# 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 that contain outliers. The core of the methodology is an algorithm for finding all the variances and within-level covariances in the wavelet table of a sequence with given covariance structure. If the original covariance matrix is band limited, then the algorithm is linear in the length of the sequence.

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. Various thresholding methods are discussed and investigated. Exact risk formulae for the mean square error of the methodology for given design are derived. Good performance is obtained by noise-proportional thresholding, with thresholds somewhat smaller than the classical universal threshold.

Outliers in the data can be removed or downweighted, and aspects of such robust techniques are developed and demonstrated in an example. Another natural application is to correlated data, where the covariance of the wavelet coefficients is not due to an initial grid transform but is an intrinsic feature. 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, Martin Löwendick and the referees for their helpful comments.

## 1. INTRODUCTION

1.1. **Background and main result.** In statistics, wavelet methods have been most widely studied in the non-parametric regression problem of estimating a function $f$ on the basis of observations $y_i$ at time points $t_i$, modelled as

$$y_i = f(t_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n \tag{1}$$

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

With some notable exceptions, the current literature mainly deals with $n$ a power of 2, independent and identically distributed errors $\varepsilon_i$, and equally spaced points $t_i$. Our methodology allows all these assumptions to be relaxed, but 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 $\varepsilon_i$.

Most wavelet-based methods use the discrete wavelet transform (DWT) described by Mallat (1989). 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 the data points $y_i$, and the variance matrix of $c_J$ is a multiple of the identity matrix. Because the DWT is an orthogonal transform, the wavelet coefficients are also uncorrelated with equal variances. Johnstone and Silverman (1997) have considered the case of wavelet thresholding where the noise is correlated but stationary. The variances of the wavelet coefficients at each level are then identical, but differ between levels, and so the coefficients can be thresholded level by level.

But what if $c_J$ has more general, and not necessarily stationary, variance matrix $\Sigma$? In general the DWT coefficients will be heteroscedastic and correlated. We set up an algorithm yielding all the variances and within-level covariances of the DWT for a wide range of variance matrices $\Sigma$. Provided $\Sigma$ is band-limited, the algorithm will be linear in $2^J$. This algorithm is the core of our methodology and has very broad potential uses. We present it and discuss its complexity properties in Section 2, before applying it in specific regression contexts, relaxing many of the classical assumptions.

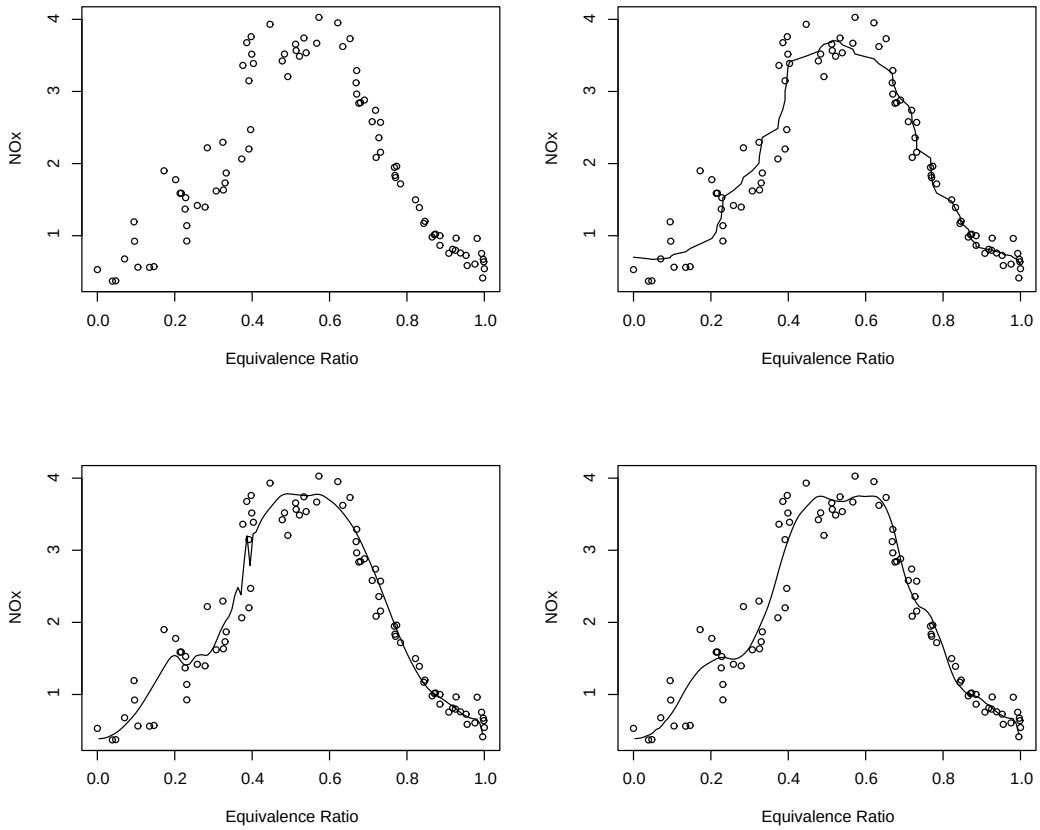


FIGURE 1. Eighty-eight measurements of exhaust from burning ethanol with three wavelet estimators. A wavelet basis with five vanishing moments was used. Top right: data regarded as lying on a regular grid, and standard universal thresholding wavelet estimator applied. (The data is extended to length 128 by reflection.) Bottom left: our method for irregular data assuming equal variances. Bottom right: our method for irregular data with local estimation of the variance. Universal thresholds are used in each case; for the bottom figures these are noise-proportional.

1.2. **An example.** Before reviewing our regression methodology, we present a particular example. Figure 1 shows a data set of Brinkmann (1981) that has been analyzed extensively, for example by Hastie (1992) and Cleveland et al. (1992). The data consist of 88 measurements from an experiment in which ethanol was burned in a single cylinder automobile test engine. The concentration of the sum of nitric oxide (NO) and nitrogen dioxide ($NO_2$) in engine exhaust, normalized by the work done by the engine, is related to the equivalence ratio, a measure of the richness of the air/ethanol mix. Because of the nature of the experiment, the observations are not available at equally-spaced design points, and the variability is larger for low equivalence ratios. The first curve is obtained by the naive approach of using only the

ranks of the $t_i$ but then applying a standard wavelet approach to obtain an estimate of $f$ at the points $t_i$. The second and third estimates are obtained using the methods of this paper. The second estimate uses the methodology of Section 4.1 to deal with the irregularity of the $t_i$, but estimates the variance globally. Generally, the performance of the estimator is much better, but the clearly spurious high-frequency feature near 0.4 has survived even the high threshold that is used in the procedure. The third estimate uses the method discussed in Section 7 to handle the heteroscedasticity in the data. It can be seen that a change in slope near 0.6 is well estimated, without other high-frequency effects intruding into the estimate.

1.3. **Review of our regression methodology.** Our regression method for generally positioned $t_i$ falls into three main parts, developed in Sections 3 and 4. First, if necessary, we 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 covariance matrix $\Sigma$ of the gridded values will in general be nonstationary but band-limited. A general covariance matrix may also arise from correlated or heteroscedastic data. The second phase is to apply the DWT to the vector $\tilde{y}$, and to use our algorithm to find its within-level covariance structure. The third phase is to threshold the DWT taking into account the heteroscedasticity of the coefficients, and to invert to find the estimate itself. Simulated examples indicate that a good approach is to use a threshold for each coefficient proportional to its standard deviation, and to use the SURE (Stein unbiased risk estimate) approach to determine the constant of proportionality. Theoretical support is provided by Johnstone and Silverman (1997).

The performance of the approach is investigated in Section 5, mostly by exact risk calculations rather than simulation. The use of the method in a robust procedure is set out and discussed in Section 6. In Section 7, we consider 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 our methodology is available from the home page of the first author.

## 2. CALCULATION OF THE VARIANCES OF WAVELET COEFFICIENTS

2.1. **Linear filters and the DWT algorithm.** We first review elementary aspects of filters and wavelets, partly to fix notation. Linear filters defined by sequences $f$ with finitely many nonzero elements are denoted

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

If $x$ is a finite vector, the definition of $\mathcal{F}x$ depends on the treatment at the boundaries; common choices are periodic continuation, reflection at the boundaries, or the boundary wavelet construction of Cohen et al. (1993).

The *binary decimation* operator $\mathcal{D}_0$ chooses every even member of a sequence:

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

Assume that $\psi$ is a finite-support mother wavelet of order $m$, with corresponding scaling function $\phi$. Standard references for wavelets include Meyer (1992), Daubechies (1992) and Chui (1992). The wavelet functions $\psi_{jk}$ defined by

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

form an orthonormal basis of $L^2(\mathbb{R})$, and are orthogonal to polynomials of degree up to $m$.

There is a sequence $(h_k)$ such that

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

where $g_k = (-1)^k h_{1-k}$. Because $\phi$ has bounded support, $h_k$ is zero outside the range $0 \le k < N$ for some integer $N$; denote by $\mathcal{G}$ and $\mathcal{H}$ the linear filters defined by $(g_k)$ and $(h_k)$ respectively.

To carry out the DWT, let $c^J = y$. Recursively for $j = J-1, \ldots, 0$, define

$$c^j = \mathcal{D}_0\mathcal{H}c^{j+1} \quad \text{and} \quad d^j = \mathcal{D}_0\mathcal{G}c^{j+1}. \tag{2}$$

Let $n_j$ denote the length of the vectors $c^j$ and $d^j$. For periodic boundary conditions, $n_j = 2^j$. The coefficients $d^{J-1}, \ldots, d^0, c^0$ make up the DWT of the data $y$. Regarding these as a single

vector $w$, we write $w = \mathcal{W}y$ where $\mathcal{W}$ is an orthogonal matrix in the periodic case. The algorithm described allows $w$ to be found in $O(N2^J)$ operations. We denote 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. **An algorithm for the computation of the variances of wavelet coefficients.** Consider the DWT of a vector $\tilde{y}$ of length $2^J$ whose elements have general covariance matrix $\Sigma$. We now set out an algorithm for finding the variances of all the wavelet coefficients $w$. In the case where $\Sigma$ is a band matrix, the number of operations will be linear in the length of $\tilde{y}$.

Let $\Sigma^j$ denote the variance matrix of $c^j$ and $\tilde{\Sigma}^j$ that of $d^j$. Then $\Sigma^J = \Sigma$ by definition. From the recursion (2), for each $j = J-1, J-2, \ldots, 0$,

$$\Sigma^j = H^{j+1}\Sigma^{j+1}(H^{j+1})^T \qquad \text{and} \qquad \tilde{\Sigma}^j = G^{j+1}\Sigma^{j+1}(G^{j+1})^T. \tag{3}$$

This gives 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.

A key aspect of our algorithm is the use of sparsity in the recursions (3). Each row of $H_{j+1}$ and $G_{j+1}$ is zero except in at most $N$ consecutive positions. Writing the first recursion of (3) in the two stages

$$A = H^{j+1}\Sigma^{j+1} \quad \text{and} \quad \Sigma^j = H^{j+1}A^T, \tag{4}$$

it follows that the total complexity of the algorithm is $O(N2^{2J})$ even for a general matrix $\Sigma$. If $\Sigma$ is a band matrix, then considerable additional economies are possible, as we show in the next section.

Vannucci and Corradi (1999) have, independently, expressed the recursion (3) in terms of the two-dimensional discrete wavelet transform of the matrix $\Sigma^{j+1}$. They thereby obtain an algorithm with complexity quadratic in the length of $\tilde{y}$; this is then used to develop a methodology for Bayesian wavelet shrinkage with priors that allow correlation among the wavelet coefficients.

2.3. **Computational complexity for bandlimited variance matrices.** For simplicity, only the periodic boundary condition case is considered in detail. Symmetric boundary conditions and boundary wavelets are slightly more complicated technically and notationally, but the resulting computational complexity is of the same order. Define the bandwidth $b_j$ to be the smallest integer such that $\Sigma^j_{i,m} = 0$ for all $i, m \in \{0, \ldots, n_j - 1\}$ with $b_j < |i - m| < n_j - b_j$, so that the bandwidth of a diagonal matrix is 0. For each $j$, we show that any band structure of $\Sigma_{j+1}$ is inherited by $\Sigma_j$. It will follow that the complexity of the proposed algorithm is $O(2^J b)$, where $b = \max(b_J, N)$ and $b_J$ is the bandwidth of $\Sigma^J$.

By symmetry, $\Sigma^j$ has at most $(b_j + 1)n_j$ entries that have to be calculated. Treating sums in the indices modulo $n_j$, the first sum in (3) can be reduced to

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

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

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

By induction on decreasing $j$, $b_j < b$, as required. By a similar argument using the second recursion of (3), the bandwidth of $\tilde{\Sigma}^j$ is subject to the same bound.

In practice further economy is possible; before calculating $\Sigma^j$ and $\tilde{\Sigma}^j$, we determine the actual bandwidth $b_{j+1}$ of $\Sigma^{j+1}$, and hence a tighter bound than $b$ on $b_j$. Considering the recursion in the form (4), each column of $\Sigma^{j+1}$ has nonzero elements in at most $1 + 2b_{j+1}$ positions, so 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 $N + 2b_{j+1}$ positions, so the complexity of the calculation of $\Sigma^j$ is $O(\{N^2 + Nb_j\}2^j) = O(Nb2^j)$. Summing over $j$ shows that the overall complexity of the algorithm is $O(Nb2^J)$. 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.

## 3. PREPROCESSING UNEQUALLY TIME-SPACED DATA

3.1. **Relaxing the basic assumptions.** The standard wavelet approach to the non-parametric regression problem (1) requires $n$ to be a power of two, and the $t_i$ to be equally spaced. One method of relaxing the first requirement is to extend the given data set to length $2^J$, by periodic extension or symmetric reflection (Smith and Eddins 1990). Though slightly artificial, this simple technique usually works well. Another approach (Kwong and Tang, 1994 , Taswell and McGill, 1994) is to construct alternatives to the fast DWT that deal with sets of arbitrary length. These techniques are of particular interest in image compression.

Relaxing the equal spacing in time raises more problems. One could use only the ranks of the $t_i$ and apply the usual threshold routines direct to the $y_i$, yielding an estimate of $f$ at the points $t_i$. Unfortunately, wavelet representations of irregularly spaced function evaluations are not in general as economical as they are if the time structure is regular; 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; see also Cai and Brown (1998). The poor performance of this procedure in Figure 1 may also be a result of this effect.

Lenarduzzi (1997) also notes that this method does not yield 'graphically pleasant' results, and suggests a modification involving spline interpolation on a small subset of time points. The coefficients 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. Cai and Brown (1998) assume that the empirical distribution of the time points approximates to a distribution with density $g_1$. They give a theoretical analysis of a method that uses a related transformation of the time domain, but essentially weights the original data to take account of this. Antoniadis and Pham (1998) make similar assumptions, and use of wavelet estimates of $fg_1$ and $g_1$. Their approach does not seem to have the properties which are expected of wavelet shrinkage estimators in the equally spaced case, but gives results very similar to linear estimators such as ordinary kernel estimators and spline smoothers.

3.2. **Mapping unequally spaced data to a regular grid.** Our method is simple and is easily combined with other developments in wavelets such as translation-invariant wavelet transforms and wavelet packets. In addition it 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. This is in contrast to Hall and Turlach (1997) who make the additional assumption that the design points are drawn from a random distribution, and then use an overall bound on the variance of the wavelet coeffi cients. We begin by choosing a fi nest resolution level $J$, usually the smallest integer such that $n \leq 2^J$.

Defi ne the grid points $\tilde{t}_k = (k + \frac{1}{2})2^{-J}$ where $k \in \{0, \ldots, 2^J - 1\}$. We calculate the gridded values $\tilde{y}_k$ by applying a linear transformation $R$ to the original data: $\tilde{y} = Ry$. One possibility is to obtain $\tilde{y}_k$ by linear interpolation of the original data, but in general this will not make use of all the original data points. Instead, we used the following procedure: For each subinterval $I_k = [k2^{-J}, (k + 1)2^{-J}]$ linear regression was applied to the observations that lie in this subinterval. If there was no observation in $I_k$ left of $\tilde{t}_k$, then the nearest data point to the left was also included in the linear regression. The right-hand side was treated analogously. Finally, the regression line was evaluated at $\tilde{t}_k$. Other grid transformations could be considered, such as higher-order local polynomial regression or interpolation. A detailed comparison between different gridding methods is left as a subject for future research.

The vector $\tilde{y}$ has length a power of two, so can be dealt with by standard DWT methods. If the original observations $y$ are uncorrelated with variance $\sigma^2$, then the covariance matrix $\Sigma$ 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 $[(k+1)2^{-J}, l2^{-J}]$. The bandwidth of $RR^T$ will essentially depend on the largest gap in the $t_i$.

We can now 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 $\sigma^2$ is not known, then the same algorithm starting with the matrix $RR^T$ will yield the variances and covariances up to an unknown common factor.

The extension to more general distributions of the original observations is straightforward. If the $y$ have variance matrix $\Sigma_Y$, then $\Sigma = R\Sigma_Y R^T$. If $\Sigma_Y$ is a diagonal matrix with unequal entries, then the bandwidth of $\Sigma$ will be the same as in the homoscedastic case. If $\Sigma_Y$ is a more general band matrix, then $\Sigma$ will 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 $\Sigma_Y$ in the general case. However, we will see in Section 7 two promising examples with heteroscedastic and correlated data.

## 4. THRESHOLDING HETEROSCEDASTIC WAVELET COEFFICIENTS

4.1. **The general approach.** Assume that $\tilde{f}$ is the $2^J$-vector of values $f(\tilde{t}_\ell)$ and that $\tilde{w}_{jk}$ is the DWT of $\tilde{f}$. Suppose that we have an array $d_k^j$ of observations of the $\tilde{w}_{jk}$ corrupted by heteroscedastic noise, and 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 as described in Section 3.2. The standard 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 apply a thresholding function $\eta$ to each wavelet coefficient yielding $\hat{w}_{jk} = \eta(d_k^j, \tau_{jk})$. The inverse DWT then gives the estimate $\hat{f} = \mathcal{W}^T\hat{w}$ for $\tilde{f}$. Here $\eta$ is either the soft thresholding function $\eta_S$ or the hard thresholding function $\eta_H$:

$$\eta_S(d_k^j, \tau) = \mathrm{sgn}(d_k^j)(|d_k^j| - \tau)_+ \qquad \eta_H(d_k^j, \tau) = d_k^j \cdot I\{|d_k^j| \geq \tau\}.$$

4.2. **Universal thresholding.** A natural approach, undergirded by theoretical work of Johnstone and Silverman (1997), is to choose each threshold $\tau_{jk}$ proportional to its standard deviation $\sigma_{jk}$. However, the noise level $\sigma$ is usually unknown. Generalizing Donoho *et al.* (1995), to take account of the unequal variances, $\sigma$ can be estimated by the median absolute deviation

of the normalized coeffi cients $d_k^{J-1}/\sqrt{\gamma_{J-1,k}}$ divided by 0.6745. Also, some of the $\gamma_{J-1,k}$, and the corresponding $d_k^{J-1}$ may be zero, for example when all grid points that have influence on the coeffi cient $d_k^{J-1}$ lie between two original observations, because of the property of vanishing moments and the linear interpolation that is used in the grid transform. Therefore we disregard coeffi cients that are numerically equal to zero. We then set

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

for all $j$ and $k$, and consider thresholds of the form

$$\hat{\tau}_{jk} = \hat{\sigma}_{jk}\tau \tag{5}$$

for some suitably chosen $\tau$. Setting $\tau = \sqrt{2\log n}$ then gives the the universal threshold or *VisuShrink* approach; this does not aim to minimize the mean square error, but tries to produce reconstructions that are 'noise-free', at least in the wavelet domain.

4.3. **An unbiased risk approach.** Another possible method for threshold selection, based on Stein (1981), was introduced to wavelet methodology as *SureShrink* by Donoho and Johnstone (1995). Let $w^*$ denote the vector of wavelet coeffi cients 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), the quantity

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

may be used as an estimate of the risk $\mathbb{E}||\hat{w} - w^*||^2$, for soft thresholding for the thresholds $\tau_{jk} = \tau\hat{\sigma}_{jk}$ defi ned in (5). Neglecting errors in the estimation of the $\sigma_{jk}$, this is an unbiased estimate of the risk. There is also an approximation 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 justifi cation for its use. In minimizing $S(\tau)$, we follow previous authors and restrict attention to $\tau$ in the range $[0, \sqrt{2\log n}]$.

## 5. PERFORMANCE OF THE THRESHOLDING METHODS

5.1. **Exact risk formulae.** Let $\hat{f}$ and $\tilde{f}$ be the vectors of the values and estimates of $f$ on the grid $\tilde{t_j}$ of $2^J$ equally-spaced points at which $f$ is to be estimated, and let $f^*$ be the vector of the true values $f(t_i)$ at the $n$ original data points. Let $\hat{w} = \mathcal{W}\hat{f}$ and $\tilde{w} = \mathcal{W}\tilde{f}$. By the orthogonality of the DWT, the mean square error satisfies

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

where $\tau_{jk}$ are the individual thresholds, and $j_0$ is a 'cut-off-level', below which no thresholding is carried out. We consider here soft thresholding only. Hard thresholding can be analyzed similarly (Kovac, 1999).

Let $w^*$ be the DWT of the sequence $Rf^*$. The individual coefficient $(\mathcal{W}R(f^*+\varepsilon))_{jk}$ is normally distributed with mean $w^*_{jk}$ and variance $\sigma_{jk}$ that can be calculated with the algorithms 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 $\mu_2$ and variance $\sigma^2$. We then have

$$\mathbb{E}\{\mathrm{MSE}(\hat{f}, \tilde{f})\} = 2^{-J} \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j} \rho(\tau_{jk}; \tilde{w}_{jk}, w^*_{jk}, \sigma_{jk}). \tag{6}$$

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 (6). The function $\rho$ can be evaluated from its definition by making use of properties of the normal distribution, generalizing results for $\mu_1 = \mu_2$ given by Donoho and Johnstone (1994) and Abramovich and Silverman (1998) to obtain

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

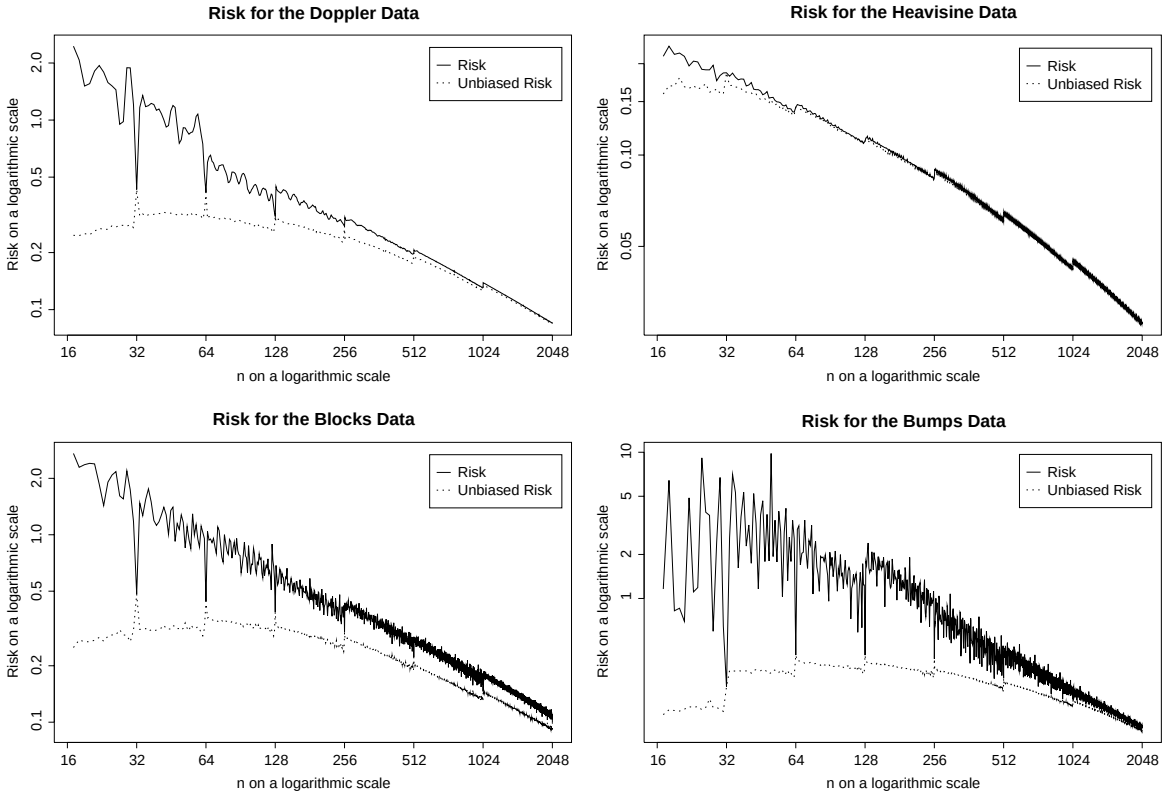


FIGURE 2. The risk for the test functions and *VisuShrink* on regular grids. Also shown as dashed lines is the "unbiased" risk which represents the amount of error which is caused by the actual thresholding step and does not contain the bias which is caused by the grid transform. The risk and the number of data points are plotted on logarithmic scales.

5.2. **Regular grids of arbitrary length.** We first consider exact risk formulae for our method where the $t_i$ are regularly placed but the number of points is not necessarily a power of two, using our algorithm to map the data to a grid of length $2^J$ and using standard DWT implementations. Our exact risk formula was applied to rescaled versions of the standard test signals of Donoho (1993) and Donoho and Johnstone (1994). For each $n$ in $\{17, \ldots, 2048\}$, calculations were performed for $n$ time points equally spaced in $[0, 1]$, 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 $\sigma$ was chosen as $0.35$ and the threshold chosen with *VisuShrink*, 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 2. It can be seen 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 the variation with $n$ is different among the four test signals.

To investigate the behavior further, 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. This is plotted as the dashed lines in the figure. The modified risk is 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. 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—hence the downward blips in the exact risk for these values of $n$.

The Blocks signal is sensitive to small changes in the time structure for all values of $n$ because it has a large number of discontinuities. 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 different, even if $n_1$ and $n_2$ are close together, but for larger $n_1$ and $n_2$ the continuity of the function ensures that the approximations are similar. Finally, the low values for powers of two for the Blocks signal are also caused by the absence of bias.

Both risk functions exhibit steps near powers of two. These are caused by the dependence of the thresholds on the number of grid points; they increase each time $n$ crosses a power of two. To check this, a separate calculation was carried out using a fixed threshold for all values of $n$. The basic shape is preserved, but the jumps near powers of two were eliminated.

5.3. **Ideal thresholds.** When comparing different choices for the thresholds it is interesting to consider the minimal risk that can be obtained with any thresholds for a specific function $f$. Such ideal thresholds are not available in practice, because the signal is unknown, but they do give a reference point against which a practical threshold choice can be judged. We consider *ideal noise-proportional thresholds* restricted to the special form $\tau_{jk} = \tau\sigma_{jk}$. For known $f$,

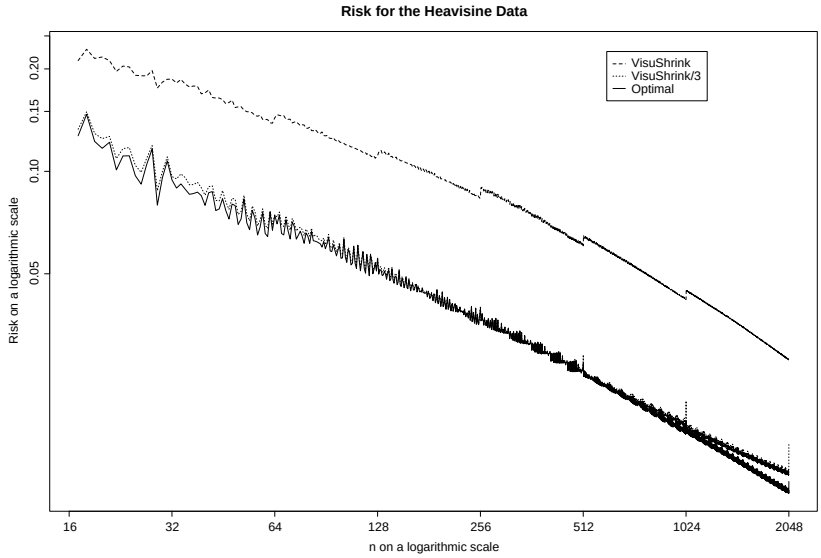


FIGURE 3. The risk for the Heavisine data and noise level 0.35 for three threshold choices on a regular grid of size $n$. The dashed and solid lines show the risk for the VisuShrink threshold choice and the ideal noise-proportional thresholds respectively. The dotted curve represents a threshold choice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

these thresholds can be found to desired accuracy by a numerical minimization of our exact risk formula.

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 3 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 ideal 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 4.7. This suggests a simple rule of thumb, to use the *VisuShrink* thresholds divided by 3. Figure 3 shows the resulting MSE which is very close to the MSE for ideal 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, where we also analyze the SureShrink method.

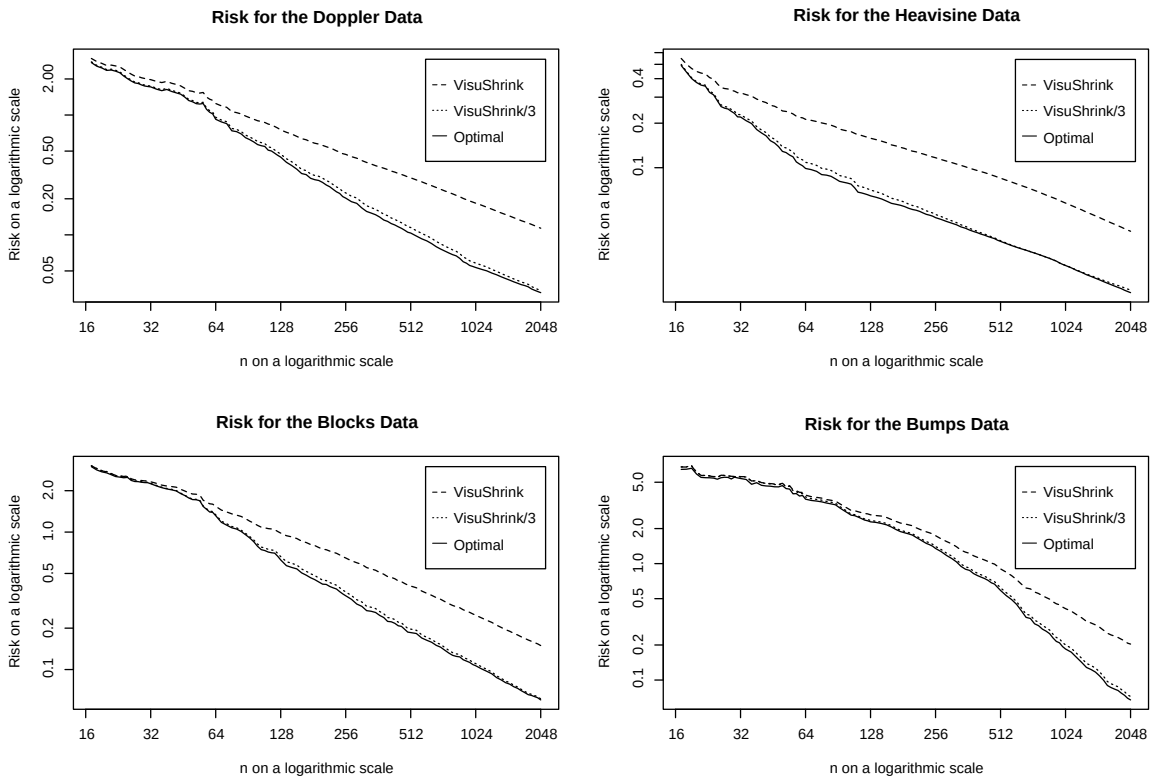


FIGURE 4. The expected MSE for the four test signals and $\sigma = 0.35$ for three threshold choices on irregular time structures of size $n$. The time points were independently drawn from a $Beta(2,2)$ distribution. The average of the exact risk over 20 replications is shown. Dashed and solid lines: risk for VisuShrink and ideal thresholds of the form $\tau\sigma_{j,k}$; dotted curve: risk for reduced Visushrink thresholds.

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 $\sigma$ was calculated, and the average over realizations was found. As in the previous section, this procedure was carried out for ideal and VisuShrink thresholds as well as for the reduced VisuShrink thresholds. The results are plotted for $\sigma = 0.35$ in Figure 4. The curves are much smoother than in Figure 2 and Figure 3, because the time structure is no longer fixed for given $n$. The slight irregularity in the curves seems to be due to sampling variation in the generation of the time points. The reduced VisuShrink thresholds still perform almost identically to the ideal thresholds.

| Signal | Time Data | IRREGSURE | IRREGVIS3 | RANKSURE | IRREGVIS | RANKVIS |
|---|---|---|---|---|---|---|
| Doppler | B(1,1) | .030 (.00034) | .034 (.00041) | .035 (.00046) | .118 (.00132) | .105 (.00100) |
| | B(2,2) | .073 (.00289) | .072 (.00230) | .077 (.00199) | .156 (.00198) | .156 (.00201) |
| | B(3,3) | .173 (.00791) | .154 (.00595) | .177 (.00647) | .225 (.00390) | .292 (.00653) |
| | B(4,4) | .317 (.01350) | .278 (.01034) | .335 (.01277) | .324 (.00689) | .509 (.01325) |
| Heavisine | B(1,1) | .014 (.00031) | .016 (.00030) | .015 (.00024) | .039 (.00047) | .032 (.00047) |
| | B(2,2) | .018 (.00051) | .019 (.00047) | .028 (.00120) | .051 (.00108) | .079 (.00247) |
| | B(3,3) | .045 (.00375) | .043 (.00295) | .088 (.00500) | .084 (.00266) | .238 (.00872) |
| | B(4,4) | .184 (.01432) | .152 (.01156) | .226 (.01195) | .149 (.00626) | .472 (.01225) |
| Blocks | B(1,1) | .064 (.00157) | .066 (.00152) | .063 (.00159) | .159 (.00176) | .137 (.00222) |
| | B(2,2) | .064 (.00155) | .066 (.00156) | .061 (.00145) | .172 (.00202) | .141 (.00226) |
| | B(3,3) | .091 (.00343) | .098 (.00364) | .085 (.00350) | .239 (.00471) | .179 (.00458) |
| | B(4,4) | .138 (.00772) | .153 (.00811) | .131 (.00741) | .352 (.00885) | .246 (.00819) |
| Bumps | B(1,1) | .070 (.00324) | .078 (.00354) | .077 (.00379) | .222 (.00511) | .218 (.00636) |
| | B(2,2) | .103 (.00588) | .111 (.00617) | .110 (.00685) | .268 (.00769) | .260 (.00957) |
| | B(3,3) | .160 (.00938) | .171 (.00982) | .166 (.01061) | .352 (.01088) | .320 (.01236) |
| | B(4,4) | .340 (.02109) | .358 (.01982) | .352 (.02064) | .558 (.01793) | .523 (.02152) |

TABLE 1. Comparison of the average Mean Square Error for five different thresholding techniques, as described in the text. For each model, 2048 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. The noise was white noise with $\sigma = 0.35$.

5.5. **Simulation comparison for randomly placed time points.** Finally, we report in Table 1 a simulation study in S-Plus to compare the MSE of five techniques for estimating a function. The three methods already considered in this section are included in the comparison, as well as two approaches based only on the ranks of the $t_i$, as described in Section 3.1. The methods are as follows:

**IRREGSURE:** Transform the data to a grid and DWT; soft threshold using noise-proportional *SureShrink* thresholds; apply inverse DWT.

**IRREGVIS:** Same as IRREGSURE, but use noise-proportional thresholds $\sigma_{jk}\sqrt{2\log n}$.

**IRREGVIS3:** As IRREGVIS, but use thresholds $\frac{1}{3}\sigma_{jk}\sqrt{2\log n}$.

**RANKSURE:** Apply standard *SureShrink* wavelet procedure to the data, taking only the ranks of the $t_i$ into account; perform the grid transform to get an estimate on a grid.

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

The methods based on *VisuShrink* cannot be expected to perform very well, because this threshold choice does not even attempt to obtain a low MSE. Except for 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. For the Blocks data, RANKSURE always attains the smallest MSE. 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

Donoho (1993) pointed out that standard thresholding techniques do not work very well if some of the observations may be considered to be outliers or when the noise follows a distribution with heavy tails. In the case of equally-spaced time points, two suggestions have been made for this problem. Bruce *et al.* (1994) use 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. Donoho and Yu (1997) construct a nonlinear multiresolution analysis based on a triadic grid; 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 'quick and dirty' approach more closely related to classical robustness methods. It 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. Classical thresholding cannot be used in such an approach, because the resulting data will not be equally spaced even if the original data were. Instead, our procedure for irregularly-spaced data can be used, and this is illustrated within a particular example.

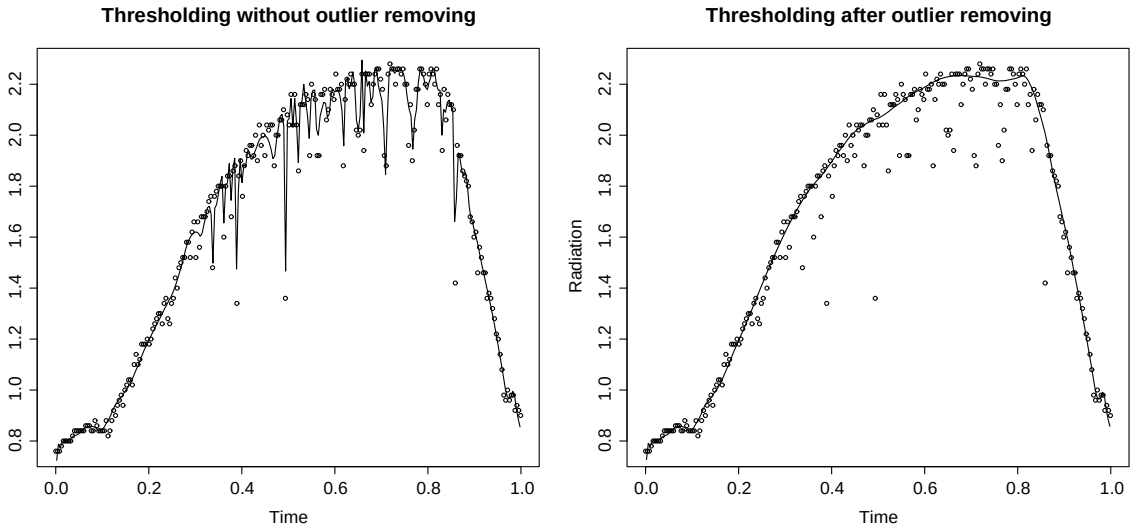


FIGURE 5. Balloon data with two wavelet estimators. Left panel: classical wavelet thresholding with VisuShrink thresholds. Right panes: Thresholding techniques for unequally spaced data applied after removing outliers.

Figure 5 shows data analyzed previously by Davies and Gather (1994). They are taken from a balloon which took measurements of radiation from the sun. The measurement device was occasionally cut off from the sun, causing individual outliers and large outlier patches. For this analysis we subsampled 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 VisuShrink method of Section 4.1 produces ugly curves like the one shown in the left panel of Figure 5. The curve exhibits several high frequency phenomena due to outliers. These have survived even the use of the VisuShrink threshold.

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 using the Haar basis.
2. For each data point the local median was calculated over a window containing 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.
3. The *VisuShrink* algorithm of Section 4.1 was applied to the modified data set, using 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 nondeleted data points were independent with variance $\hat{\sigma}_0^2$.

In Step 1, it would not be advisable to base the variance estimation on a higher-order wavelet basis expansion than the Haar basis, because the filters $\mathcal{G}$ and $\mathcal{H}$ would have wider support, so more wavelet coefficients would be contaminated by the outliers. Experiments showed that a re-estimation of the variance in Step 3 from the cleaned data set will typically underestimate the noise level, and so is not to be recommended.

The results can be seen in the right panel of Figure 5. Note that the abrupt changes in slope in the curve are well modeled, but the spurious high frequency effects have been removed.

## 7. HETEROSCEDASTIC AND CORRELATED DATA

7.1. **Heteroscedastic data.** A common problem in nonparametric regression is the handling of data with non-constant variance. Intuitively one should apply heavier smoothing in regions where the variance is larger, and an estimate of the local variance of the data will allow our methodology to be used. Our variance calculation algorithm will determine appropriate adjustments to the thresholds in different parts of the wavelet table. Rather than being completely prescriptive, we present a possible approach in the context of the example discussed in Section 1.2. In a similar situation Silverman (1985) used an iterative method to give estimates of the local variance and hence an improved estimate of the whole curve.

To give a noniterative wavelet-based method, we adopt an approach similar to suggestions of von Sachs and MacGibbon (1997) and Neumann and von Sachs (1997). 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 width $0.2$ around each point. These values for the variances of the individual data were then plugged into the derivation of the initial covariance of the gridded data. Using VisuShrink noise-proportional thresholds then gives the bottom right panel of Figure 1.

In this example, the individual variances were estimated from the data themselves. 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 our algorithm is for data with known covariance structure, whether stationary or nonstationary. A simple example of such data is 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. 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.

It is reasonable to suppose that the variance structure of the data is stationary and known; in constructing their synthetic data, Eisenberg and Levis used known properties of real laboratory data and filtering instrumentation. 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 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. VisuShrink noise-proportional thresholding was applied to the DWT of the data at levels 7 and above. The variances of the wavelet coefficients were calculated as described in Section 2, using the estimated autocorrelations to construct the covariance matrix of the data. 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 $1$. Figure 6 shows the 'true' signal and the estimate obtained. 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. Note that the pattern of

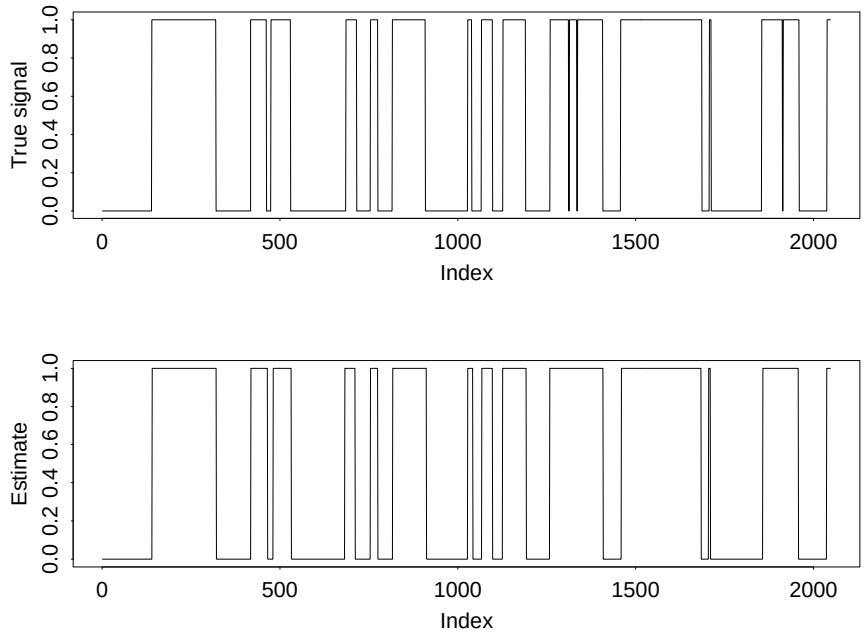


FIGURE 6. Upper panel: The 'true' signal synthesized by Eisenberg and Levis, plotted for time points 1 to 2048. Lower panel: Estimate obtained by noise- proportional thresholding at levels 7 and above, as described in the text.

transitions between 0 and 1 is well estimated; the only effects 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 (Coifman and Donoho, 1995; Nason and Silverman, 1995). This essentially constructs estimates for every position of the wavelet grid, and then averages. 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 best error rate, but only by a small margin. A smaller threshold is desirable 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

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.

A range of ideas including irregular data, nonstationary dependence, correlated data and robust methods have been considered, for the main part, separately from one another, and an interesting 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.

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 uses knowledge of the 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 will 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*, 85:115–129.

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

Brinkman, N. D. (1981). Ethanol—a single-cylinder engine study of effi ciency and exhaust emissions. *SAE Transactions*, 90:1410–1424.

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

Cai, T. and Brown, L. D. (1998). Wavelet shrinkage for nonequispaced samples. *Annals of Statistics*, 26:1783–1799.

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

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local regression models. In Chambers, J. M. and Hastie, T. J., editors, *Statistical Models in S*, pages 309–376. Wadsworth and Brooks, Pacifi c Grove, California.

Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, 1:54–81.

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

Davies, L. and Gather, U. (1993). The identifi cation 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. 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.

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

Hastie, T. J. (1992). Generalized additive models. In Chambers, J. M. and Hastie, T. J., editors, *Statistical Models in S*, pages 195–247. Wadsworth and Brooks, Pacifi c Grove, California.

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.

Kovac, A. (1999). *Wavelet Thresholding for Unequally Spaced Data*. PhD thesis, University of Bristol, Bristol.

Kwong, M. K. and Tang, P. T. P. (1994). W-matrices, nonorthogonal multiresolution analysis, and fi nite 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.

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

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fi tting (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.

Vannucci, M. and Corradi, F. (1999). Covariance structure of wavelet coefficients: Theory and models in a bayesian perspective. *Journal of the Royal Statistical Society Series B*, 61:971–986.

von Sachs, R. and MacGibbon, B. (1997). Nonparametric curve estimation by wavelet thresholding for with locally stationary errors. *Technical Report "Berichte der AG Technomathematik 179, University of Kaiserslautern"*. submitted for publication.

von Sachs, R. and Neumann, M. (1995). Beyond the gaussian i.i.d. situation. In Antoniadis, A. and Oppenheim, G., editors, *Wavelets and Statistics*, pages 301–329. Springer, New York.