively for $j = J-1, J-2, \ldots, 0$, the smooth c^j at level j and the detail d^j at level j by

$$c^j = \mathcal{D}_0 \mathcal{H} c^{j+1} \quad \text{and} \quad d^j = \mathcal{D}_0 \mathcal{G} c^{j+1}. \tag{3}$$

Let n_j denote the length of the vectors c^j and d^j. If we use periodic boundary conditions, then $n_j = 2^j$. If symmetric boundary conditions are used, then the vectors need to be slightly longer; for each j we have $2^j \leq n_j < 2^j + N$. See Nason and Silverman (1994) for further details.

The DWT of the data y is defined to be the coefficients $d^{J-1}, \ldots, d^0, c^0$. Regarding these coefficients as a single vector w, we can write $w = \mathcal{W} y$ where $\mathcal{W}$ is an orthogonal matrix in the periodic case; the algorithm we have set out then allows w to be found in $O(N2^J)$ operations.

In the symmetric case, the number of operations is $O(N\{2^J + JN\})$ which is $O(N2^J)$ provided the very mild condition $NJ < 2^J$ is satisfied. We denote for future reference by H_j and G_j the $n_{j-1} \times n_j$ matrices such that $c^{j-1} = H_j c^j$ and $d^{j-1} = G_j c^j$.

2.2. **Calculation of the variances of wavelet coefficients.** Consider the discrete wavelet transform of a vector $\tilde{y}$ of length 2^J whose elements have general covariance matrix Σ. We will set out an algorithm for finding the variances of all the wavelet coefficients w. In the case where Σ is a band matrix, the number of operations will be linear in the length of $\tilde{y}$, while even in the case of general Σ only a quadratic number of operations is required.

The algorithm is straightforward. Let Σ^j denote the variance matrix of c^j and $\tilde{\Sigma}^j$ that of d^j. Then

$$\Sigma^J = \Sigma$$

by definition. From the recursion (3) it follows that, for each $j = J-1, J-2, \dots, 0$,

$$\Sigma^j = H^{j+1}\Sigma^{j+1}(H^{j+1})^T \tag{4}$$

and

$$\tilde{\Sigma}^j = G^{j+1}\Sigma^{j+1}(G^{j+1})^T. \tag{5}$$

Note that this gives us not only the variances $\sigma_{jk} = \tilde{\Sigma}^j_{k,k}$ of the individual wavelet coefficients d^j_k, but also the covariance structure of the wavelet coefficients at each level. In most of our work we shall only be interested in the variances themselves, and so only the main diagonal elements of $\tilde{\Sigma}^j$ will be required; however there may be further developments where the within-level covariances are also important.

The key to the economy of the algorithm is to make use of the sparsity structure of the matrices in the recursions (4) and (5). By the assumption that the filters $\mathcal{H}$ and $\mathcal{G}$ are of finite length N, each row of the matrices H_{j+1} and G_{j+1} is zero except in at most N consecutive positions. Consequently, the calculation of each element of Σ^j and $\tilde{\Sigma}^j$ requires at most a constant multiple of N^2 operations, and a more economical method will be discussed in Section 2.3 below. Summing over j, the total number of elements of all the matrices Σ^j and $\tilde{\Sigma}^j$ is of order 2^{2J}, and so the total complexity of the algorithm even for a general matrix Σ is at most $O(N^2 2^{2J})$. If Σ is a band matrix, then considerable additional economies are possible, and these will also be discussed in Section 2.3 below.

2.3. **Computational complexity for bandlimited variance matrices.** The key to our fast algorithm for finding all the covariance matrices Σ_j recursively in the case where Σ is a band matrix is that the band structure of each Σ_{j+1} is inherited by Σ_j. In this section we examine this aspect in detail, and conclude that the proposed method is essentially of order $2^J b_J$. The precise order is given in the theorem proved below. The main conclusion is that, as long as we assume the bandwidth of Σ^J to be bounded by a constant, the computation time of the algorithm increases only linearly with the number of points 2^J. Furthermore, the dependence on b_J is also linear, and (although we will normally only be concerned with a fixed N) the dependence on N is linear rather than quadratic.

In the periodic case, define the bandwidth b_j to be the smallest integer such that $\Sigma^j_{i,m} = 0$ for all $i, m \in \{0, \dots, n_j - 1\}$ with

$$b_j < |i-m| < n_j - b_j.$$

In the symmetric case, we require the conventional condition that $\Sigma^j_{i,m} = 0$ whenever $|i-m| < b_j$. Note that the bandwidth of the identity matrix is 0. In either case, Σ^j has no more than $(2b_j+1)\cdot n_j$ entries, and by symmetry only $(b_j+1)\cdot n_j$ entries have to be calculated and stored.

Consider the periodic case first. Treating sums in the indices modulo n_j, the sum in (4) can be reduced to

$$\Sigma^j_{i,m} = \sum_{k,l=0}^{N-1} h_k h_l \Sigma^{j+1}_{2i+k,2m+l}. \tag{6}$$

It now follows easily that every term in the sum is zero if $|i-m|$ modulo 2^j is greater than $(b_{j+1}+N-1)/2$, and hence

$$b_j \leq (b_{j+1}+N-1)/2. \tag{7}$$

For the variance matrix of the d^j we obtain the expression

$$\tilde{\Sigma}^j_{i,m} = \sum_{k,l=0}^{N-1} g_k g_l \Sigma^{j+1}_{2i+k-N+2,2m+l-N+2} \tag{8}$$

from which it can again be shown that the bandwidth of $\tilde{\Sigma}^j$ is subject to the same bound.

A direct consequence of (7) is that we have the uniform bound

$$b_j \leq \max\{b_J, N-1\} \tag{9}$$

for all j.

For the symmetric case the calculations are a little more complicated. The reflection at the boundaries means that the first and last few rows of the matrices H_j and G_j will have nonzero elements in reflected versions of the positions obtained merely by shifting the filter coefficients to the appropriate starting points. This means that, near the boundary, $\Sigma^j_{i,m}$ can be nonzero for i and m that are somewhat more separated than in the periodic case, but by consideration of the various cases that can occur it can be seen that the number of additional nonzero bands in Σ^j is at most $(N-1)/2$. It follows that we have the bound

$$b_j \leq (b_{j+1}+2N-2)/2. \tag{10}$$

Again, the bandwidth of $\tilde{\Sigma}^j$ satisfies the same bound, and we have a uniform bound

$$b_j \leq \max\{b_J, 2(N-1)\} \tag{11}$$

for all j.

These results can be used to control the number of multiplications in our algorithm: Before calculating the covariance structures of the j-th level Σ^j and $\tilde{\Sigma}^j$, we determine the actual bandwidth b_{j+1} of Σ^{j+1}, and hence a bound on b_j. We then compute only the main diagonal, the b_j diagonals next to it and (in the periodic case) the corners of Σ^j.

The direct evaluation of each element in (6) and (8) requires a computation of order N^2 operations. However economy is possible, for example, by writing (4) in the two stages

$$A = H^{j+1}\Sigma^{j+1} \tag{12}$$

and

$$\Sigma^j = H^{j+1}A^T. \tag{13}$$

Consider the periodic case first. Since each column of Σ^{j+1} has nonzero elements in at most $1+2b_{j+1}$ positions, the nonzero elements of A can be found in $O(N\{1+2b_{j+1}\}2^j)$ operations. Each row of A then has nonzero elements in at most $N+2bj+1$ positions, so the complexity of the calculation of Σ^j is $O(\{N^2 + Nb_j\}2^j)$. Because of symmetry, only the elements on and above the diagonal have to be calculated, but this does not affect the order of magnitude of the calculation.

In the symmetric case, the quantity 2^j has to be replaced by n_j, since there will be n_j rows of A and of Σ^j to be calculated. We can now state the main theorem which gives the complexity of the overall algorithm for finding all the Σ^j and $\tilde{\Sigma}^j$.

Theorem 2.1. *Define the bound b by $b = \max(b_J, N)$. In the symmetric case, assume that $NJ < 2^J$. The computational complexity of the algorithm described is then of order $Nb2^J$.*

Proof. In the periodic case, we have from (9) that $b_j \leq b$, and hence $b_j + N \leq 2b$ for each j. In the symmetric case, from (11) we have $b_j \leq 2b$ but again $b_j + N$ is bounded by a constant multiple of b. Therefore in both cases the argument set out above shows that the complexity of the calculation of the variance matrices at level j is $O(Nbn_j)$.

To obtain the overall complexity of the algorithm, we sum over j. In the periodic case $\sum_j n_j < 2^J$, while in the symmetric case there are at most N additional coefficients on each of J levels, so, under the extremely mild assumption that $NJ < 2^J$ we again have $\sum_j n_j = 2^J + O(NJ) = O(2^J)$. In either case, the overall complexity is $O(Nb2^J)$, completing the proof of the theorem. □

In the case where Σ does not have a band structure, the complexity of the jth step in the recursions (4) and (5) will be $O(N2^{2j})$ in the periodic case, since each of (12) and (13) requires 2^j applications of a filter of length N to a vector of length 2^{j+1}. So the total complexity will be $O(N2^{2J})$. In the symmetric case the additional calculations will again not affect the order of the calculation provided $NJ < 2^J$. If N is large then some economy may be possible by the use of the fast Fourier transform to perform the convolutions, but we shall not pursue this possibility.

2.4. **Computational complexity for stationary variance matrices.**

The main emphasis of this paper is on cases where Σ is nonstationary and bandlimited. However, we note in passing that if Σ is periodically stationary (i.e. a circulant matrix) then the principle of our algorithm can immediately be applied. Since all the variance matrices are circulant, it is only necessary to store one row of each variance matrix. We define sequences s^j and $\tilde{s}^j$ of length 2^j, such that $\Sigma^j_{lm} = s^j_{l-m}$, and similarly for $\tilde{\Sigma}^j$, with $l-m$ being interpreted modulo 2^j. Then straightforward matrix algebra shows that

$$s^j = \mathcal{D}_0\mathcal{H}\mathcal{H}s^{j+1} \quad \text{and} \quad \tilde{s}^j = \mathcal{D}_0\mathcal{G}\mathcal{G}s^{j+1}. \tag{14}$$

The calculations in (14) can be carried out in $O(2^j \min(j, N))$ operations, because a fast Fourier transform can in principle be used to carry out the convolutions if N is large. So the overall burden of the algorithm, summing over J, is $O(2^J \min(J, N))$, comparable to the nonstationary bandlimited case for fixed b.

3. PREPROCESSING UNEQUALLY TIME-SPACED DATA

3.1. **Relaxing the basic assumptions.**

Wavelet thresholding is a very promising approach in non-parametric regression. However, it requires that the number of data points is a power of two, and the time points must be spaced equally $t_i = i/n$. The first requirement is often met by extending the given data set to length 2^J using for example periodic extension or symmetric reflection (Smith and Eddins 1990). At first sight, this looks a little bit artificial, but we have seen that the DWT itself makes use of boundary conditions. In fact, the results that are obtained with this simple techniques are usually very good.

There are other ways for dealing with data sets that are equally spaced in time but do not contain a power of two elements. Kwong and Tang (1994) as well as Taswell and McGill (1994) propose alternatives to the fast DWT that deal with sets of arbitrary length. These techniques are of particular interest in image compression.

The condition of equal spacing in time raises more problems. One possible approach is only to use the ranks of the t_i and to apply the usual threshold routines direct to the y_i, ignoring the details of the time structure. This would give us an estimate of f at the time points t_i. Unfortunately, wavelet representations of irregularly spaced data are not as economical as they are if the time structure is regular, since the unevenness of the t_i means that regularity properties of a function f are not necessarily inherited by the vector of values $f(t_i)$. As a consequence, the mean square error can be relatively high, as we will see in Section 5.5.

Lenarduzzi (1997) discusses this issue in more detail and suggests a modification involving spline interpolation on a small subset of time points. The coefficients which are cut off by the thresholding function are replaced by the wavelet coefficients of the spline. Her approach does not yield an economical wavelet expansion of the function estimate, and it would be difficult to assess its performance other than by simulation.

Another approach is taken by Antoniadis and Pham (1997). They assume that either the time points are random with a density function g_1 or that they are fixed, but that their empirical distribution converges to a distribution with density g_1 as $n \to \infty$. Instead of estimating f directly they compute estimates of fg_1 and g_1, constructing wavelet estimators on a regular grid of both functions. Although Antoniadis and Pham's algorithms make use of wavelets, they do not seem to have the properties which are expected of wavelet shrinkage estimators in the equally spaced case. They seem to give results very similar to linear estimators like ordinary kernel estimators and spline smoothers.

3.2. **Mapping unequally spaced data to a regular grid.** The method we develop is very simple and can easily be combined with other developments in wavelet methodology such as translation-invariant wavelet transforms and wavelet packets. In addition our method is easily generalized to the case where the original data y_i are themselves correlated or heteroscedastic.

The basic approach is to interpolate the given data onto a regular grid, keeping track of the effect on the correlation structure. We begin by choosing a finest resolution level J. In our experience it is usually reasonable to choose 2^J as the smallest power of two such that $n \leq 2^J$.

Define the grid points

$$\tilde{t}_k = (k + \frac{1}{2})2^{-J}$$

where $k \in \{0, \ldots, 2^J - 1\}$. We calculate a new "observation" $\tilde{y}_k$ by evaluating the straight line that interpolates the nearest data point on the left and on the right at $\tilde{t}_k$. More precisely,

$$\tilde{y}_k = \begin{cases} y_0 & \text{if } \tilde{t}_k < t_0, \\ y_{n-1} & \text{if } \tilde{t}_k \geq t_{n-1} \\ y_i + (\tilde{t}_k - t_i) \cdot \frac{y_{i+1} - y_i}{t_{i+1} - t_i} & \text{otherwise.} \end{cases}$$

In the third case, the value of i is chosen as the smallest integer such that

$$t_i \leq \tilde{t}_k \leq t_{i+1}.$$

The linear function that maps for given $t_0, \ldots, t_{n-1}$ and resolution level J the original data to the grid data will be denoted by R, so that

$$\tilde{y} = Ry.$$

The vector $\tilde{y}$ has length a power of two, and so can be dealt with by standard DWT methods. If the original observations y are uncorrelated with variance σ^2, then the covariance matrix Σ of $\tilde{y}$ is given by

$$\Sigma = \sigma^2 \cdot RR^T.$$

The matrix RR^T is a band matrix, because, for any k and l, the linear interpolation scheme we have set out ensures that $\tilde{y}_k$ and $\tilde{y}_l$ are uncorrelated if at least two of the original time points t_i lie in the interval $[\tilde{t}_k, \tilde{t}_l]$. The bandwidth of the matrix RR^T will essentially depend on the largest gap in the t_i.

Once the matrix Σ has been found, we can carry out a DWT of the sequence $\tilde{y}$ to obtain coefficients d_k^j. The algorithm set out in Section 2.2 will allow us to find the variances and within-level covariances of all these coefficients. If the variance σ^2 is not known, then the same algorithm starting with the matrix RR^T will yield constants γ_{jk} depending only on the time points, the wavelet filter and the length of the grid such that

$$\sigma_{jk}^2 = \operatorname{var} d_k^j = \sigma^2 \cdot \gamma_{jk}.$$

In this paper, we shall consider in detail the case where the original observations y are homoscedastic and uncorrelated. However, the extension of the material of this section to more general distributions is straightforward. If the y have variance matrix Σ_Y, then we have

$$\Sigma = R\Sigma_Y R^T.$$

If Σ_Y is a diagonal matrix with unequal entries, then the bandwidth of Σ will be the same as in the homoscedastic case. If Σ_Y is a more general band matrix, then Σ will obviously still be a band matrix with a somewhat larger bandwidth. The detailed development of these cases is a topic for future research; one issue is of course the specification or estimation of a suitable matrix Σ_Y in the general case. However, we will see in Section 7 two examples with heteroscedastic and correlated data where these ideas perform very well.

4. THRESHOLDING HETEROSCEDASTIC WAVELET COEFFICIENTS

4.1. **The general approach.** Let us now assume that $\tilde{f}$ is the 2^J-vector of values $f(\tilde{t}_\ell)$ and that $\tilde{w}_{jk}$ is the array of discrete wavelet coefficients of the vector $\tilde{f}$. Suppose that we have an array of wavelet coefficients d_k^j that may be considered as observations of the $\tilde{w}_{jk}$ corrupted by heteroscedastic noise, and furthermore that we know the variances $(\sigma_k^j)^2$ of the d_k^j at least up to a constant. Typically these coefficients and variances will have been obtained from homoscedastic but irregularly spaced data in the manner described in Section 3.2. In order to reconstruct the function f, the general approach within the wavelet literature is to threshold the coefficients in some way, and then to invert the DWT to complete the estimation of f.

In general, we shall apply a thresholding function η to each wavelet coefficient yielding modified wavelet coefficients $\hat{w}_{jk}$

$$\hat{w}_{jk} = \eta(d_k^j, \tau_{jk})$$

where η is either the soft thresholding function

$$\eta_S(d_k^j, \tau) = \operatorname{sgn}(d_k^j)(|d_k^j| - \tau)_+$$

or the hard thresholding function

$$\eta_H(d_k^j, \tau) = d_k^j \cdot I\{|(|d_k^j) \geq \tau\}.$$

Performing the inverse transform on $\hat{w}$ gives us the estimate $\hat{f}$ for $\tilde{f}$:

$$\hat{f} = \mathcal{W}^T \hat{w}$$

4.2. **Universal thresholding.** A natural approach is to choose each threshold τ_{jk} proportional to its standard deviation σ_{jk}. One possible approach is to adapt the universal threshold or *VisuShrink* approach and use

$$\tau_{jk} = \sqrt{2\log(n)} \cdot \sigma_{jk}.$$

The *VisuShrink* approach does not aim to minimize the mean square error, but it tries to produce noise-free reconstructions. Some theoretical basis for the idea of thresholding proportional to noise standard deviation is given by Johnstone and Silverman (1997).

The noise level σ is usually unknown. In the case where the time points are equally spaced, Donoho, Johnstone, Kerkyacharian and Picard (1995) have suggested the estimation of the the noise level by taking the median absolute deviation of the coefficients at the finest scale of resolution, and dividing by 0.6745. However, in our setting it is necessary to divide each coefficient d_k^{J-1} by $\sqrt{\gamma_{J-1,k}}$ beforehand. Moreover, it can happen that some of the $\gamma_{J-1,k}$ are zero, for example when all grid points that have influence on the coefficient d_k^{J-1} lie between two original observations, because of the property of vanishing moments and the linear interpolation that is used in the grid transform. If the variance factor is zero, simple linear algebra arguments show that the corresponding wavelet coefficient will also be zero whatever the original input vector $\tilde{y}$. To avoid numerical difficulties we separate out this case by testing for very small variance factors. Taking this into account we suggest the following estimator:

$$\hat{\sigma} = \mathrm{MAD}\{d_k^{J-1}/\sqrt{\gamma_{J-1,k}} : \gamma_{J-1,k} > 0.0001\}/0.6745. \tag{15}$$

We then set

$$\hat{\sigma}_{jk}^2 = \hat{\sigma}^2 \gamma_{jk}$$

for all j and k.

4.3. **An unbiased risk approach.** Another possible method for threshold selection is based on a method of Stein (1981). It was introduced as *SureShrink* by Donoho and Johnstone (1995) for the i.i.d. Gaussian error setting, and adapted for correlated noise by Johnstone and Silverman (1997).

Let w^* denote the vector of wavelet coefficients obtained by interpolating the values $f(t_i)$ to the regular grid $\tilde{t}_j$. By a simple extension of the argument given in Section 2.3.2 of Johnstone and Silverman (1997), for any array $\mathcal{T} = \{\tau_{jk}\}$ of thresholds, the quantity

$$S(\mathcal{T}) = \sum_{j,k} [\hat{\sigma}_{jk}^2 + \min\{(d_k^j)^2, \tau_{jk}^2\} - 2\hat{\sigma}_{jk}^2 I\{|d_k^j \leq \tau_{jk}|\}]$$

may be used as an estimate of the risk $\mathbb{E}||\hat{w} - w^*||^2$, for soft thresholding. If the estimates $\hat{\sigma}_{jk}^2$ were replaced by the true values σ_{jk}^2, then $S(\mathcal{T})$ would be an unbiased estimate of this risk. In addition, this is not exactly the risk we are interested in, because of the error involved in replacing the wavelet transform of the true grid values of the signal by the values w^*. Because of these approximations, we regard the unbiased risk property of the criterion $S(\tau)$ as heuristic rather than rigorous justification for its use.

In using the unbiased risk criterion, we restrict attention to thresholds proportional to estimated standard deviation, and define $S(\tau)$ to be $S(\mathcal{T})$ for $\tau_{jk} = \hat{\sigma}_{jk}\tau$. We then set

$$\hat{\tau}_{jk} = \hat{\sigma}_{jk}\hat{\tau} \tag{16}$$

where $\hat{\tau}$ is chosen to minimize $S(\tau)$ over the range $[0, \sqrt{2\log n}]$.

5. Performance of the Thresholding Methods

In this section we discuss in detail the performance of the methods we have introduced. The expected mean squared error is used as a measure of the performance throughout this section. First, we derive an exact risk formula for given time points and threshold choice. This is used in Section 5.2 to analyze the risk of the VisuShrink method for time points that are equally spaced on a grid of arbitrary length. In Section 5.3 comparisons are made with the minimum of the risk over all threshold values. It turns out that dividing the VisuShrink thresholds by 3 gives mean square error performance nearly as good as this unattainable minimum.

Irregularly spaced time points are considered in Section 5.4. Finally we present in Section 5.5 simulation results which compare an adaptation of the SureShrink estimator with usual and reduced VisuShrink thresholds and with methods where usual thresholding is applied to the data and the grid transform is used to obtain an estimate at the time points.

5.1. **Exact risk formulae.** We assume throughout this section that periodic boundary conditions are applied, so that the discrete wavelet transform is orthogonal, and function $f(t)$ is estimated at 2^J equally-spaced points. Let $\tilde{f}$ and $\hat{f}$ be vectors of length 2^J giving the values and estimates of the function on this grid, and let $\hat{w} = \mathcal{W}\hat{f}$ and $\tilde{w} = \mathcal{W}\tilde{f}$ be the wavelet transforms of these vectors. Using Parseval's equation, the mean square error $\mathrm{MSE}(\hat{f}, \tilde{f}) = 2^{-J}\|\tilde{f} - \hat{f}\|_2^2$ can be written as

$$\mathrm{MSE}(\hat{f}, \tilde{f}) = 2^{-J}\|\hat{w} - \tilde{w}\|_2^2 = 2^{-J} \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j} \Big(\eta_S[\{\mathcal{W}R(f^* + \varepsilon)\}_{jk}, \tau_{jk}] - \tilde{w}_{jk}\Big)^2 \tag{17}$$

where τ_{jk} are the individual thresholds, j_0 the "cut-off-level", below which no thresholding is carried out, and f^* the true values $f(t_i)$ on a given sequence t_i of length n.

Let w^* be the wavelet coefficients of the sequence Rf^*. The individual coefficient $(\mathcal{W}R(f^* + \varepsilon))_{jk}$ is normally distributed with mean w^*_{jk} and variance σ_{jk} that can be calculated with the algorithms that were introduced above. To explore the mean square error, we define

$$\rho(\tau; \mu_1, \mu_2, \sigma) = \mathbb{E}(\mu_1 - \eta_S(X, \tau))^2$$

where X is a normally distributed random variable with mean μ_2 and variance σ^2.

We then have

$$\mathbb{E}\{\mathrm{MSE}(\hat{f}, \tilde{f})\} = 2^{-J} \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j} \rho(\tau_{jk}; \tilde{w}_{jk}, w^*_{jk}, \sigma_{jk}). \tag{18}$$

To carry out an exact risk calculation for any particular function and time sequence, the vectors $\tilde{f}$ and Rf^* are calculated, and their wavelet transforms substituted into (18). The function ρ can be evaluated from its definition by making use of properties of the normal distribution, to obtain

$$\begin{aligned}\rho(\tau; \mu_1,&\mu_2, 1) \\ &= \mu_1^2 + \{1 + (\tau + \mu_2 - \mu_1)^2 - \mu_1^2\}\Phi(-\mu_2 - \tau) \\ &+ \{1 + (\tau + \mu_1 - \mu_2)^2 - \mu_1^2\}\Phi(\mu_2 - \tau) \\ &+ (2\mu_1 - \mu_2 - \tau)\varphi(-\mu_2 - \tau) \\ &- (2\mu_1 - \mu_2 + \tau)\varphi(\mu_2 - \tau).\end{aligned} \tag{19}$$

This formula generalizes results that were obtained by other authors in the conventional setting where $\mu_1 = \mu_2$ (Donoho and Johnstone 1994; Abramovich and Silverman 1998).

5.2. **Regular grids of arbitrary length.** In this section we consider the performance of the methods on data where the t_i are regularly placed but the number of points is not necessarily a power of two. The basic idea is to use our algorithm to map the data to a grid of length 2^J so that standard DWT implementations can be used. We use our exact risk formula to examine the mean square error for rescaled versions of four test signals *Doppler, Heavisine*, *Blocks* and *Bumps* (see Figure 2) that were analyzed by Donoho (1993a) and Donoho and Johnstone (1994). Calculations were carried out for each n in $\{17, \ldots, 2048\}$. For each n, the time points were given by

$$t_i = (i + 0.5)/n$$

for $i = 1, \ldots, n$, and the size 2^J of the grid $(\tilde{t}_j)$ was chosen to be the smallest power of two not less than n. The noise level σ was chosen as 0.35 (corresponding to a root signal-to-noise ratio $SD(f)/\sigma \approx 6.3$) and the threshold chosen with *VisuShrink*, so that

$$\tau_{jk} = \sqrt{2\log(n)} \cdot \sigma_{jk},$$

assuming the variances to be known.

Daubechies' wavelet with two vanishing moments, periodic boundary conditions and a cut-off-point j_0 of 3 were used. The results are plotted as solid lines in Figure 3.

It can be seen from the figure that the expected mean square error decays very fast; note the use of logarithmic scales on both axes. The risk does not decrease monotonically and

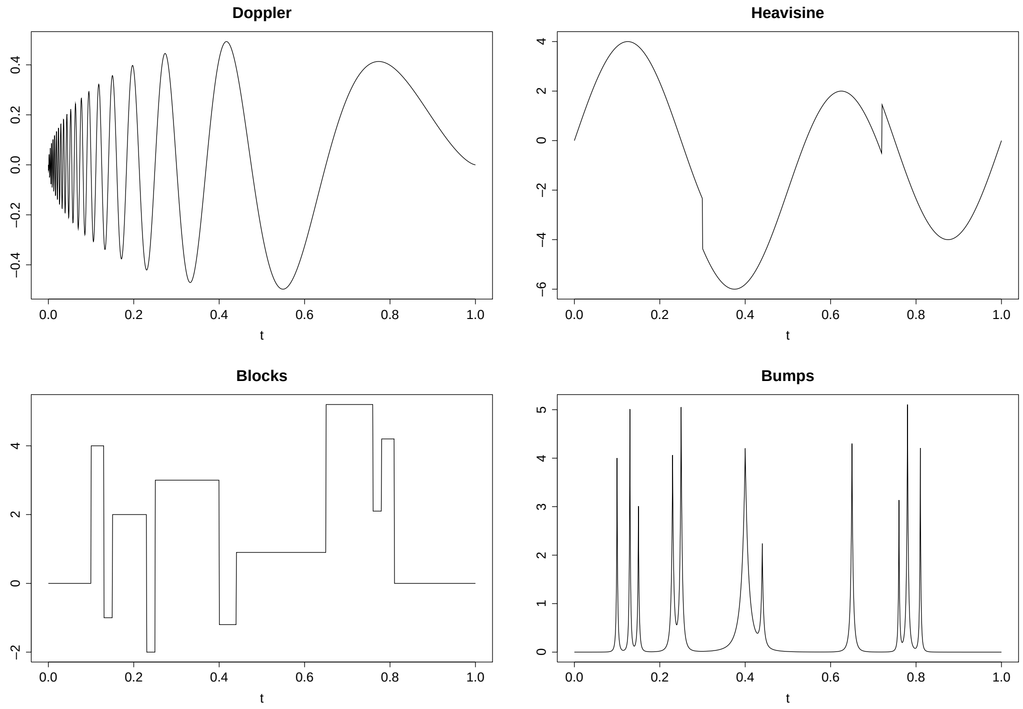

FIGURE 2. Donoho's and Johnstone's test signals: Doppler, Heavisine, Blocks, Bumps

the amount of variation with n is different among the four test signals. The variation in the risk of the Heavisine and Doppler data is relatively small while the Blocks and Bumps data exhibit a lot more oscillation, in particular for small n with the Bumps data and for large n with the Blocks data. Finally, the risk seems often to be exceptionally small when n is a power of two.

To investigate the source of these variation properties we calculated the modified risk

$$E2^{-J}\|\hat{f} - Rf^*\|_2^2,$$

the expected mean square error obtained by supposing the interpolation of $f(t)$ from the original observations to the grid to be exact. An exact formula for this quantity was changing the replacing $\tilde{w}_{jk}$ by w^*_{jk} in equation (18) above. It is plotted as the dashed lines in the figure. The modified risk looks much smoother than the total risk, so indeed the variation can be considered as being due to the bias in the wavelet coefficients caused by the grid transform. Of course, for powers of two both curves take the same value, because the grid transform is then the identity function and causes no bias at all.

Considering the various signals individually, the Blocks signal is very sensitive to small changes in the time structure because it has a large number of discontinuities. This explains why the variation in the error of its estimation is considerable even for large n. On the other hand, the risk for the Bumps signal depends strongly on how well its peaks can be approximated. For small numbers n_1 and n_2 the quality of this approximation can be very

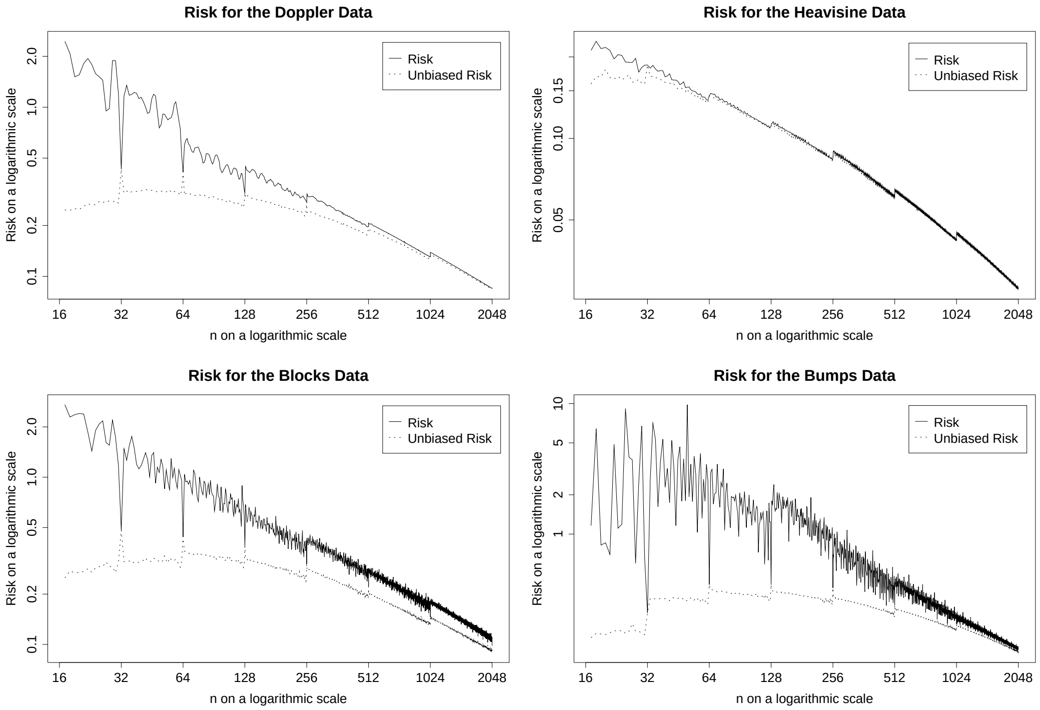

FIGURE 3. The risk for the test functions and *VisuShrink* on regular grids. Also shown as dashed lines is the “unbiased” risk which represents the amount of error which is caused by the actual thresholding step and does not contain the bias which is caused by the grid transform. The risk and the number of data points are always plotted on logarithmic scales.

different, even if n_1 and n_2 are close together. However, for larger n_1 and n_2 the continuity of the Bumps function ensures that the approximations are similar, so that the amount of variation is relatively small for larger sampling rates.

Similar arguments can be put forward for the Heavisine and Doppler signals which are much smoother than the other two signals. Finally, the low values for powers of two that can be seen very clearly for the Blocks signal are also caused by the absence of bias.

Both risk functions exhibit small steps near powers of two. These can be explained by the values of the thresholds which depend explicitly on the number of grid points and increase each time n crosses a power of two. This increment in threshold tends to cause an increase in expected mean square error, not surprisingly since the visual thresholds are already usually larger than the thresholds that minimize expected mean square error, as we shall see below. To confirm this effect, see Figure 4. The solid line shows the risk for the Doppler data using VisuShrink as already seen in Figure 3. The dashed line represents the risk for fixed thresholds

$$\tau_{jk} = \sqrt{2\log(256)} \cdot \sigma_{jk}.$$

Note that the steps have disappeared.

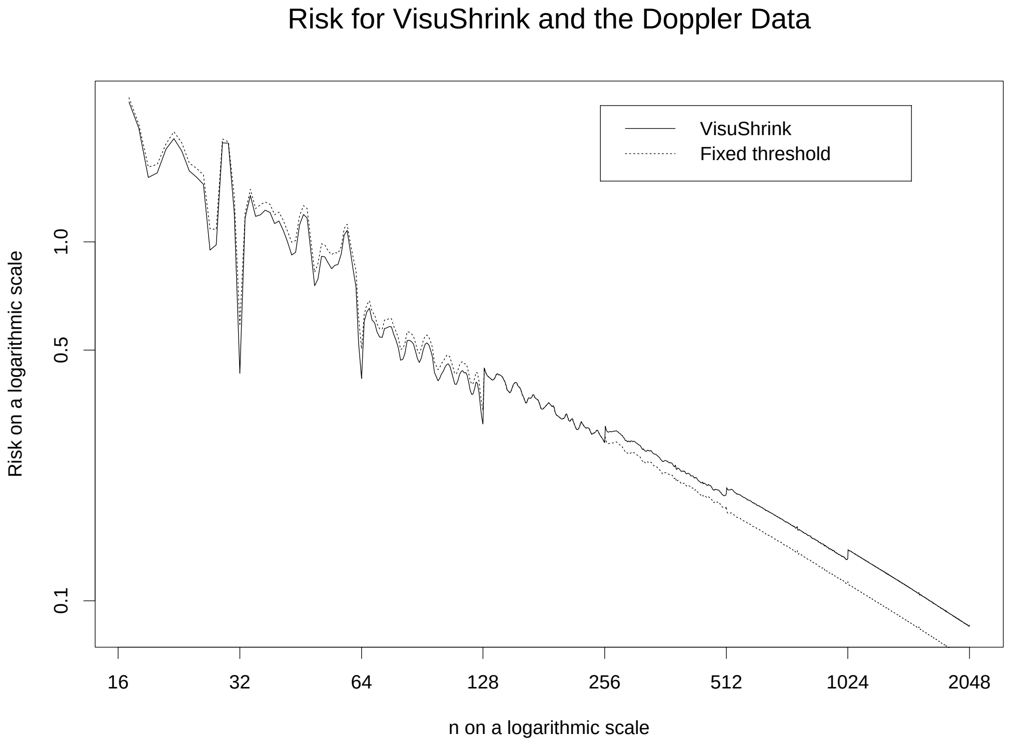

FIGURE 4. The risk for the Doppler data with *VisuShrink* compared with a fixed threshold choice on regular grids. The number of data points n and the risk are both plotted on logarithmic scales.

5.3. **Optimal thresholds.** When comparing different choices for the thresholds it is interesting to ask for the minimal risk that can be obtained with any thresholds for a specific function f. In practice, of course, such optimal thresholds are not available because the signal is unknown, but they do give a reference point against which a practical threshold choice can be judged.

We consider *optimal noise-proportional thresholds* where the thresholds are restricted to the special form $\tau_{jk} = \alpha \cdot \sigma_{jk}$. The thresholds can then be found to desired accuracy by a numerical minimization of the exact risk formula we have derived.

For four different noise levels ($\sigma = 0.05, 0.15, 0.25, 0.35$) and for the same test signals as in Section 5.1 we calculated the minimum risk for n regularly spaced time points with n taking all values between 17 and 2048. The solid lines in figure 5 show the results for the Heavisine data and $\sigma = 0.35$. The results for the other data sets and noise levels do not look substantially different.

It is well known that the optimal thresholds are usually much smaller than those specified by the *VisuShrink* method. We calculated the ratio of the *VisuShrink* threshold to the optimal thresholds for all 32512 test cases (four noise levels, four signals, 2032 values for n) and got a median of 3.4 and the mean 3.9. The 15%-quantile was 2.6 while the 85%-quantile was

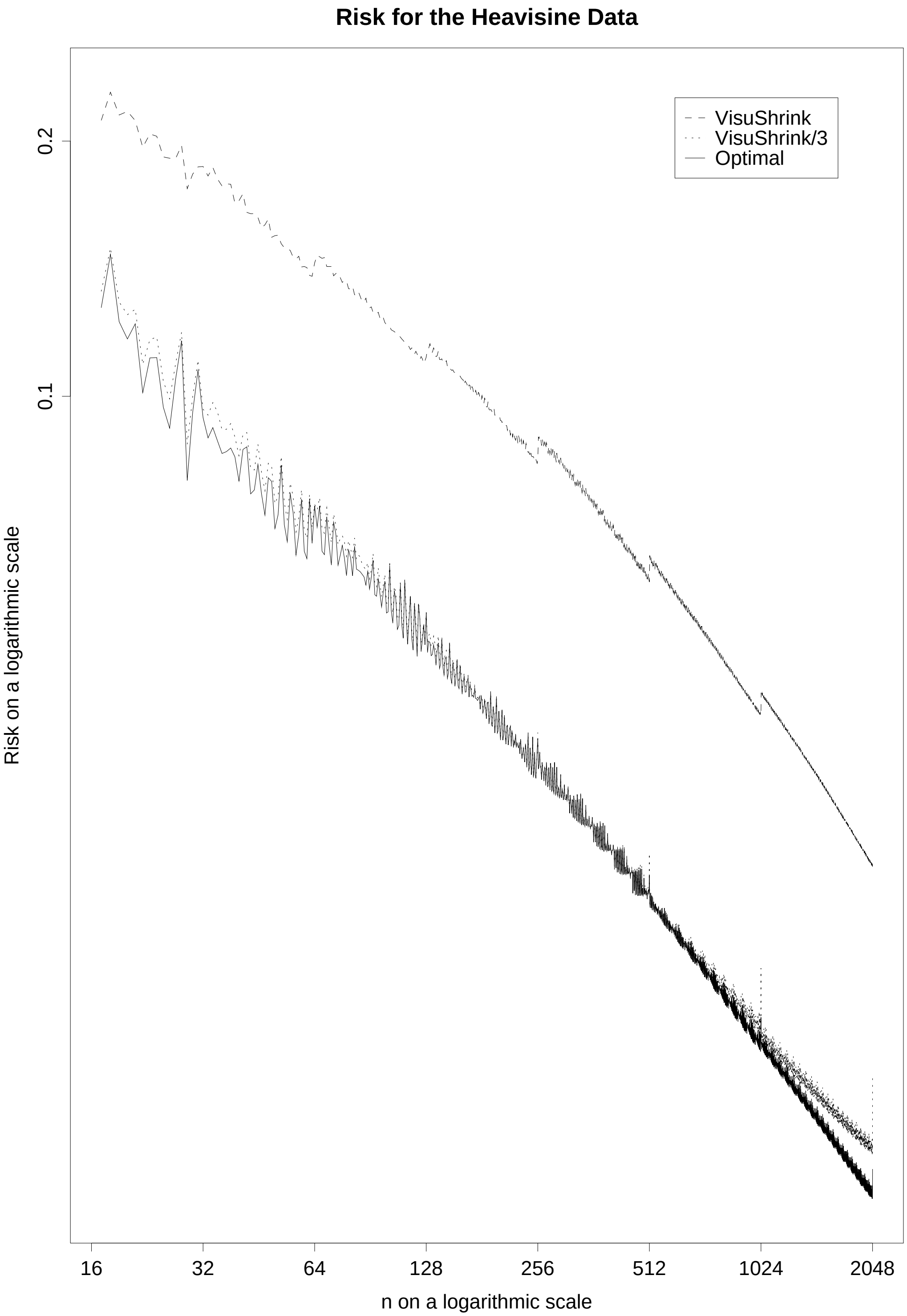

FIGURE 5. The risk for the Heavisine data and noise level 0.35 for three threshold choices on a regular grid of size n. The dashed and solid lines show the risk for the VisuShrink threshold choice and optimal chosen thresholds respectively. The dotted curve represents a threshold choice where the VisuShrink threshold is divided by 3.

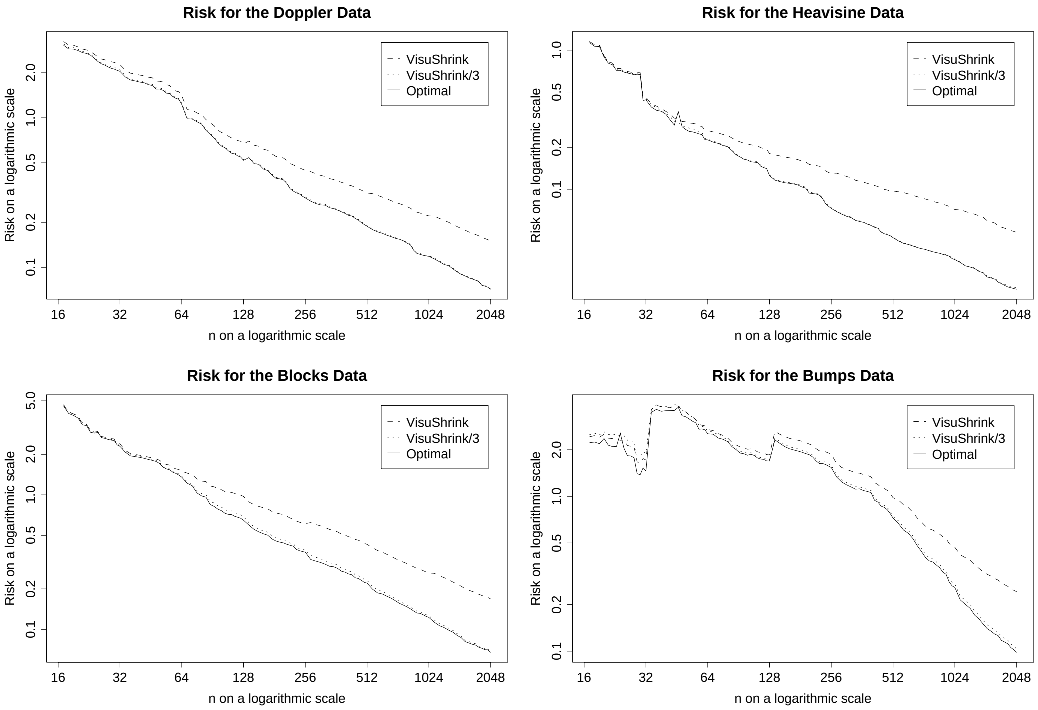

FIGURE 6. The expected MSE for DJ's test signals and noise level 0.35 for three threshold choices on irregular time structures of size n. Time points were independently drawn from a $Beta(2, 2)$ distribution and the exact risk calculated. Shown is the average over 20 replications. The dashed and solid lines represent the risk for VisuShrink and L^2-optimal thresholds of the form $\alpha\sigma_{j,k}$ while the dotted curve shows the risk for reduced Visushrink thresholds. The risk and n are both plotted on logarithmic scales.

4.7. This suggests a simple rule of thumb for small MSE, to use the *VisuShrink* thresholds divided by 3. The dotted line in figure 5 shows the resulting MSE which is very close to the MSE for optimal noise-proportional thresholds. The good behavior of this approach is confirmed in a different context by a simulation study which we present in Section 5.5 below.

5.4. **Irregular time structures of arbitrary length.** We now turn to the case where the time structure is no longer regular. For a range of values of n we simulated 20 different time structures as samples of size n from a Beta(2,2) distribution. For each time structure the exact MSE of a signal f and a fixed noise level σ was calculated, and the average over realizations was found. As in the previous section, this procedure was carried out for optimal and VisuShrink thresholds as well as for the reduced VisuShrink thresholds. The results are plotted for $\sigma = 0.35$ in Figure 6.

The most obvious difference from Figure 3 and Figure 5 is that the curves are much smoother now. This a consequence of the change in the time structure which is no longer fixed for given n. On the other hand, the performance of the reduced VisuShrink thresholds is, again, virtually indistinguishable from that of the optimal thresholds.

Somewhat more explanation is needed for the effects near some powers of two which occur for the Bumps signal in particular. The argument which was given earlier in Section 5.2 in connection with Figure 4 has to be ruled out since the jumps are not only visible for VisuShrink thresholds which do change near powers of two, but also for the other threshold types. In fact, the reason seems to be that the discrete version of the mean square error which we consider in this chapter depends on the grid length. The quality of the estimates does not appear to be much different for $n = 32$ and $n = 33$: most of the peaks of the Bumps signal are either not resolved or massively flattened. However, the quantification via the mean square error is achieved by a comparison with two very different samples of the original signal: While the sample on a regular grid of size 32 exhibits the same features as the reconstructions, the peaks are much better resolved on a grid of size 64, leading to large residuals near peaks. The preceding statements were confirmed by simulation.

5.5. **Simulation comparison for randomly placed time points.** The aim of this final subsection is to compare five techniques for estimating a function in terms of their MSE. Among them are two new approaches that were not analyzed above, but were mentioned in Section 3.1.

This is the idea: We suppose that the number of data points is a power of two. Then, instead of applying the grid transform, we ignore the time structure and perform the DWT directly on the given data (y_i), followed by a thresholding method. This gives us an estimate for the function on the irregularly spaced time points. Finally we carry out the grid transform to get an estimate on a grid.

This algorithm clearly has the advantage that it does not require the calculation of the variances and can easily be generalized to arbitrary numbers of data points by imposing boundary conditions. On the other hand, we will see that these methods often do not work very well as far as their MSE is concerned. This is due to the fact that wavelet transforms of a sample which is taken from irregularly spaced time points are usually not as economical as in the regular setting, because the property of vanishing moments does not apply any more. There are other disadvantages as well: Lenarduzzi (1997) shows that the resulting curves do not look "graphically pleasant", but we do not consider this in more detail.

Table 5.5 shows the result of a simulation study that was carried out in S-Plus to compare five methods:

IRREGSURE: Transform the data to a grid, apply the DWT to the grid data, calculate the *SureShrink* thresholds, apply soft thresholding, and apply the inverse DWT.

IRREGVIS: Same as IRREGSURE, but use noise-proportional thresholds $\alpha \cdot \sigma_{jk}$ with factor α equal to $\sqrt{2\log(n)}$.

IRREGVIS3: Like IRREGVIS, but divide the factor α by 3, ie use thresholds $\tau_{jk} = \sqrt{2\log(n)} \cdot \sigma_{jk}/3$.

RANKSURE: Apply the usual *SureShrink* wavelet procedure of Donoho and Johnstone to the data, taking only the time order of the observations into account, and then perform the grid transform to get an estimate on a grid.

Signal	Time Data	IRREGSURE	IRREGVIS3	RANKSURE	IRREGVIS	RANKVIS
Doppler	B(1,1)	0.032	0.036	0.036	0.119	0.103
	B(2,2)	0.070	0.069	0.073	0.156	0.154
	B(3,3)	0.176	0.159	0.185	0.230	0.292
	B(4,4)	0.343	0.302	0.362	0.335	0.546
Heavisine	B(1,1)	0.014	0.016	0.014	0.039	0.033
	B(2,2)	0.018	0.019	0.028	0.049	0.077
	B(3,3)	0.060	0.054	0.099	0.089	0.250
	B(4,4)	0.152	0.126	0.192	0.133	0.429
Bumps	B(1,1)	0.076	0.084	0.085	0.231	0.231
	B(2,2)	0.094	0.101	0.098	0.254	0.239
	B(3,3)	0.173	0.187	0.183	0.375	0.347
	B(4,4)	0.371	0.385	0.382	0.575	0.555
Blocks	B(1,1)	0.061	0.064	0.061	0.159	0.137
	B(2,2)	0.065	0.067	0.060	0.176	0.138
	B(3,3)	0.091	0.099	0.086	0.246	0.177
	B(4,4)	0.141	0.155	0.137	0.356	0.253

TABLE 1. Comparison of the average Mean Square Error for five different thresholding techniques. The methods IRREGSURE, IRREGVIS and IRREGVIS3 are based on the grid transform and coefficient-dependent thresholds as introduced in Section 3.2 and Section 4.1. RANKSURE and RANKVIS do not take the irregular time structure into account at first and perform ordinary thresholding followed by a grid transform. The time points were chosen randomly from four different Beta distributions. For each thresholding technique, signal, and model for the time structure, the average of the MSE over 50 replications was calculated.

RANKVIS: Perform *VisuShrink* on the data, again taking only the time order into account, and follow by the grid transform.

For each of the four test functions, and for each of the time data sets, as well as for each method, we calculated the average mean square error over 50 replications. The noise was white noise with $\sigma = 0.35$ and the time points samples of size 2048 from four Beta distributions B(1,1) (identical with a uniform distribution), B(2,2), B(3,3) and B(4,4).

Obviously, the methods based on *VisuShrink* do not perform well. But that is not very surprising, because the conservative threshold choice does not attempt to obtain a low MSE. As described above, the main idea of *VisuShrink* is to produce “noise-free” reconstructions.

For all the signals except the Blocks signal, IRREGSURE always performs better than RANKSURE. The simple IRREGVIS3 method also exhibits a small MSE which is in 5 cases even smaller than the MSE for IRREGSURE.

It is interesting that RANKSURE always attains the smallest MSE for the Blocks data. As we pointed out above, wavelet decompositions of smooth functions that are sampled on an irregular grid are usually not as economical as in the equally spaced setting, and the poor performance of RANKSURE on the Heavisine data confirms this. However, for the piecewise constant Blocks signal, samples on irregular grids have the same general properties as those on a regular grid, and so it is not surprising that RANKSURE performs well.

6. ROBUST WAVELET REGRESSION

In this section we discuss the application of our method to the problem of nonparametric regression when some of the observations may be considered to be outliers or when the noise follows a distribution with heavy tails. Donoho (1993) pointed out that standard thresholding techniques do not work very well in this situation.

In the case of equally-spaced time points, we note two suggestions that have been made for dealing with this problem. The approach of Bruce, Donoho, Gao and Martin (1994) is based on an alternative discrete wavelet transform of the data. At each level, the sequence $\mathcal{H}c^{j+1}$ is preprocessed before the decimation step $\mathcal{D}_0$ is carried out to obtain c^j. The algorithm performs in $O(n)$ and is included with the *S+Wavelets* toolkit that is available from MathSoft. A related approach is due to Donoho and Yu (1997). They construct a nonlinear multiresolution analysis based on a triadic grid, so that the present version of their method is restricted to $n = 3^J$ data points for some integer J. The computational time required for their method is $O(n \log_3 n)$.

We propose a more direct approach that is more closely related to classical robustness methods, and is equally applicable to regularly or irregularly spaced data. We identify outliers, remove them from the data, and apply wavelet thresholding to the remaining data points. Of course, classical thresholding cannot be used in such an approach, because the resulting data will no longer be equally spaced even if the original data were. However, our procedure for irregularly-spaced data can be used, and this is illustrated within a particular example.

Figure 7 shows data from a weather balloon. They were analysed previously by Davies and Gather (1994) and are taken from a balloon which took measurements of radiation from the sun. Unfortunately, it happened occasionally that the measurement device was cut off from the sun causing large outlier patches. There are also individual outliers in the data.

The data are highly correlated and therefore for this analysis we decided to subsample the data by working with every 20th data point only, reducing the sample size from 4984 to 250. When applied directly to these data, the method of Section 4.1, using the VisuShrink threshold, produces ugly curves like the one shown in Figure 7. The curve exhibits several high frequency phenomena due to outliers. These have survived even the use of the VisuShrink threshold.

In order to try to remove the outliers and extreme observations, the following procedure was carried out:

1. The variance of the data was estimated from the median absolute deviation of the differences $d_i = (y_{i+1} - y_i)/2$, giving the estimate $\hat{\sigma}_0^2$, say. This corresponds to the usual variance estimation via a wavelet decomposition when the Haar basis is used.

   In principle, wavelet bases of higher order could be used, but for these bases more wavelet coefficients are contaminated by the outliers, because of the wider support of the filters $\mathcal{G}$ and $\mathcal{H}$.
2. For each data point the median over a small window was computed. The window contained the point itself and its five left and right neighbors. If the difference between data point and median was greater than $1.96\,\hat{\sigma}_0$, the point was removed.

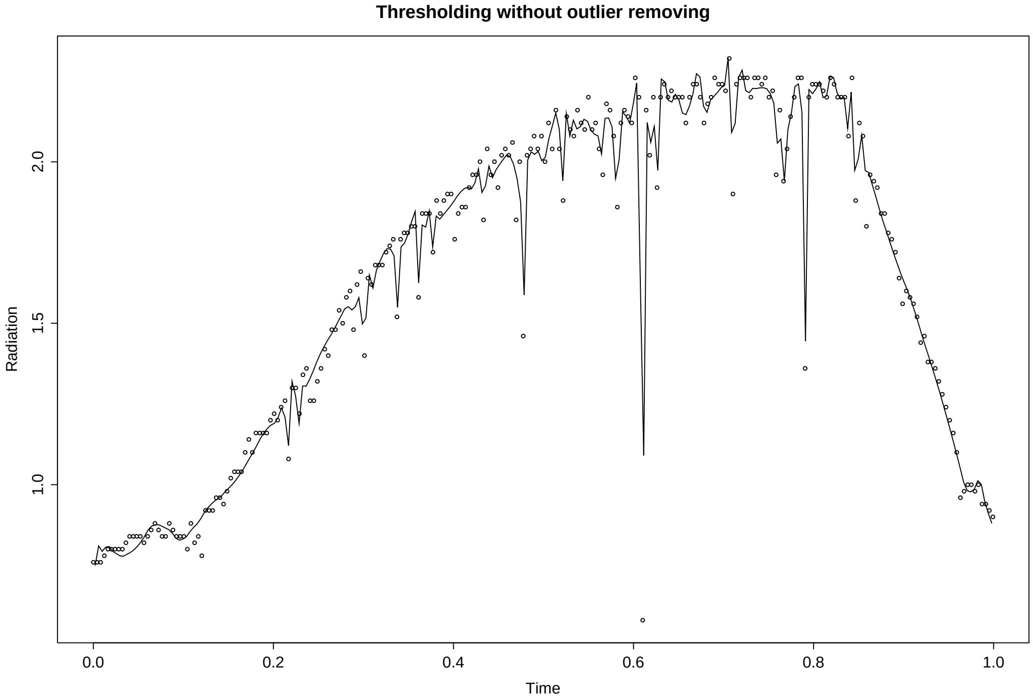

FIGURE 7. Balloon data and a wavelet estimator with VisuShrink thresholds applied to it. The data are taken from a weather balloon and describe the radiation of the sun. The high frequency phenomena in the estimated signal are due to outlier patches in the data which may be caused by a rope which cut the measuring device from the direct sunlight. The wavelet basis with four vanishing moments and extremal phase was used.

3. The thresholding algorithm of Section 4.1 was applied to the modified data set, now taking into account the values of t_i. At this stage, we used *VisuShrink* threshold

$$\tau_{j,k} = \sqrt{2\log(256)}\hat{\sigma}_{j,k},$$

and a wavelet basis with four vanishing moments and extremal phase. The variances of the wavelet coefficients $\hat{\sigma}_{j,k}$ were determined under the assumption that the non-deleted data points were independent with variance $\hat{\sigma}_0^2$. Experiments showed that a re-estimation of the variance from the cleaned data set will typically underestimate the noise level, and so is not to be recommended.

The results can be seen in Figure 8. It is interesting to note that the abrupt changes in slope in the curve are well modeled, but the spurious high frequency effects have been removed. Note that all the calculations can be performed in $O(n)$ operations.

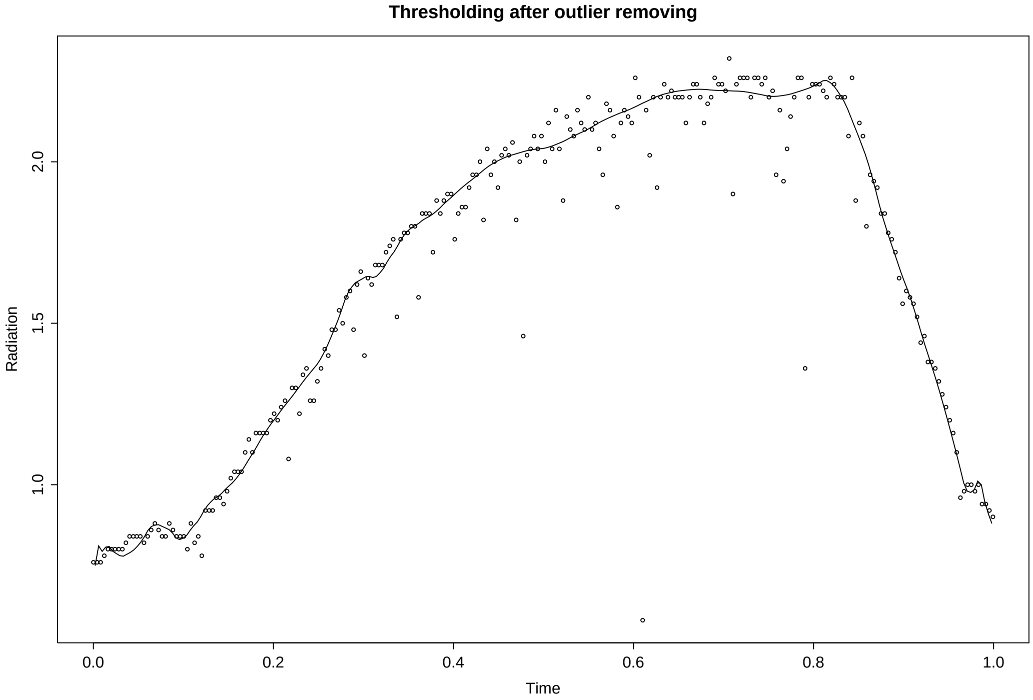

FIGURE 8. Balloon data with a robust wavelet estimator. The thresholding techniques for unequally spaced data were applied to a data set from which outliers had been removed by the procedure explained in the text.

7. HETEROSCEDASTIC AND CORRELATED DATA

7.1. **Heteroscedastic data.** A common problem in nonparametric regression arises when the variance of the data is not constant. Intuitively one might want to adapt the smoothing to remove the noise in regions where the variance is large, but retain as much structure of the signal as possible where the amount of random variation is rather small. Provided we can get some reasonable estimate of the local variance of the data, the methodology we have developed will allow a wavelet thresholding method to be used; the variance calculation algorithm will determine appropriate factors to be used for adjusting the thresholds in different parts of the wavelet table.

Rather than being completely prescriptive, we present a possible approach in the context of an example. Figure 9 shows a data set that has been analyzed extensively in the field of nonparametric regression. It consists of 133 observations of the acceleration of a motorcyclist's head during a crash test. Because of the nature of the experiment, the observations are not available at equally-spaced time points. When applied directly to these data, the method of Section 4.1, using the VisuShrink threshold, a wavelet basis with six vanishing moments and globally estimated variance, produces ugly curves like the one shown in Figure 9. The worst feature is of course the high frequency phenomena around time 0.5, which has survived even the use of the VisuShrink threshold.

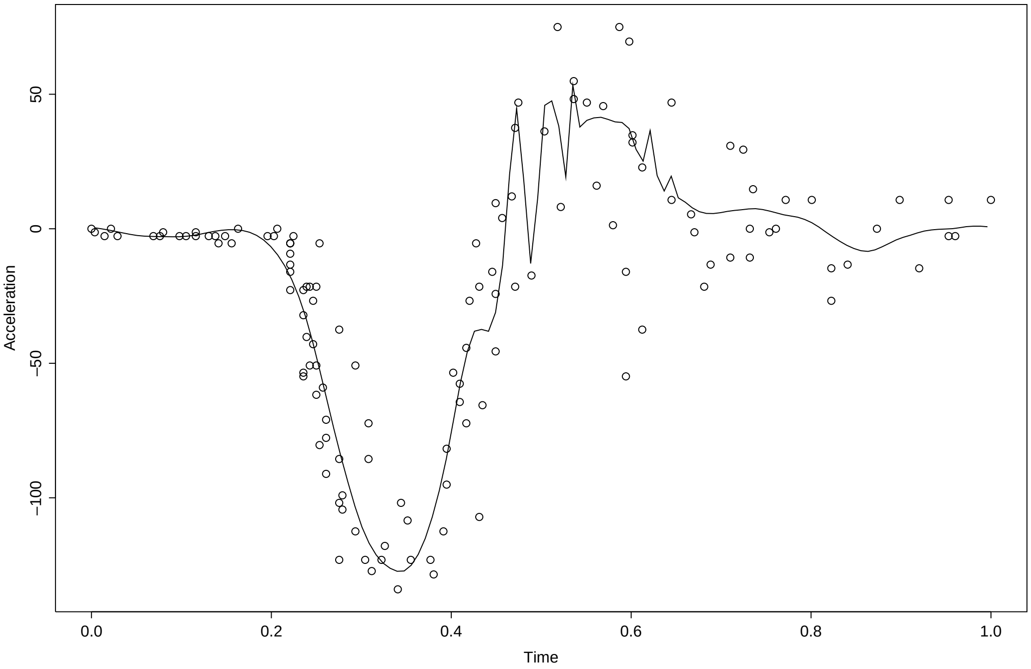

FIGURE 9. Crash data with a wavelet estimator. The data were gathered during a crash test and show the acceleration of a motorcyclist's head. The wavelet estimator was calculated with VisuShrink and a wavelet basis with six vanishing moments. The variance was estimated globally.

Examination of the figure makes it clear that the variance is very small initially, but increases very fast after the impact of the motorcycle and decreases again later. Silverman (1985) used an iterative method based on smoothing the residuals from a weighted spline smoothing to give estimates of the local variance and hence an improved estimate of the whole curve. To give a noniterative wavelet-based method, a slightly different approach was used. For each $i = 1, \ldots, n-1$ the difference $d_i = (y_{i+1} - y_i)/\sqrt{2}$ was calculated, and ascribed to the point $r_i = (t_{i+1} + t_i)/2$. The estimated standard deviation $\hat{\sigma}_i$ of the i-th data point was based on the median of the absolute values of d_i over a small window of size 0.1:

$$\hat{\sigma}_i = \mathrm{med}\{|d_j| : |t_i - r_j| \leq 0.1\}/0.6745.$$

These values for the variances of the individual data were then plugged into the derivation of the initial covariance of the gridded data as described in Section 3. Using VisuShrink noise-proportional thresholds now gives the result shown in Figure 10. Again a wavelet basis with six vanishing moments was used. It can be seen that the spurious high frequency effects have been removed, that the initial part of the curve is no longer oversmoothed, and that the damped oscillations towards the end of the time period are now visible. A comparison with Figure 6 of Silverman (1985) is instructive; as one might hope, the 'elbow' near

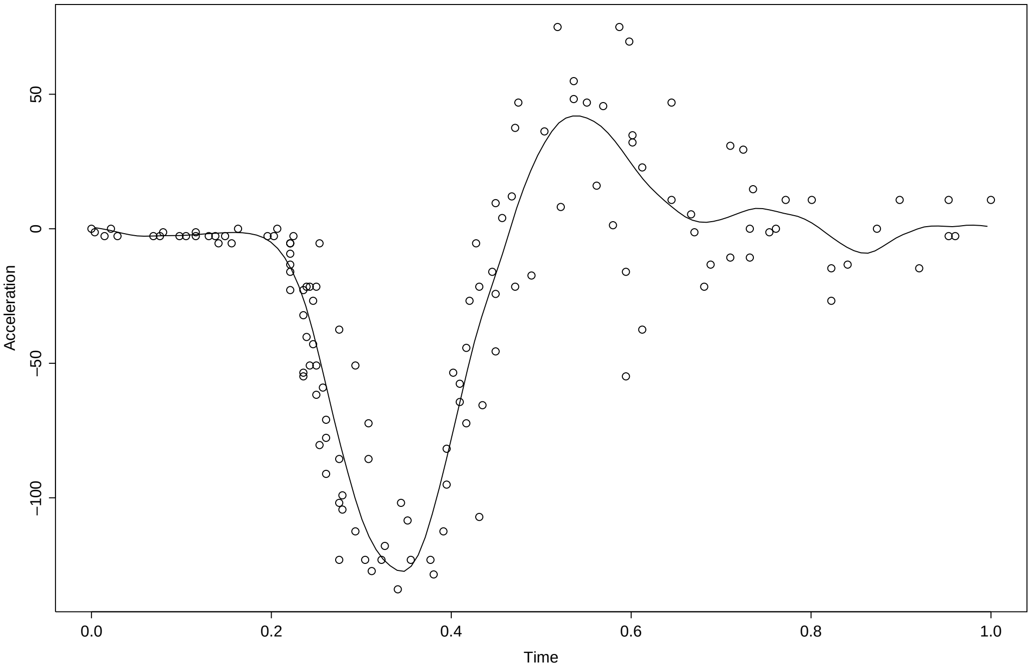

FIGURE 10. Crash data with wavelet estimator for heteroscedastic data, constructed using local estimates of the variance of individual data points, and VisuShrink noise-proportional thresholds. A wavelet basis with six vanishing moments was used.

time 0.2 is better fitted by the wavelet method, and the wavelet method also avoids some of the smoothing out of the overall minimum of the acceleration near time 0.35.

In the example we have presented, the individual variances had to be estimated from the data themselves. Of course the basic methodology is equally applicable if the variances are known, or can be estimated from external considerations.

7.2. **Correlated data.** Another obvious use of the algorithm we have set out is in the processing of data with known covariance structure, whether stationary or nonstationary. To provide a simple example of such data, we considered a segment of the synthetic ion channel gating data described in Johnstone and Silverman (1997). These data, due to Eisenberg and Levis (see Eisenberg, 1994) are designed to represent the challenges posed by real ion channel gating data. The true signal is a step function taking values 0 and 1, corresponding to closings and openings of single membrane channels in cells. It should be stressed that these are not simulated data in the usual statistical sense, but are synthetic data carefully constructed by practitioners with direct experience of the collection and analysis of real data.

For data of this kind, it is reasonable to suppose that the variance structure is stationary and known; in constructing their synthetic data, Eisenberg and Levis used known properties of laboratory data and of the instrumentation used for filtering these data in practice. Working from a very long 'noise' sequence provided by the authors, we estimated the noise variance to be 0.8 and the autocorrelations to be 0.31, -0.36, -0.26, -0.08 at lags 1 to 4 respectively, and zero for larger lags. A section of the original data is plotted in Figure 4 of Johnstone and Silverman (1997); the standard deviation of the noise is nearly 1, and it is difficult to detect the changes in overall level by eye.

The segment of the first 2048 data values was examined in more detail. The wavelet transform of the data was thresholded using VisuShrink noise-proportional thresholds at levels 7 and above. The Daubechies extremal phase wavelet of order 6 was used. As a final step, the estimated function was rounded to the nearest integer, which was always 0 or 1. In Figure 11 we show the 'true' signal and the signal estimated by this procedure. The number of discrepancies between the true signal and the estimate is 54 out of 2048, a 2.6% error rate which is far better than any performance obtained by Johnstone and Silverman (1997) for a standard wavelet transform. It is also interesting to note that the pattern of transitions between 0 and 1 is well estimated; the only effects that are missed are three sojourns in state 0, each of length 2.

Johnstone and Silverman (1997) obtained considerable improvements by the use of a translation-invariant method (see Coifman and Donoho, 1995, or Nason and Silverman, 1995). This essentially constructs estimates for every position of the wavelet grid, and then averages. We have not, in this work, considered translation-invariant transforms in any detail, but for this case we tried a translation-invariant prescription using periodic boundary conditions, a primary resolution level of 7, and thresholds proportional to standard deviation. If the VisuShrink constant of proportionality is used, the results are not as good as in the simple wavelet transform case. However if these thresholds are divided by 2, the misclassification rate improves to 47 out of 2048, which actually surpasses Johnstone and Silverman's error rate, but only by a small margin. It is interesting that a smaller threshold is desirable; this is because of the smoothing effect of the averaging step in the recovery part of the translation-invariant procedure.

8. CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

In this paper we have set out an algorithm for finding the variances and within-level covariances of the wavelet transform starting from a rather general covariance structure. Several possible applications of this method have been considered, but obviously there are many avenues that we have not explored.

For example, generalized linear models (GLIMs) (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) have been one of the major advances in statistical methodology of the last 25 years. Nonparametric smoothing ideas can be incorporated into the GLIM framework by assuming one or more of the dependences on the covariates to be a curve rather than simply linear. For a detailed synthesis of work in this area, and further extensions, see Chapter 5 and 6 of Green and Silverman (1994). It is explained there how to fit nonparametric GLIMs by solving a sequence of weighted regression problems, each having the same structure as a standard nonparametric regression problem with unequal variances. Because the GLIM fitting is posed as a maximum likelihood problem, it is very natural to

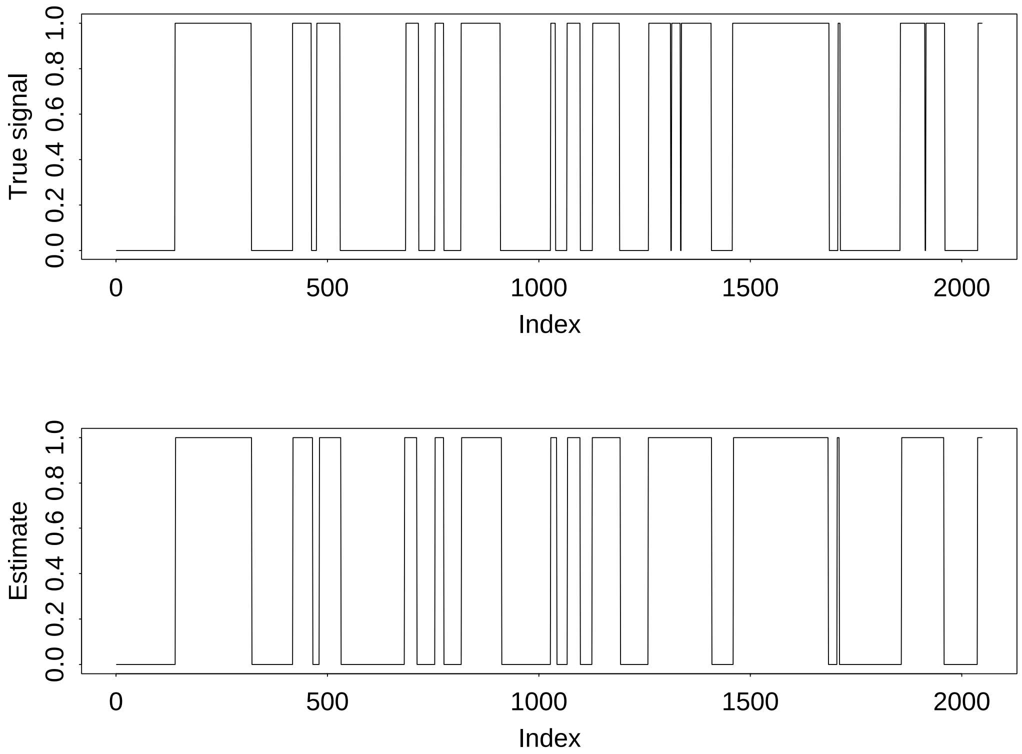

FIGURE 11. Upper panel: The 'true' signal synthesized by Eisenberg and Levis, plotted for time points 1 to 2048. Lower panel: Estimate obtained by noise- proportional thresholding at levels 7 and above, as described in the text.

use a penalized likelihood approach, typically using a quadratic penalty such as integrated squared second derivative. The algorithm we have set out in this paper allows the possibility of using a wavelet curve estimate at each stage instead; this would allow for the fitting of dependences whose behavior had inhomogeneous smoothness properties. Because of the nonlinear nature of the wavelet smoothing, there may be problems with the convergence of this iteration, and the practical and theoretical investigation of this convergence is left as a subject for future research.

In this paper, we have considered a range of ideas including irregular data, nonstationary dependence, correlated data and robust methods. For the most part, these have been considered separately from one another, but another area of investigation is a synthesis between them. Conceptually, it is fairly obvious how one would proceed, but the combination of the different aspects may well need care in practice.

In our application of the algorithm we have almost entirely concentrated on the variances of the individual wavelet coefficients, while the algorithm itself also yields a great deal of information about covariance. Even though the wavelet transform often has a decorrelating effect (see, for example, Johnstone and Silverman, 1997, Section 2.2) it would be interesting to devise ways of processing the coefficients in a way that makes use of our knowledge of their correlation structure. This may well be more burdensome computationally, but would possibly produce more accurate statistical estimates.

Finally, we have concentrated on the one-dimensional case, but wavelets are of growing importance in the analysis of image data. The basic principles of our method can be easily extended to deal with two-dimensional wavelet transforms of data showing short-range correlation. Of course, the operational details are likely to depend on the specific field of application, but the need for efficient algorithms is likely to be even more crucial than in the one-dimensional case.

9. References

Abramovich, F. and Silverman, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika*. in press.

Antoniadis, A. and Pham, D. T. (1997). Wavelet regression for random or irregular design. Technical report, Laboratory of Modelling and Computation, IMAG, Grenoble, France.

Bruce, A., Donoho, D. L., Gao, H.-Y., and Martin, R. (1994). Smoothing and robust wavelet analysis. In *Proceedings CompStat*, Vienna, Austria.

Chui, C. K. (1992). *An Introduction to Wavelets*. Academic Press, Inc., San Diego.

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

Davies, L. and Gather, U. (1993). The identification of multiple outliers (with discussion). *Journal of the American Statistical Association*, 88:782–801.

Donoho, D. L. (1993). Nonlinear wavelet methods for recovery of signals, images, and densities from noisy and incomplete data. In Daubechies, I., editor, *Different Perspectives on Wavelets*, pages 173–205. American Mathematical Society.

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

Donoho, D. L. and 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., and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *Journal of the Royal Statistical Society Series B*, 57:301–369.

Donoho, D. L. and Yu, T. P. Y. (1998). Nonlinear "wavelet transforms" based on median-thresholding. *SIAM journal of mathematical analysis*. submitted for publication.

Eisenberg, R. (1994). Biological signals that need detection: Currents through single membrane channels. In Norman, J. and Sheppard, F., editors, *Proceedings of the 16th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 32a–33a.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society Series B*, 59:319–351.

Kwong, M. K. and Tang, P. T. P. (1994). W-matrices, nonorthogonal multiresolution analysis, and finite signals of arbitrary length. Technical Report MCS-P449-0794, Argonne National Laboratory.

Lenarduzzi, L. (1997). Denoising not equispaced data with wavelets. Technical Report IAMI 97.1, Istituto Applicazioni Matematica ed Informatica C.N.R., via Ampere 56, 20131 Milano, Italy.

Mallat, S. G. (1989). Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Transactions of the American Mathematical Society*, 315:69–89.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London.

Meyer, Y. (1992). *Wavelets and Operators*. Cambridge University Press, Cambridge.

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

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, 135:370–384.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society Series B*, 47:1–52.

Smith, M. J. T. and Eddins, S. L. (1990). Analysis/synthesis techniques for subband image coding. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38:1446–1456.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9:1135–1151.

Taswell, C. and McGill, K. C. (1994). Wavelet transform algorithms for finite duration discrete-time signals. *ACM Transactions on Mathematical Software*, 20:398–412.