MathSoft

Research Report
No. 41

# Wavelet Shrinkage for Unequally Spaced Data

**Sylvain Sardy, Donald B. Percival, Andrew G. Bruce,
Hong-Ye Gao and Werner Stuetzle**

Last revised

April 28, 1998

**Acknowledgements:** This work was partially supported by ONR contract N00014–93–C–0106 and by NSF Grant DMI–94–61370.

MathSoft, Inc.

1700 Westlake Ave. N, Suite 500

Seattle, WA 98109.9891, USA

Tel: (206) 283–8802

FAX: (206) 283–6310

E-mail: `sardy@stat.washington.edu`
`dbp@statsci.com`
`andrew@statsci.com`
`gao@statsci.com`
`wxs@stat.washington.edu`

ftp: `ftp.statsci.com`
(`pub/WAVELETS`)

# Wavelet Shrinkage for Unequally Spaced Data

**Sylvain Sardy[1,2], Donald B. Percival[1,3], Andrew G. Bruce[1], Hong-Ye Gao[1,4] and Werner Stuetzle[2]**

[1]MathSoft, Inc., 1700 Westlake Avenue North, Seattle, WA 98109–9891, U.S.A.

[2]Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195–4322, U.S.A.

[3]Applied Physics Laboratory, Box 355640, University of Washington, Seattle, WA 98195–5640, U.S.A.

[4]TeraLogic, Inc., 707 California Street, Mountain View, CA 94041–2005, U.S.A.

Correspondence to Dr. Donald B. Percival, MathSoft, Inc., 1700 Westlake Avenue North, Suite 500, Seattle, Washington 98109–9891, U.S.A.; Phone: (206)–283–8802, ext. 258; Fax: (206)–283–8691; E-mail: `dbp@statsci.com`

## Abstract

Wavelet shrinkage (WaveShrink) is a relatively new technique for nonparametric function estimation that has been shown to have asymptotic near-optimality properties over a wide class of functions. As originally formulated by Donoho and Johnstone, WaveShrink assumes equally spaced data. Because so many statistical applications (e.g., scatterplot smoothing) naturally involve unequally spaced data, we investigate in this paper how WaveShrink can be adapted to handle such data. Focusing on the Haar wavelet, we propose four approaches that extend the Haar wavelet transform to the unequally spaced case. Each approach is formulated in terms of continuous wavelet basis functions applied to a piecewise constant interpolation of the observed data, and each approach leads to wavelet coefficients that can be computed via a matrix transform of the original data. For each approach, we propose a practical way of adapting WaveShrink. We compare the four approaches in a Monte Carlo study and find them to be quite comparable in performance. The computationally simplest approach (isometric wavelets) has an appealing justification in terms of a weighted mean square error criterion and readily generalizes to wavelets of higher order than the Haar.



**Key Words**: Nonparametric Regression; Wavelet Transform; WaveShrink

# 1 Introduction

Suppose we observe data $Y_1, Y_2, \ldots, Y_n$ generated according to the model

$$Y_i = f(x_i) + \epsilon_i \qquad i = 1, 2, \ldots, n, \tag{1}$$

where $f$ is an unknown function, the $x_i$'s are fixed known sampling times satisfying $x_i < x_{i+1}$, and the $\epsilon_i$'s are independent and identically distributed random variables from the distribution $N(0,\ \sigma^2)$. The goal is to estimate the $f(x_i)$'s such that the estimates $\hat{f}(x_i)$ have small risk, where the risk is defined as

$$R(\hat{f}, f) = \frac{1}{n} \sum_{i=1}^{n} E\left(\hat{f}(x_i) - f(x_i)\right)^2. \tag{2}$$

Many techniques have been developed to estimate $f(x_i)$. When the spacings $x_{i+1} - x_i$ between sampling times are all equal, Donoho and Johnstone (1994) have recently proposed the use of wavelet shrinkage (WaveShrink), a wavelet-based technique for nonparametric function estimation that is particularly valuable for large amounts of data and for functions exhibiting locally nonsmooth behavior (e.g., jumps, cusps or peaks). On a theoretical level, WaveShrink has been shown to have very broad near-optimality properties. For example, WaveShrink achieves, within a factor of $\log n$, the optimal minimax risk over each functional class in a variety of smoothness classes and with respect to a variety of loss functions (Donoho et al. (1995)).

To date, the methodology and algorithms for WaveShrink have almost exclusively focused on equally spaced $x_i$'s – this is a severe restriction because many statistical applications do not have an equally spaced design. The goal of this paper is to evaluate the effectiveness of very simple schemes for extending WaveShrink to unequally spaced samples. We investigate four such schemes, all of which use piecewise constant interpolation in conjunction with the Haar wavelet. Our results establish a benchmark against which to evaluate more elaborate schemes. We demonstrate that, while the differences in the performance of the four schemes are small overall, the simplest scheme (isometric wavelets) has an appealing justification in terms of a weighted mean square error criterion and can be readily generalized to wavelets of higher order than the Haar.

There are several previous papers in the literature that focus on defining and computing the wavelet transform for unequally sampled data. Antoniadis et al. (1994) consider curve estimation using wavelet methods for unequally sampled data in which estimation is accomplished by keeping or eliminating all wavelet coefficients on a given scale. Delyon and Juditsky (1995) discuss fast algorithms for computing wavelet coefficients in the unequally sampled case. Sweldens (1995) proposes a 'lifting' scheme that defines a wavelet transform for arbitrary sampling schemes over one and higher dimensional surfaces (in fact, the isometric wavelets we introduce in section 3.1 can be formulated as a special case of lifting). Foster (1996) considers a weighted wavelet transform that takes irregular spacing into account. Antoniadis et al. (1997) propose replacing unequally sampled data by averages over regularly spaced bins and then applying the usual wavelet transform to the binned data. Scargle (1997b) formulates a similar idea in the context of the Haar wavelet transform. None of the above papers deal specifically with WaveShrink, but there are two previous papers that do. Scargle (1997a) gives an example of applying wavelet shrinkage with a universal threshold to wavelet coefficients computed from a binned radial velocity time series, but does not study the properties of this estimator. Hall and Turlach (1997) study wavelet shrinkage based on two different interpolation methods (local averaging and local linear interpolation) in conjunction with sampling on a dyadic grid and wavelets of possibly higher order than the Haar (sampling of interpolated data on a grid is also studied by Kovac and Silverman (1998), who developed fast algorithms for computing the variance and covariance of such samples). Sampling on a grid is one of the four schemes that we consider in this paper, but we do so only in the context of a much simpler interpolation scheme (piecewise constant interpolation between midpoints of the observations) in conjunction with the Haar wavelet.

The remainder of this paper is organized as follows. Following a review of Haar wavelet analysis for equally spaced data in section 2, we define four simple wavelet analysis techniques appropriate for unequally spaced data in section 3. We review the basics of WaveShrink in section 4 and adapt it in section 5 to work with the four analysis techniques, hence producing nonparametric function estimates through wavelet coefficient shrinkage. In section 6 we report the results of a simulation study comparing the approaches using synthetic signals studied by Donoho et al. (1995). To demonstrate how our methodology works on an actual unequally sampled time series, we estimate the light curve for a variable star in section 7. We summarize our main conclusions and discuss directions for further research in section 8. All results in this paper can be reproduced through software that is accessible by anonymous ftp: see section 9.

## 2 Haar Wavelet Analysis for Equally Spaced Data

In section 3 we propose four definitions for a wavelet transform, all of which extend the usual Haar wavelet transform so that it is suitable for denoising unequally spaced data. To motivate these definitions, we review here the Haar continuous wavelet transform (CWT) and note that, when the Haar CWT is applied to a function $\bar{f}$ that is a piecewise constant interpolation of the elements of a column vector $\mathbf{f}$ containing equally spaced samples from the function $f$, the resulting wavelet series is identical to the Haar discrete wavelet transform (DWT) of $\mathbf{f}$. This correspondence between the CWT wavelet series for $\bar{f}$ and the DWT of $\mathbf{f}$ will motivate our definitions of wavelet analysis appropriate for unequally spaced data.

Let $\mathbf{f} \equiv [\, f(x_1) \quad f(x_2) \quad \cdots \quad f(x_n)\,]^T$, where for this section we assume $x_i = i$. For simplicity assume $n = 2^J$ for some integer $J > 0$. Define the piecewise constant approximation $\bar{f}$ to the function $f$ as

$$\bar{f}(x) \equiv \begin{cases} f(x_i), & x_i - \frac{1}{2} \leq x < x_i + \frac{1}{2},\ i = 1, \ldots, n; \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

Let $j = 1, 2, \ldots, J$ be indices for scales, and let $k = 1, 2, \ldots, n/2^j$ be indices for locations within the $j$th scale. Define $n-1$ Haar mother wavelets as

$$\psi_{j,k}(x) \equiv \begin{cases} \frac{1}{\sqrt{2^j}}, & 2^j(k-1) + \frac{1}{2} \leq x < 2^j(k - \frac{1}{2}) + \frac{1}{2}; \\ -\frac{1}{\sqrt{2^j}}, & 2^j(k - \frac{1}{2}) + \frac{1}{2} \leq x < 2^j k + \frac{1}{2}; \\ 0, & \text{otherwise}; \end{cases} \tag{4}$$

and define a single Haar father wavelet (or scaling function) as

$$\phi_{J,1}(x) \equiv \begin{cases} \frac{1}{\sqrt{n}}, & \frac{1}{2} \leq x < n + \frac{1}{2}; \\ 0, & \text{otherwise}. \end{cases} \tag{5}$$

The wavelet series for $\bar{f}$ is formed from the inner products of $\bar{f}$ with the $\psi_{j,k}$'s and $\phi_{J,1}$, i.e.,

$$\langle \bar{f}, \psi_{j,k} \rangle \equiv \int_{\frac{1}{2}}^{n+\frac{1}{2}} \bar{f}(x)\psi_{j,k}(x)\, dx \quad \text{and} \quad \langle \bar{f}, \phi_{J,1} \rangle \equiv \int_{\frac{1}{2}}^{n+\frac{1}{2}} \bar{f}(x)\phi_{J,1}(x)\, dx. \tag{6}$$

Because it is piecewise constant by construction, the function $\bar{f}$ can be represented exactly as

$$\bar{f}(x) = \sum_{j=1}^{J} \sum_{k=1}^{n/2^j} \langle \bar{f}, \psi_{j,k} \rangle \psi_{j,k}(x) + \langle \bar{f}, \phi_{J,1} \rangle \phi_{J,1}(x).$$

Let us now place the $n$ values of the wavelet series into the $n$ dimensional column vector $\boldsymbol{\alpha}_{\mathbf{f}}$. The last element of $\boldsymbol{\alpha}_{\mathbf{f}}$ is taken to be $\langle \bar{f}, \phi_{J,1} \rangle$, while the first $n-1$ elements are the $\langle \bar{f}, \psi_{j,k} \rangle$'s ordered as follows: $\langle \bar{f}, \psi_{1,k} \rangle, k = 1, \ldots, n/2$; $\langle \bar{f}, \psi_{2,k} \rangle, k = 1, \ldots, n/4$; $\ldots$;

$\langle \bar{f}, \psi_{J-1,k} \rangle, k = 1, 2$; and $\langle \bar{f}, \psi_{J,1} \rangle$. For example, when $n = 4$ we have

$$\boldsymbol{\alpha}_{\mathbf{f}} = \begin{pmatrix} \langle \bar{f}, \psi_{1,1} \rangle \\ \langle \bar{f}, \psi_{1,2} \rangle \\ \langle \bar{f}, \psi_{2,1} \rangle \\ \langle \bar{f}, \phi_{2,1} \rangle \end{pmatrix}.$$

By construction, the vector $\boldsymbol{\alpha}_{\mathbf{f}}$ can be expressed as an orthonormal transformation of the vector $\mathbf{f}$, namely, the Haar DWT of $\mathbf{f}$; i.e., we can write $\boldsymbol{\alpha}_{\mathbf{f}} = W\mathbf{f}$, where $W$ is an $n \times n$ orthonormal matrix that defines the Haar DWT. For example, the DWT matrix for $n = 4$ is given by

$$W = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{w}_{1,1}^T \\ \mathbf{w}_{1,2}^T \\ \mathbf{w}_{2,1}^T \\ \mathbf{v}_{2,1}^T \end{pmatrix} \quad \text{and hence} \quad \boldsymbol{\alpha}_{\mathbf{f}} = \begin{pmatrix} \langle \mathbf{f}, \mathbf{w}_{1,1} \rangle \\ \langle \mathbf{f}, \mathbf{w}_{1,2} \rangle \\ \langle \mathbf{f}, \mathbf{w}_{2,1} \rangle \\ \langle \mathbf{f}, \mathbf{v}_{2,1} \rangle \end{pmatrix}, \tag{7}$$

where $\langle \mathbf{f}, \mathbf{w} \rangle \equiv \mathbf{f}^T\mathbf{w}$ is the usual inner product between two column vectors $\mathbf{f}$ and $\mathbf{w}$. Thus, although $\boldsymbol{\alpha}_{\mathbf{f}}$ was defined in terms of the Haar CWT of the function $\bar{f}$, it has an alternative interpretation as the Haar DWT wavelet coefficients for the vector $\mathbf{f}$.

# 3 Wavelet Analysis for Unequally Spaced Data

Let us now consider the case of unequally sampled data so that we assume just the ordering $x_i < x_{i+1}$, $i = 1, \ldots, n-1$. For convenience we also assume that $x_1 = 1$ and $x_n = n$ (this convention is handy because it forces the average spacing between adjacent $x_i$'s to be unity, thus greatly simplifying the mathematics that follow). The vector $\mathbf{f}$ now consists of samples of $f$ at the $x_i$'s. We start by defining $\bar{f}$ in a manner analogous to equation (3). Define

$x_0 = 0$ and $x_{n+1} = n + 1$, and let $m_i \equiv (x_i + x_{i+1})/2$, $i = 0, \ldots, n$, represent the mid-points of the augmented $x_i$'s. Let

$$\bar{f}(x) \equiv \begin{cases} f(x_i), & m_{i-1} \le x < m_i,\ i = 1, \ldots, n; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Note that, when $x_i = i$, we have $m_{i-1} = x_i - \frac{1}{2}$ and $m_i = x_i + \frac{1}{2}$, so the above definition is consistent with (3) for the equally spaced case. With $\bar{f}$ so defined, we now consider how to define Haar-like mother and father wavelets appropriate for unequally sampled data. In the next four subsections, we explore the following four ideas, each of which can be used with wavelet shrinkage.

- *Isometric Wavelets*: Here we define a Haar-like orthogonal (but not in general orthonormal) CWT such that the corresponding wavelet series for $\bar{f}$ is *identical* to the Haar DWT of $\mathbf{f}$. This scheme amounts to just taking the Haar DWT of $\mathbf{f}$, so formally we treat the data as if they were equally spaced; however, as shown below, this procedure has an appealing justification in terms of an isometry involving a risk measure with a non-Euclidean norm. A generalization of this scheme to wavelets other than the Haar is to apply any DWT to $\mathbf{f}$.

- *Asymmetric Haar*: Here we take the Haar-like orthogonal isometric wavelets and adjust the mother and father wavelets to define an orthonormal set of functions, which we call the 'asymmetric Haar' functions. The corresponding wavelet series for $\bar{f}$ can be interpreted as a (in general) nonorthonormal transform of the vector $\mathbf{f}$; however, the resulting transform is easy to compute and readily invertible. This scheme cannot be readily generalized beyond the Haar case.

- *Sampling on a Grid*: Here we sample the interpolated function $\bar{f}$ over a grid of equally spaced points and apply the Haar DWT to these samples. This scheme can be generalized by using other DWTs and other interpolation schemes, and it can readily handle sample sizes $n$ that are not a power of 2.

- *Exact Integration*: Here we form the wavelet series for $\bar{f}$ using the Haar CWT defined over the interval of support for $\bar{f}$. This scheme can be generalized to other CWTs and other interpolation schemes, and it can handle $n$'s that are not a power of 2.

## 3.1 Isometric Wavelets

Let $X$ and $Y$ be two random variables with joint density $p(x, y)$ and marginal densities $p_X(x)$ and $p_Y(y)$. Suppose we construct a rule $\tilde{f}(X)$ for predicting $Y$ given $X$. The mean squared prediction error for this rule is given by

$$E\{(\tilde{f}(X) - Y)^2\} = \int \mathrm{var}\{Y \mid x\}\, p_X(x)\, dx + \int \left(\tilde{f}(x) - E\{Y \mid x\}\right)^2 p_X(x)\, dx.$$

Suppose further that $X$ and $Y$ are related via a 'random sampling time' version of equation (1), namely, $Y = f(X) + \epsilon$, where $\epsilon$ is a random variable independent of $X$ with variance $\sigma^2$. The above then becomes

$$E\{(\tilde{f}(X) - Y)^2\} = \sigma^2 + \int \left(\tilde{f}(x) - f(x)\right)^2 p_X(x)\, dx.$$

The above suggests that the natural way to measure the distance between a function $f$ and its estimate $\tilde{f}$ is via a norm defined in terms of an inner product involving $p_X$ as a weighting function:

$$\langle g, h \rangle_{p_X} \equiv \int g(x) h(x) p_X(x)\, dx.$$

Note that, when the sampling times are fixed and equally spaced, then $p_X(x) = \frac{1}{n} \sum \delta_{x_i}(x)$ (where $\delta_a$ is the Dirac measure at $a$) is the measure of choice and leads to the usual definition of risk (cf. Equation 2); on the other hand, when the sampling times are random and hence unequally spaced, the above derivation suggests using a measure involving the density $p_X$ of the random variable $X$. In practice, we do not know $p_X$, but suppose we consider the $x_i$'s of equation (1) to be ordered observations from a random sample of size $n$ from $X$ (for this argument we drop the assumptions $x_1 = 1$ and $x_n = n$, and we now augment the $x_i$'s using $x_0 = x_1 - (x_n - x_1)/(n-1)$ and $x_{n+1} = x_n + (x_n - x_1)/(n-1)$). Under the assumption that the mid-points $m_i$ are distinct, we can then estimate $p_X$ using

$$\hat{p}_X(x) \equiv \begin{cases} \frac{1}{n(m_i - m_{i-1})}, & m_{i-1} \le x < m_i,\ i = 1, \ldots, n; \\ 0, & \text{otherwise} \end{cases}$$

(note that $\int \hat{p}_X(x)\, dx = 1$, as required). The weighted inner product thus becomes

$$\begin{aligned} \langle g, h \rangle_{\hat{p}_X} &= \sum_{i=1}^{n} \frac{1}{n(m_i - m_{i-1})} \int_{m_{i-1}}^{m_i} g(x)h(x)\, dx \\ &= \sum_{i=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} g(a_i + b_i y) h(a_i + b_i y)\, dy \\ &= \int_0^1 \tilde{g}(y)\tilde{h}(y)\, dy = \langle \tilde{g}, \tilde{h} \rangle \end{aligned}$$

where $a_i \equiv i m_{i-1} - (i-1) m_i$; $b_i \equiv n(m_i - m_{i-1})$;

$$\tilde{g}(y) \equiv \begin{cases} g(a_i + b_i y), & \frac{i-1}{n} \le y < \frac{i}{n},\ i = 1, \ldots, n; \\ 0, & \text{otherwise}; \end{cases} \tag{9}$$

$\tilde{h}$ is defined in an analogous manner; and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. Let $L^2[c, d]$ represent the set of all square integrable functions defined over the interval $[c, d]$. The above argument shows that the metric spaces $(L^2[m_0, m_n], \langle \cdot, \cdot \rangle_{\hat{p}_X})$ and $(L^2[0, 1], \langle \cdot, \cdot \rangle)$ are isometric, with an isometry given by the invertible mapping $\Phi : L^2[m_0, m_n] \to L^2[0, 1]$ defined by (9). Thus, any orthonormal wavelet basis for $L^2[0, 1]$ with respect to the Euclidean metric is equivalent to an orthonormal basis for $L^2[m_0, m_n]$ with respect to the empirical $\hat{p}_X$ metric, and hence expansion of the unequally spaced piecewise constant function $\bar{f}$ over a wavelet basis for $(L^2[m_0, m_n], \langle\cdot,\cdot\rangle_{\hat{p}_X})$ is equivalent to expansion of the equispaced piecewise constant function $\Phi(\bar{f})$ over a wavelet basis for $(L^2[0,1], \langle\cdot,\cdot\rangle)$. (Note that, as long as we use the same estimate for $p_X$, the same argument holds both for interpolation schemes other than piecewise constant and for wavelets other than the Haar.)

Based upon the above argument, we can define 'isometric' Haar mother and father wavelets $\psi^{(1)}_{j,k}$ and $\phi^{(1)}_{J,1}$ that treat unequally spaced data in a manner consistent with using an inner product weighted by our simple estimate of $p_X$. Isometric wavelets force the wavelet series for $\bar{f}$, say $\boldsymbol{\alpha}^{(1)}_{\mathbf{f}}$, to be identical to coefficients obtained from the usual Haar DWT of the vector $\mathbf{f}$. The appropriate definitions for the isometric Haar mother wavelets are

$$\psi^{(1)}_{j,k}(x) = \begin{cases} \frac{1}{\sqrt{2^j}}, & m_{2^j(k-1)} \le x < m_{2^j(k-\frac{1}{2})}; \\ -\frac{1}{\sqrt{2^j}}, & m_{2^j(k-\frac{1}{2})} \le x < m_{2^j k}; \\ 0, & \text{otherwise}, \end{cases}$$

while the father wavelet $\phi^{(1)}_{J,1}$ is defined as in equation (5). These wavelets are plotted in the left-hand column of figure 1 for a simple example involving $n = 4$ points. By construction $\psi^{(1)}_{j,k}$ is nonzero over exactly $2^j$ of the sampling points $x_i$, so the notion of scale maintained by these wavelets does not depend on the distances between the $x_i$'s. The mother wavelet $\psi^{(1)}_{j,k}$ takes on the same values as the corresponding wavelet for the usual Haar CWT, and in a similar fashion $\langle\psi^{(1)}_{j,k}, \psi^{(1)}_{j,k'}\rangle = 0$ for $k \ne k'$; however, the isometric Haar CWT differs from the usual CWT in the following ways. Let $\delta^+_{j,k}$ and $\delta^-_{j,k}$ be the widths of the strictly positive and negative portions of $\psi^{(1)}_{j,k}$:

$$\delta^+_{j,k} \equiv m_{2^j(k-\frac{1}{2})} - m_{2^j(k-1)} \quad \text{and} \quad \delta^-_{j,k} \equiv m_{2^j k} - m_{2^j(k-\frac{1}{2})}.$$

We then have

$$\int_{\frac{1}{2}}^{n+\frac{1}{2}} \psi^{(1)}_{j,k}(x)\,dx = (\delta^+_{j,k} - \delta^-_{j,k})/2^{j/2} \;\text{ and }\; \int_{\frac{1}{2}}^{n+\frac{1}{2}} \left[\psi^{(1)}_{j,k}(x)\right]^2 dx = (\delta^+_{j,k} + \delta^-_{j,k})/2^j.$$

Because in general $\delta^+_{j,k}$ is not equal to $\delta^-_{j,k}$ and also $\delta^+_{j,k} + \delta^-_{j,k}$ is not equal to $2^j$, it follows that the mother wavelets need not either integrate to zero or have unit norm and that $\langle \psi^{(1)}_{j,k}, \psi^{(1)}_{j',k'} \rangle$ need not be zero when $j$ differs from $j'$. Thus the $\psi^{(1)}_{j,k}$'s and $\phi^{(1)}_{J,1}$ do not constitute an orthonormal set of functions. By construction, we have, e.g., when $n = 4$,

$$\boldsymbol{\alpha}^{(1)}_{\mathbf{f}} \equiv \begin{pmatrix} \langle \bar{f}, \psi^{(1)}_{1,1} \rangle \\ \langle \bar{f}, \psi^{(1)}_{1,2} \rangle \\ \langle \bar{f}, \psi^{(1)}_{2,1} \rangle \\ \langle \bar{f}, \phi^{(1)}_{2,1} \rangle \end{pmatrix} = \begin{pmatrix} \langle \mathbf{f}, \mathbf{w}^{(1)}_{1,1} \rangle \\ \langle \mathbf{f}, \mathbf{w}^{(1)}_{1,2} \rangle \\ \langle \mathbf{f}, \mathbf{w}^{(1)}_{2,1} \rangle \\ \langle \mathbf{f}, \mathbf{v}^{(1)}_{2,1} \rangle \end{pmatrix} \equiv W_1 \mathbf{f},$$

where $W_1$ is by definition the Haar DWT matrix $W$ of equation (7), and the transposes of $\mathbf{w}^{(1)}_{j,k}$ and $\mathbf{v}^{(1)}_{J,1}$ are the rows of $W_1$. Thus, while $\boldsymbol{\alpha}^{(1)}_{\mathbf{f}}$ is not produced by an orthonormal transform of $\bar{f}$, by design it can be interpreted as an orthonormal transform of the vector $\mathbf{f}$.

## 3.2 Asymmetric Haar

As noted in the previous section, isometric Haar wavelets are not orthonormal with respect to the Euclidean inner product (although they are orthonormal with respect to the empirical $\hat{p}_X$ inner product). The basic idea behind the 'asymmetric Haar' approach is to adjust the heights of the isometric Haar wavelets so that the resulting functions $\psi^{(2)}_{j,k}$ and $\phi^{(2)}_{J,1}$ are orthonormal with respect to the Euclidean inner product. Thus, as was true for the isometric Haar wavelets, the widths of the asymmetric Haar wavelets are defined by the observed $x_i$'s; in contrast, the heights of the mother asymmetric Haar wavelets are set to ensure unit norms and integration to zero, while the height of the father wavelet is set to yield unit norm. The appropriate definitions for the asymmetric Haar mother wavelets are

$$\psi_{j,k}^{(2)}(x) = \begin{cases} \psi_{j,k}^{+} \equiv \sqrt{\frac{\delta_{j,k}^{-}}{\delta_{j,k}^{+}(\delta_{j,k}^{+}+\delta_{j,k}^{-})}}, & m_{2^j(k-1)} \leq x < m_{2^j(k-\frac{1}{2})}; \\ \psi_{j,k}^{-} \equiv -\sqrt{\frac{\delta_{j,k}^{+}}{\delta_{j,k}^{-}(\delta_{j,k}^{+}+\delta_{j,k}^{-})}}, & m_{2^j(k-\frac{1}{2})} \leq x < m_{2^j k}; \\ 0, & \text{otherwise}, \end{cases}$$

while the father wavelet $\phi_{J,1}^{(2)}$ is given as in equation (5). An example of these wavelets is plotted in the second column of figure 1. By construction each mother wavelet integrates to zero and has unit norm. It is easy to check that the $\psi_{j,k}^{(2)}$'s and $\phi_{J,1}^{(2)}$ are pairwise orthogonal.

The wavelet series $\boldsymbol{\alpha}_{\mathbf{f}}^{(2)}$ for $\bar{f}$ with respect to the asymmetric Haar wavelets is given by, e.g., when $n = 4$,

$$\boldsymbol{\alpha}_{\mathbf{f}}^{(2)} \equiv \begin{pmatrix} \langle \bar{f}, \psi_{1,1}^{(2)} \rangle \\ \langle \bar{f}, \psi_{1,2}^{(2)} \rangle \\ \langle \bar{f}, \psi_{2,1}^{(2)} \rangle \\ \langle \bar{f}, \phi_{2,1}^{(2)} \rangle \end{pmatrix} \equiv W_2 \mathbf{f},$$

where $W_2$ is an $n \times n$ matrix that can be defined as follows. Let $\delta_i \equiv m_i - m_{i-1}$ be the distances between midpoints. We then have

$$\langle \bar{f}, \psi_{j,k}^{(2)} \rangle = \int_{\frac{1}{2}}^{n+\frac{1}{2}} \psi_{j,k}^{(2)}(x) \bar{f}(x)\, dx = \sum_{i=1}^{n} \psi_{j,k}^{(2)}(x_i) f(x_i) \delta_i,$$

with a similar expression for $\langle \bar{f}, \phi_{J,1}^{(2)} \rangle$. We can thus write $W_2 = VD$, where $V$ is an $n \times n$ matrix that has the same pattern of zeros as the usual Haar DWT matrix $W$, while $D$ is a diagonal matrix with diagonal elements given by $\delta_1, \ldots, \delta_n$. As an example, for $n = 4$ we have

$$V = \begin{pmatrix} \psi_{1,1}^{+} & \psi_{1,1}^{-} & 0 & 0 \\ 0 & 0 & \psi_{1,2}^{+} & \psi_{1,2}^{-} \\ \psi_{2,1}^{+} & \psi_{2,1}^{+} & \psi_{2,1}^{-} & \psi_{2,1}^{-} \\ \phi_{2,1}^{+} & \phi_{2,1}^{+} & \phi_{2,1}^{+} & \phi_{2,1}^{+} \end{pmatrix} \text{ and } D = \begin{pmatrix} \delta_1 & 0 & 0 & 0 \\ 0 & \delta_2 & 0 & 0 \\ 0 & 0 & \delta_3 & 0 \\ 0 & 0 & 0 & \delta_4 \end{pmatrix}.$$

Note that, when the $x_i$'s are equally spaced, $W_2$ for this example is equal to the $W$ of equation (7). Define $\widetilde{W}_2 = VD^{1/2}$, where $D^{1/2}$ is the square root matrix for $D$. A straightforward argument shows that $\widetilde{W}_2^T \widetilde{W}_2 = I$. The orthonormality of $\widetilde{W}_2^T$ implies that $W_2^T W_2 = D$ and hence that $W_2^{-1} = V^T$. Thus, while the columns of $W_2$ are pairwise orthogonal, they do not have unit norm; nonetheless, the $W_2$ transform can be readily inverted so that we can recover $\mathbf{f}$ from $\boldsymbol{\alpha}_{\mathbf{f}}^{(2)}$ using $\mathbf{f} = V^T \boldsymbol{\alpha}_{\mathbf{f}}^{(2)}$.

## 3.3 Sampling on a Grid

The third simple scheme we consider is to sample the interpolated function $\bar{f}$ over a set of $n' = 2^{J'} \geq n$ equally spaced points, to use these sampled values to construct a new interpolated function and then to apply the usual Haar CWT to this new function to obtain the vector of $n'$ CWT coefficients $\boldsymbol{\alpha}_{\mathbf{f}}^{(3)}$. To specify the $n'$ interpolated values, let $\Delta \equiv n/n' = 2^{J-J'}$ and $x_i' \equiv \frac{1}{2} + \left(i - \frac{1}{2}\right)\Delta$, $i = 0, \ldots, n'+1$. Define the vector of $n'$ interpolated values as $\mathbf{f}^{(3)} \equiv [\, \bar{f}(x_1') \quad \bar{f}(x_2') \quad \cdots \quad \bar{f}(x_{n'}')\,]^T$. Let $m_i' = \frac{1}{2} + i\Delta$, $i = 0, \ldots, n'$ represent the midpoints of the $x_i'$'s (note that $m_0' = \frac{1}{2}$ and $m_{n'}' = n + \frac{1}{2}$). Define the new interpolated function as

$$\bar{f}^{(3)}(x) = \begin{cases} \bar{f}(x_i'), & m_{i-1}' \leq x < m_i',\ i = 1, \ldots, n'; \\ 0, & \text{otherwise.} \end{cases}$$

Define $n' - 1$ Haar mother wavelets $\psi_{j,k}^{(3)}$ as in equation (4), with the distinction that now $j = J_0, J_0 + 1, \ldots, J$ with $J_0 \equiv 1 + J - J'$ (note that $J_0 \leq 1$). The father wavelet $\phi_{J,1}^{(3)}$ is given by (5). Let $W_{n'}$ be the $n' \times n'$ Haar DWT matrix. The CWT coefficients $\boldsymbol{\alpha}_{\mathbf{f}}^{(3)}$ can be expressed as

$$\boldsymbol{\alpha}_{\mathbf{f}}^{(3)} = \sqrt{\Delta} W_{n'} \mathbf{f}^{(3)} = W_3 \mathbf{f} \ \text{ with } \ W_3 \equiv \sqrt{\Delta} W_{n'} G,$$

where $G$ is an $n' \times n$ matrix expressing the interpolation. In the example shown in the third column of figure 1 (for which $n = 4$ and $n' = 8$), the $G$ matrix is given by

$$G^T = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

Note that the number of ones in the $i$th column of $G$ indicates the number of times $f(x_i)$ is used in the interpolation scheme. We can define a pseudo inverse for this $G$ as follows:

$$G^{\#} \equiv \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Note that, if the $i$th column of $G$ contains $n_i$ ones with $n_i \geq 1$, then the $i$th row of $G^{\#}$ contains $1/n_i$ in $n_i$ places. If each $f(x_i)$ is used at least once so that $G$ has rank $n$, then $G^{\#}G = I_n$, where $I_k$ refers to the $k \times k$ identity matrix. Since $W_{n'}^T W_{n'} = I_{n'}$, we can recover $\mathbf{f}$ from $\boldsymbol{\alpha}_{\mathbf{f}}^{(3)}$ in the full rank case via $\mathbf{f} = W_3^{\#}\boldsymbol{\alpha}_{\mathbf{f}}^{(3)}$, where $W_3^{\#} \equiv G^{\#}W_{n'}^T/\sqrt{\Delta}$.

Although we can always ensure that $G$ has rank $n$ by making $n'$ sufficiently large, a practical problem is that $n'$ so chosen is driven by the smallest spacings between the original $x_i$'s and hence can be prohibitively large. When this is a concern, a practical – but suboptimal – procedure is to pick $n'$ so that the resulting $\Delta$ is less than or equal to, say, the lower $p\%$ quantile $q_p$ of the observed spacings $x_i - x_{i-1}$'s, (this yields $J' = \lceil \log_2 \left(\frac{n-1}{q_p} + 1\right) \rceil$). We used $p = 20$ in the Monte Carlo simulations reported in section 6. The matrix $G$ typically now has a rank less than $n$ because certain columns can have all zero elements (with $p = 20$ we found the rank of $G$ to be about 90% to 95% of $n$ in the Monte Carlo simulations, which means that 5% to 10% of the observations were being effectively discarded). By defining the corresponding rows of $G^{\#}$ to be zero, we will have $G^{\#}G = K_n$, where $K_n$ is an $n \times n$ diagonal matrix, all of whose diagonal elements are either 1 or 0 (note that, if the $i$th diagonal element of $K_n$ is zero, then $f(x_i)$ is not used at all in the interpolation scheme). When $G^{\#}G \neq I_n$, we can only recover portions of the vector $\mathbf{f}$ perfectly since $W_3^{\#}\boldsymbol{\alpha}_{\mathbf{f}}^{(3)} = K_n\mathbf{f}$. Elements of $K_n\mathbf{f}$ that are zero due to a zero diagonal element in $K_n$ can be filled in by setting them equal to the value of their nearest neighbor, thus yielding an approximate reconstruction of $\mathbf{f}$ given by

$$\mathbf{f}^{(3)} \equiv M_n W_3^{\#}\boldsymbol{\alpha}_{\mathbf{f}}^{(3)},$$

where $M_n$ is an $n \times n$ matrix whose diagonal elements are identical to those of $K_n$ and whose off-diagonal elements are defined by the nearest neighbors. For example, if the $i$th diagonal element of $K_n$ is zero while elements $i-1$ and $i+1$ are nonzero, and if $x_i - x_{i-1} < x_{i+1} - x_i$, then the $i$th row of $M_n$ will be unity at element $i-1$ and zero elsewhere.

## 3.4 Exact Integration

The fourth simple scheme we consider is to analyze $\bar{f}$ using the Haar CWT defined by equations (4) and (5) with $j = J_0, J_0+1, \ldots, J$, where $J_0 \leq 1$. This yields the Haar CWT coefficients $\boldsymbol{\alpha}_{\mathbf{f}}^{(4)}$, each of which is obtained by an inner product of $\bar{f}$ with $\psi_{j,k}$ or $\phi_{J,1}$ (see equation (6)). Because these functions are all piecewise constant, we can express each coefficient in $\boldsymbol{\alpha}_{\mathbf{f}}^{(4)}$ as a linear combination of elements in $\mathbf{f}$, and hence we can write $\boldsymbol{\alpha}_{\mathbf{f}}^{(4)} = W_4\mathbf{f}$, where $W_4$ is a $n' \times n$ matrix with $n' \equiv 2^{J'}$ and $J' \equiv J - J_0 + 1$.

To recover $\mathbf{f}$ using the elements of $\boldsymbol{\alpha}_{\mathbf{f}}^{(4)}$, we form $\mathbf{f}^{(4)}$, whose $i$th element is by definition

$$\bar{f}^{(4)}(x_i) \equiv \sum_{j=J_0}^{J} \sum_{k=1}^{n/2^j} \langle \bar{f}, \psi_{j,k}\rangle \psi_{j,k}(x_i) + \langle \bar{f}, \phi_{J,1}\rangle \phi_{J,1}(x_i).$$

An efficient way to obtain $\mathbf{f}^{(4)}$ is to first compute $\mathbf{g} \equiv W_{n'}^T \boldsymbol{\alpha}_{\mathbf{f}}^{(4)}/\sqrt{\Delta}$ (as before, $W_n'$ is the $n' \times n'$ Haar DWT matrix); to associate the $j$th value $g_j$ in $\mathbf{g}$ with the $j$th interval $\left[\frac{1}{2} + (j-1)\frac{n}{n'}, \frac{1}{2} + j\frac{n}{n'}\right)$, $j = 1, \ldots, n$; and then to set the $i$th value of $\mathbf{f}^{(4)}$ equal to the $g_j$ such that $x_i$ is contained in the $j$th interval. (An alternative reconstruction procedure is to form the pseudo-inverse of $W_4$, but this becomes computationally impractical for large $n'$. It should be noted this alternative procedure does *not* in general yield $\mathbf{f}^{(4)}$.)

As was true for sampling on a grid, we can insure that $\mathbf{f}^{(4)} = \mathbf{f}$ by making $J'$ sufficiently large, but again the practical problem is that the resulting $n'$ can be prohibitively large. The choice of $J'$ is obviously quite important. If it is set too small, we will not get a reasonable approximation to $\bar{f}$ when projected upon the basis functions $\psi_{j,k}$ and $\phi_{J,1}$. As a practical procedure, we set $J'$ using the same heuristic as we did for sampling on a grid.

# 4 WaveShrink for Equally Spaced Data

In this section we review the key steps in the WaveShrink algorithm for denoising equally spaced data as formulated in Donoho and Johnstone (1994). Given $n$ samples generated according to equation (1) with $x_i = i$, we can express our model in vector notation as $\mathbf{Y} = \mathbf{f} + \boldsymbol{\epsilon}$, where $\mathbf{Y} \equiv [\,Y_1 \quad Y_2 \quad \cdots \quad Y_n\,]^T$ and $\boldsymbol{\epsilon} \equiv [\,\epsilon_1 \quad \epsilon_2 \quad \cdots \quad \epsilon_n\,]^T$. The WaveShrink algorithm estimates $\mathbf{f}$ using the following three steps.

1. *Computation of the Wavelet Coefficients.* As before, let $W$ be the $n \times n$ DWT matrix. Using this matrix, we compute

$$\boldsymbol{\alpha}_{\mathbf{Y}} \equiv W\mathbf{Y} = W\mathbf{f} + W\boldsymbol{\epsilon} \equiv \boldsymbol{\alpha}_{\mathbf{f}} + \boldsymbol{\alpha}_{\boldsymbol{\epsilon}}.$$

   Because the DWT is an orthonormal transform and because the $\epsilon_i$'s are independent and identically distributed normal random variables (iid rv's), the vector $\boldsymbol{\alpha}_{\boldsymbol{\epsilon}}$ has exactly the same distribution as $\boldsymbol{\epsilon}$. The wavelet transform thus converts a 'function plus noise' model into another such model, and the statistical properties of the noise are identical in both models.

2. *Shrinkage of Wavelet Coefficients.* The underlying heuristic of WaveShrink is that the wavelet transforms of the sampled function and the noise are quantitatively different. On the one hand, WaveShrink postulates that $\mathbf{f}$ is well represented in the wavelet domain by a combination of all coefficients associated with the large scales indexed by $j = J_s + 1, \ldots, J$ together with a few important large coefficients associated with scales $j = 1, \ldots, J_s$. On the other hand, the elements of $\boldsymbol{\alpha_\epsilon}$ are iid rv's, and hence the wavelet-domain representation for the noise consists of coefficients of roughly the same size. The WaveShrink algorithm is thus to leave untouched all coefficients in $\boldsymbol{\alpha_{\mathbf{Y}}}$ corresponding to scales $j = J_s + 1, \ldots, J$ and to shrink toward zero the $n_s \equiv n - n/2^{J_s}$ coefficients at scales $j = 1, \ldots, J_s$. If the few large coefficients in $\boldsymbol{\alpha_{\mathbf{f}}}$ dominate all the coefficients in $\boldsymbol{\alpha_\epsilon}$ and if each coefficient is shrunk appropriately by an amount that depends on its magnitude, WaveShrink can suppress noise coefficients while retaining the large important coefficients in $\boldsymbol{\alpha_{\mathbf{f}}}$ with little alteration. A number of shrinkage functions have been proposed and studied in the literature (see, e.g., Bruce and Gao (1996)). For simplicity and for comparison with results in the equally spaced case, we use the soft shrinkage function due to Donoho and Johnstone (1994):

$$\delta_\lambda^S(x) = \operatorname{sgn}(x)(|x| - \lambda)_+, \tag{10}$$

where $\lambda$ is a (to be determined) positive threshold level; $\operatorname{sgn}(x) = 1$ if $x > 1$ and $= -1$ otherwise; and $(x)_+ = x$ if $x > 0$ and $= 0$ otherwise.

3. *Estimation of the Function.* Let $\alpha_{\mathbf{Y},i}$ be one of the wavelet coefficients in $\boldsymbol{\alpha}_{\mathbf{Y}}$ that is to be shrunk. The shrunk coefficient is then given by

$$\hat{\alpha}_{\mathbf{Y},i} \equiv \sigma\delta_{\lambda}^{S}(\alpha_{\mathbf{Y},i}/\sigma). \tag{11}$$

On the other hand, if $\alpha_{\mathbf{Y},i}$ is one of the coefficients that is to be left alone, then $\hat{\alpha}_{\mathbf{Y},i} \equiv \alpha_{\mathbf{Y},i}$. Let $\hat{\boldsymbol{\alpha}}_{\mathbf{Y}}$ be a vector containing the $\hat{\alpha}_{\mathbf{Y},i}$'s. Because $W$ is an orthonormal transform, its inverse is its transpose, so we can estimate $\mathbf{f}$ via $\hat{\mathbf{f}} \equiv W^T\hat{\boldsymbol{\alpha}}_{\mathbf{Y}}$.

The above formulation for WaveShrink requires that we know the noise variance $\sigma^2$ and that we set $J_s$ and $\lambda$. Estimation of $\sigma^2$ is not difficult: because the heuristic behind WaveShrink says that these coefficients mainly depend on $\boldsymbol{\epsilon}$ with the exception of a few large values ('outliers') due to $\mathbf{f}$, we can use a robust estimate of the variance of the $j = 1$ scale wavelet coefficients (Donoho et al. (1995)). To set $J_s$, we follow Bruce and Gao (1996), who found via empirical experiments that leaving the 16 highest scale wavelet coefficients untouched is generally a good choice; i.e., $J_s = J - 4$, and $n_s = n - 16$ (see, however, our discussion of the constant function in section 8).

The most difficult task in specifying WaveShrink is to determine the threshold level $\lambda$. Since the goal is to estimate $\mathbf{f}$ such that the risk measure of equation (2) is small, Donoho and Johnstone (1995) propose estimating the risk measure as a function of $\lambda$ and then picking $\lambda$ such that the estimated risk is minimized. The proposed estimator is Stein's unbiased risk estimator (SURE), which is defined as follows. Let $\alpha_{\mathbf{Y},i}$, $i = 1, \ldots, n_s$ represent the subset of the wavelet coefficients to be shrunk. For a given $\lambda$, the risk estimator is given by

$$\mathrm{SURE}(\lambda) = \sigma^2\left(n - 2\sum_{i=1}^{n_s} I\left\{|\alpha_{\mathbf{Y},i}| \le \lambda\sigma\right\} + \sum_{i=1}^{n_s}\min\left\{\alpha_{\mathbf{Y},i}^2/\sigma^2, \lambda^2\right\}\right), \tag{12}$$

where $I\left\{|\alpha_{\mathbf{Y},i}| \le \lambda\sigma\right\} = 1$ if $|\alpha_{\mathbf{Y},i}| \le \lambda\sigma$ and $= 0$ otherwise. Finding the minimizing $\lambda$ is easy because the minimum of $\mathrm{SURE}(\lambda)$ can only occur over a set of $n_s$ discrete values.

It should be noted that WaveShrink is computationally very efficient because the DWT and its inverse can be performed using fast 'pyramid' algorithms (Mallat (1989)). These algorithms require $O(n)$ floating point operations (by comparison, the well-known fast Fourier transform algorithm requires $O(n \log_2 n)$)). In addition, because WaveShrink leaves unaltered all wavelet coefficients for the $J - J_s$ largest scales, it is only necessary to use $J_s$ repetitions of the pyramid algorithm (in fact, this allows us to relax the requirement that $n$ be a power of 2: we actually only need that $n$ be divisible by $2^{J_s}$).

# 5 WaveShrink for Unequally Spaced Data

In order to estimate $\mathbf{f}$ from noisy data $\mathbf{Y}$ using the $l$th of the four Haar-like transforms described in section 3, we transform the data to obtain the empirical wavelet coefficients $\boldsymbol{\alpha}_{\mathbf{Y}}^{(l)} \equiv W_l \mathbf{Y}$, shrink the wavelet coefficients to obtain $\hat{\boldsymbol{\alpha}}_{\mathbf{Y}}^{(l)}$, and then produce the estimate $\hat{\mathbf{f}}^{(l)}$ from $\hat{\boldsymbol{\alpha}}_{\mathbf{Y}}^{(l)}$. While the material in section 3.$l$ completely describes how to compute $\boldsymbol{\alpha}_{\mathbf{Y}}^{(l)}$ and to produce the estimate $\hat{\mathbf{f}}^{(l)}$ once $\hat{\boldsymbol{\alpha}}_{\mathbf{Y}}^{(l)}$ is known, we need to adapt the shrinkage step to handle unequally sampled data.

Although the DWT of equally spaced data leaves the noise structure unaltered, this is not generally true for three of the methods we have discussed. Under the model $\mathbf{Y} = \mathbf{f} + \boldsymbol{\epsilon}$, we have

$$\boldsymbol{\alpha}_{\mathbf{Y}}^{(l)} = \boldsymbol{\alpha}_{\mathbf{f}}^{(l)} + \boldsymbol{\alpha}_{\boldsymbol{\epsilon}}^{(l)} \text{ with } \boldsymbol{\alpha}_{\boldsymbol{\epsilon}}^{(l)} \equiv W_l \boldsymbol{\epsilon}.$$

Since the covariance matrix of $\boldsymbol{\epsilon}$ is by assumption $\sigma^2 I$, the covariance matrix $\Sigma^{(l)}$ of $\boldsymbol{\alpha}_{\mathbf{Y}}^{(l)}$ is $\Sigma^{(l)} = \sigma^2 W_l W_l^T$. In particular we have

$$\Sigma^{(1)} = \sigma^2 I, \ \Sigma^{(2)} = \sigma^2 V D D V^T, \ \Sigma^{(3)} = \sigma^2 \Delta W_{n'} G G^T W_{n'}^T \text{ and } \Sigma^{(4)} = \sigma^2 W_4 W_4^T.$$

Except for $\Sigma^{(1)}$, the diagonal elements for these covariance matrices are in general not equal, and the matrices in general have nonzero off-diagonal elements. Thus the elements of $\boldsymbol{\alpha}_{\boldsymbol{\epsilon}}^{(l)}$, $l = 2$, 3 or 4, in general have heterogeneous variance and are pairwise correlated. To correct for the heterogeneous variance, we use 'diagonal' shrinkage so that equation (11) becomes

$$\hat{\alpha}_{\mathbf{Y},i}^{(l)} \equiv \sigma_i \delta_\lambda^S(\alpha_{\mathbf{Y},i}^{(l)}/\sigma_i),$$

where $\sigma_i^2$ is the $i$th diagonal element of $\Sigma^{(l)}$; likewise, the risk estimator given in equation (12) is adjusted by replacing both occurrences of $\alpha_{\mathbf{Y},i}$ by $\alpha_{\mathbf{Y},i}^{(l)}\sigma/\sigma_i$. Note that, because diagonal shrinkage requires knowledge only of the diagonal elements of $\Sigma^{(l)}$, we can use efficient algorithms to compute these elements rather than forming all of $\Sigma^{(l)}$ using inefficient matrix multiplications (these algorithms are implemented in the computer code referenced in section 9).

Although diagonal shrinkage forces homogeneous variances for $\boldsymbol{\alpha}_{\mathbf{Y}}^{(l)}$, its elements are in general still pairwise correlated. For ideal shrinkage, we would need to account for these correlations; however, under certain conditions, Johnstone and Silverman (1997) show that diagonal shrinkage is asymptotically equivalent to ideal shrinkage. For simplicity we confine ourselves in this paper to diagonal shrinkage (this simplification is valid as long as the covariance matrix is diagonally dominant – this is always true for isometric wavelets, but might be violated for the other techniques if the sample distribution of the differences $x_{i+1}-x_i$ is highly skewed).

WaveShrink also requires knowledge of $\sigma^2$. As noted previously, it is easy to estimate $\sigma^2$ in the equally spaced case using a robust estimator such as the median absolute deviation (MAD) scale estimator. For unequally spaced data, we can estimate $\sigma^2$ using the same robust estimator in the case of the isometric wavelet scheme, but the other three methods require more complicated estimators. In the example discussed in section 7, we used a MAD estimate based on the finest scale wavelet coefficients rescaled by the diagonal of the covariance matrix, but it is beyond the scope of this article to ascertain what is the best way of estimating $\sigma^2$ for each method. To focus our study on the methods themselves rather than on more complicated interplay between the methods and various estimators of $\sigma^2$, we will evaluate the four methods in the next section under the assumption that $\sigma^2$ is known *a priori*.

# 6 Monte Carlo Study

Here we report on Monte Carlo experiments conducted to compare the performance of WaveShrink based upon the four techniques described above. We used five test functions: the zero function $f(x) = 0$ and functions proportional to the four plotted in figure 2, which are called the blocks, bumps, heavisine and Doppler functions. The latter four functions were used in Donoho and Johnstone (1994) for testing WaveShrink on equally spaced data and were chosen to be caricatures of ones arising in imaging, spectroscopy and other scientific applications. These functions are defined precisely in table 1 of Donoho and Johnstone (1994) over the interval $[0, 1]$, which we have mapped to the interval $[1, n]$ to match the convention adopted in this paper; additionally, each of these nonzero test functions was normalized such that its 'standard deviation' is equal to 5:

$$\frac{1}{n-1}\int_1^n \left(f(x) - \bar{f}\right)^2 dx = 25, \text{ where } \bar{f} \equiv \frac{1}{n-1}\int_1^n f(x)\, dx.$$

We consider four different sample sizes: $n = 64$, 128, 256 and 512.

For all except the Doppler function, we set the $x_i$'s by choosing $n$ samples from a standard normal distribution and then rescaling and relocating their order statistics such that the first and last values were 1 and $n$; for the Doppler function, the order statistics of the absolute values of the $n$ samples were used instead (this yields a denser sampling of the first portion of the Doppler signal, which is the region over which it is varying rapidly). The rationale for using the normal distribution was to simulate the sampling scheme that would arise in scatter plots involving normal deviates. For each set of $x_i$'s so chosen, we simulated a noisy function from equation 1, with $\sigma^2 = 1$ (i.e., the standard deviation of the noise is 5 times smaller than the 'standard deviation' of the function). Each noisy function was then used to estimate the true $f$ over the selected $x_i$'s using each of the four WaveShrink techniques. The quality of the estimate was measured by computing the observed risk, namely,

$$\hat{R}(\hat{f}, f) = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{f}(x_i) - f(x_i) \right)^2 . \tag{13}$$

To eliminate effects that might be attributable to a particular choice of the $x_i$'s, we reselected the $x_i$'s for each function $n_r = 16384/n$ times (this makes the total number of generated noisy functions the same for each $n$). We repeated the above to obtain $n_r$ observed risks for each function and each WaveShrink technique. We then averaged all $n_r$ observed risks for each function/technique combination. (The risk function that we are actually estimating thus differs from the one presented in equation 2: the $x_i$'s are replaced by random variables $X_i$, and the expectation is over the joint distribution of the $X_i$'s and $Y_i$'s.)

As noted above, we used SURE to select the threshold $\lambda$ for each noisy function. We assumed knowledge of the noise variance $\sigma^2$ in computing SURE, so the resulting observed risk can be called *oracle* based (Donoho and Johnstone (1994)). In order to assess how well SURE picks an appropriate $\lambda$, we also determined the $\lambda$ such that the right-hand side of equation 13 is minimized as a function of $\lambda$. We denote this second procedure as a *super oracle*.

Table 1 lists the oracle-based average observed risks along with estimated standard deviations for these averages. For each sample size $n$ and each test function, we have used a bold font to indicate the smallest average observed risk (in several cases two risks are so indicated because they agree to the two decimal points shown in the table). With the exception of one case ($n = 128$ and heavisine, for which exact integration tied with sampling on a grid), either isometric wavelets or sampling on a grid had the smallest average observed risk. The general pattern is for isometric wavelets to be the best technique for large $n$, while sampling on a grid is better for small $n$. If we examine the difference between observed risks, there is little separation between isometric wavelets and sampling on grid: in cases where the former has a larger risk than the latter, the difference in observed risks is always less than 16%. Moreover, if we take sampling variations into account, there are very few cases where we can claim that the difference between average observed risks is statistically significant. We can thus conclude the isometric wavelet scheme works at least as well as the other three methods and thus is the preferred technique because of its inherent computational simplicity.

When the super oracle was used, we found that the observed risk did not decrease substantially. For the blocks, bumps, heavisine and Doppler functions, the improvement that was gained by using the super oracle was 5% as averaged across the 64 combinations of sample size and method displayed in Table 1 (the improvements were all less than or equal to 12%). For the constant function, the improvements were somewhat larger, ranging between 8% and 18% with an average of 12% over the 16 sample size/method combinations in Table 1, a result that is consistent with the discussion in Donoho and Johnstone (1995) concerning the performance of SURE with extremely sparse signals. We can conclude that the SURE procedure works well in picking out an appropriate $\lambda$ for thresholding.

# 7 Example: An Astronomical Time Series

As an example of wavelet shrinkage of an actual unequally sampled time series, we consider the problem of estimating the light curve for the variable star RU Andromeda (this time series was obtained from the American Association of Variable Star Observers (AAVSO) International Database, which is maintained by J. A. Mattei and is accessible on the World Wide Web at `www.aavso.org`). The observed magnitude values for this star are indicated in Figure 3 by small dots, which range in time from Julian Day 2,449,004 to 2,450,352 (January 1993 to mid-1996). The magnitudes of this star are measured at irregularly spaced times due to blockage of the star by sunlight, weather conditions and availability of telescope time. There were 295 observations in all, three of which were reported as upper limits on the star's magnitude and hence were eliminated since their error properties are quite different from the remaining observations. Out of the 292 remaining observations we selected 256 values at random to conform to the power of two assumption made throughout this paper (we did the random selection just once, but a slight improvement to this procedure would be to make many such selections and then to average the resulting estimated light curves).

The four different estimated light curves are indicated on Figure 3 by connected lines. Qualitatively the four estimated light curves are quite similar, although the one given by asymmetric Haar is markedly noisier in appearance. The estimated light curves generally track the overall light variations quite nicely.

# 8 Conclusions and Discussion

We have demonstrated that the simple isometric wavelet scheme works just as well as three other schemes for constructing a Haar-like wavelet shrinkage estimator for unequally sampled data. We have provided some theoretical justification as to why the isometric wavelet scheme works in terms of a nonstandard – but intuitively appealing – metric, a result that is of interest even for smoothing schemes that are not based on wavelets. The isometric wavelet idea has additional appeal in that it is trivial to extend it to wavelets other than the Haar, which will be considered in future research.

Isometric wavelets can be extended to higher dimension of the predictor space. Of particular interest is the 2-D denoising problem. Images are 2-D signals on a grid. Because the data are equally spaced at the pixel locations, WaveShrink can be used for the image denoising problem and enjoys the same properties as in the 1-D case. Among the four techniques proposed in this paper to generalize WaveShrink to the unequally spaced setting, the isometric wavelet technique can easily be extended to the 2-D problem. In the 1-D case, the segmentation of the predictor space (the line) from mid-point to mid-point leads to both the definition of the compact support of the wavelets, and the empirical marginal density $\hat{p}_X(x)$. Similarly, the planar tessellation (Okabe et al. (1992)) of the predictor space in Voronoi diagrams offers the same properties: the compact support of the wavelets is the union of neighboring Voronoi polygons, and the area of the Voronoi polygon reflects the density of the explanatory data.

It should be noted that, with $J_s$ fixed *a priori*, wavelet shrinkage estimation of a constant function is *not* competitive with the best parametric procedure (i.e., the sample mean), but it can be made so if $J_s$ is set so that there is but a single scaling coefficient (proportional to the sample mean) and if threshold levels are set such that wavelet shrinkage sets all the wavelet coefficients to be zero. More research is needed on data-based setting of $J_s$ and the threshold levels to ascertain whether wavelet shrinkage can compare favorably to the best parametric procedure in this case and similar scenarios.

# 9 Software Availability

The figures and tables in this paper are reproducible through software that can be obtained by anonymous ftp to `ftp.statsci.com` in the directory `pub/WAVELETS/unequal` (this software makes use of the S+WAVELETS toolkit in S-PLUS).

## 10 Acknowledgments

This work was partially supported by ONR Contract N00014–93–C–0106 and by NSF Grants DMI–94–61370 and DMI–96–28595. The authors would like to thank the editor and an anonymous referee for helpful comments that improved this paper.

| | isometric | asymmetric | sampling | exact |
|---|---|---|---|---|
| **blocks** | | | | |
| $n = 64$ | 0.66 ±0.03 | 0.71 ±0.03 | **0.63** ±0.04 | 0.66 ±0.06 |
| $n = 128$ | 0.50 ±0.02 | 0.55 ±0.02 | **0.49** ±0.03 | 0.52 ±0.04 |
| $n = 256$ | **0.38** ±0.02 | 0.45 ±0.02 | **0.38** ±0.02 | 0.39 ±0.01 |
| $n = 512$ | **0.28** ±0.01 | 0.33 ±0.01 | 0.29 ±0.01 | 0.29 ±0.01 |
| **bumps** | | | | |
| $n = 64$ | **0.75** ±0.04 | 0.79 ±0.04 | 0.83 ±0.06 | 0.88 ±0.09 |
| $n = 128$ | 0.66 ±0.02 | 0.71 ±0.02 | **0.65** ±0.03 | 0.68 ±0.03 |
| $n = 256$ | **0.58** ±0.02 | 0.63 ±0.02 | **0.58** ±0.02 | 0.61 ±0.02 |
| $n = 512$ | **0.49** ±0.01 | 0.54 ±0.01 | **0.49** ±0.01 | 0.50 ±0.01 |
| **heavisine** | | | | |
| $n = 64$ | 0.75 ±0.03 | 0.80 ±0.03 | **0.67** ±0.03 | 0.70 ±0.03 |
| $n = 128$ | 0.57 ±0.02 | 0.64 ±0.02 | **0.56** ±0.02 | **0.56** ±0.02 |
| $n = 256$ | **0.43** ±0.01 | 0.50 ±0.01 | 0.46 ±0.02 | 0.46 ±0.02 |
| $n = 512$ | **0.31** ±0.01 | 0.38 ±0.01 | 0.36 ±0.01 | 0.36 ±0.01 |
| **Doppler** | | | | |
| $n = 64$ | 0.93 ±0.03 | 0.94 ±0.03 | **0.88** ±0.04 | 0.95 ±0.04 |
| $n = 128$ | 0.87 ±0.02 | 0.89 ±0.02 | **0.82** ±0.02 | 0.88 ±0.02 |
| $n = 256$ | 0.80 ±0.01 | 0.84 ±0.01 | **0.78** ±0.02 | 0.82 ±0.02 |
| $n = 512$ | **0.68** ±0.01 | 0.74 ±0.01 | 0.73 ±0.02 | 0.71 ±0.01 |
| **pure noise** | | | | |
| $n = 64$ | 0.30 ±0.02 | 0.35 ±0.03 | **0.28** ±0.02 | 0.29 ±0.02 |
| $n = 128$ | **0.16** ±0.01 | 0.19 ±0.01 | **0.16** ±0.01 | 0.16 ±0.01 |
| $n = 256$ | **0.08** ±0.01 | 0.10 ±0.01 | 0.09 ±0.01 | 0.09 ±0.01 |
| $n = 512$ | **0.04** ±0.00 | 0.06 ±0.00 | 0.05 ±0.00 | 0.05 ±0.00 |

Table 1: Oracle-based average observed risk ($n_r = 16384/b$ repetitions)

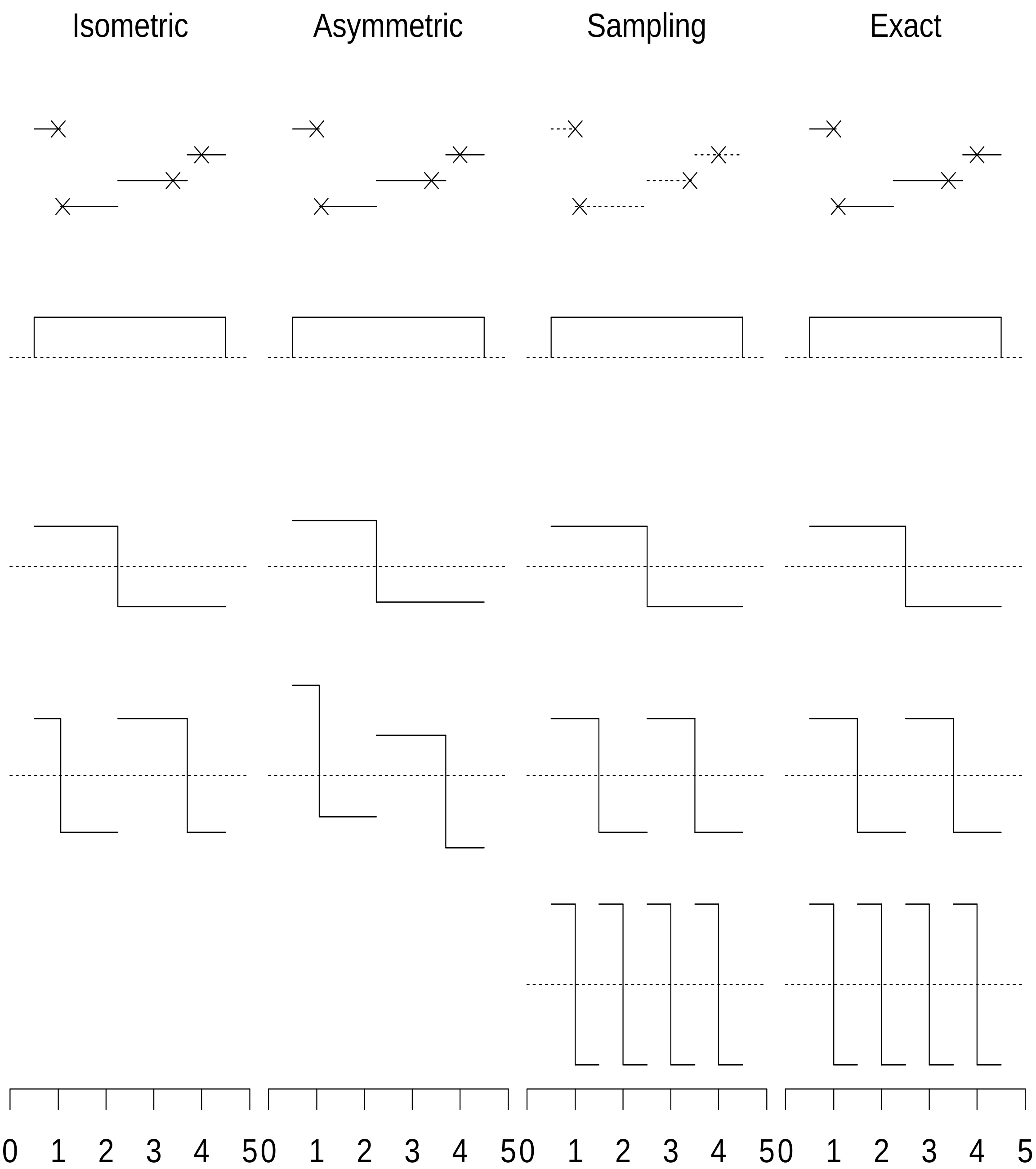


Figure 1: The above figure illustrates the different wavelets used in function space for the four techniques (here $x_1 = 1$, $x_2 = 1.1$, $x_3 = 3.4$ and $x_4 = 4$).

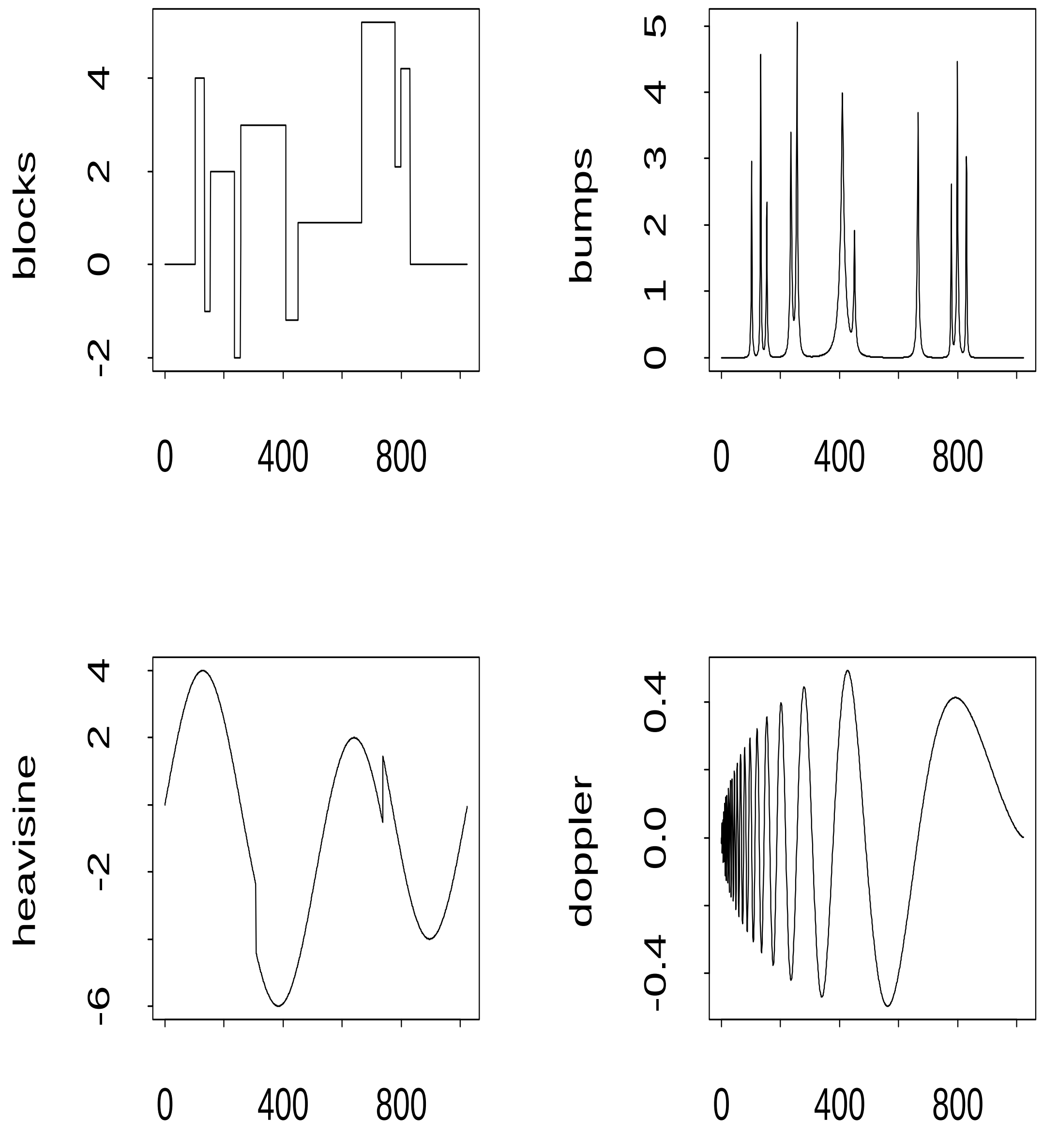


Figure 2: The above figure shows the four test functions used in the Monte Carlo experiments.

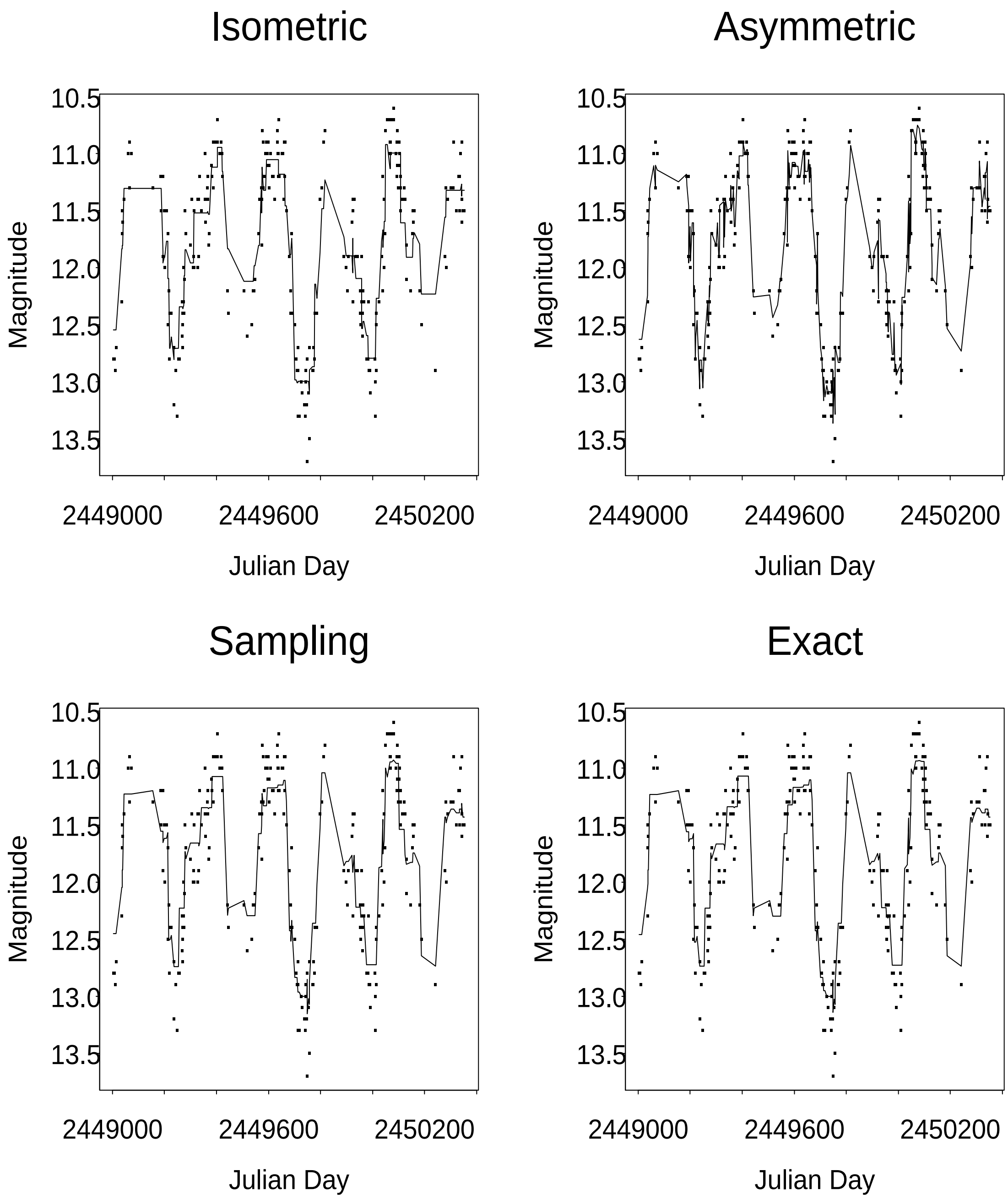


Figure 3: Wavelet shrinkage estimates of light curve for the variable star RU Andromeda (this time series was obtained from the American Association of Variable Star Observers (AAVSO) International Database maintained by J. A. Mattei).

# References


Antoniadis, A., Gregoire, G., and McKeague, I. W. (1994). Wavelet Methods for Curve Estimation. *Journal of the American Statistical Association*, 89(428):1340–1353.

Antoniadis, A., Gregoire, G., and Vial, P. (1997). Random Design Wavelet Curve Smoothing. *Statistics & Probablilty Letters*, 35:225–232.

Bruce, A. G. and Gao, H.-Y. (1996). Understanding WaveShrink: Variance and Bias Estimation. *Biometrika*, 83(4):727–745.

Delyon, B. and Juditsky, A. (1995). Estimating Wavelet Coefficients. In Antoniadis, A. and Oppenheim, G., editors, *Wavelets and Statistics*, pages 151–168. Springer–Verlag, New York. Lecture Notes in Statistics.

Donoho, D., Johnstone, I., Kerkyacharian, G., and Picard, D. (1995). Wavelet Shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, 57:301–369. (with discussion).

Donoho, D. L. and Johnstone, I. M. (1994). Ideal Spatial Adaptation via 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(432):1200–1224.

Foster, G. (1996). Wavelets for Period Analysis of Unequally Sampled Time Series. *Astronomical Journal*, 112(4):1709–1729.

Hall, P. and Turlach, B. A. (1997). Interpolation Methods for Nonlinear Wavelet Regression with Irregularly Spaced Design. *Annals of Statistics*, 25(5):1912–1925.

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(2):319–351.

Kovac, A. and Silverman, B. W. (1998). Extending the Scope of Wavelet Regression Methods by Coefficient-Dependent Thresholding. Technical Report SFB 474, University of Dortmund.

Mallat, S. (1989). A Theory for Multiresolution Signal Decomposition: the Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693.

Okabe, A., Boots, B., and Sugihara, K. (1992). *Spatial Tesselations: Concepts and Applications of Voronoi Diagrams*. John Wiley, New York.

Scargle, J. D. (1997a). Astronomical Time Series Analysis: New Methods for Studying Periodic and Aperiodic Systems. In Maoz, D., Sternberg, A., and Leibowitz, E. M., editors, *Astronomical Time Series*. Kluwer Academic Publishers, Dordrecht.

Scargle, J. D. (1997b). Wavelet Methods in Astronomical Time Series Analysis. In Rao, T. S., Priestley, M. B., and Lessi, O., editors, *Applications of Time Series Analysis in Astronomy and Meteorology*, pages 226–248. Chapman & Hall, London.

Sweldens, W. (1995). The Lifting Scheme: a New Philosophy in Biorthogonal Wavelet Constructions. In *SPIE Proceedings, Signal and Image Processing III*, volume 2569, pages 68–79, San Diego, CA.