Wavelet Methods in Statistics

Piotr Fryzlewicz

Department of Mathematics

University of Bristol

Bristol BS8 1TW

UK

p.z.fryzlewicz@bristol.ac.uk

http://www.maths.bris.ac.uk/~mapzf/

July 18, 2007

1 Motivation

The main statistical application of wavelets is in signal denoising (a.k.a. smoothing, non-parametric function estimation). As a motivating example, consider the noisy signal in Figure 1. Our objective is to try to remove the noise and get as close as possible to revealing the "true" structure of the signal. If you squint, you can probably tell that the signal is composed of at least 5 different pieces. Let us see if some established smoothing methods can tell us more.

The solid line in the top plot of Figure 2 is the result of smoothing the signal with a rectangular kernel whose bandwidth has been optimised to minimise the mean-square error. As we can see the smooth still has a noisy appearance. The rounded mean-square error is 784. If you want to try it yourselves, try the `ksmooth` function in R.

The solid line in the bottom plot of Figure 2 is the result of smoothing the signal using the "taut string" methodology of Davies & Kovac (2001). The reconstruction is good except that two of the "dips" are not detected correctly. The rounded mean-square error is 1051. The function to use is `pmreg` from the R package `ftnonpar`.

The solid line in the top plot of Figure 3 is the result of applying the "adaptive weights" smoothing technique of Polzehl & Spokoiny (2000). The reconstruction is very good but some considerable bias is apparent. The rounded mean-square error is 419. The function I used was `awsuni` from the R package `aws`.

Finally, the solid line in the bottom plot of Figure 3 is a reconstruction which uses nonlinear wavelet shrinkage with Haar wavelets, with a little twist. The reconstruction is probably as good as it can be! The rounded mean square error is 176. We will cover the standard nonlinear Haar shrinkage in this lecture course. We might or might not cover the little twist, depending on our time.

2 Wavelets

Wavelets can be informally described as localised, oscillatory functions designed to have several "attractive" properties not enjoyed by "big waves" — sines and cosines. Since their "invention" in the early eighties (the term "wavelet" first appeared in Morlet *et al.*, 1982), wavelets have received enormous attention both in the mathematical community and in the applied sciences. Several monographs on the mathematical theory of wavelets appeared: for example Daubechies (1992), Meyer (1992), Mallat (1998) and Cohen (2003). Some of the material in this section has been adapted from Vidakovic (1999), an excellent monograph on the statistical applications of wavelets.

Formally, let $\psi_{a,b}(x)$, $a \in \mathbb{R} \setminus \{0\}$, $b \in \mathbb{R}$ be a family of functions being translations and

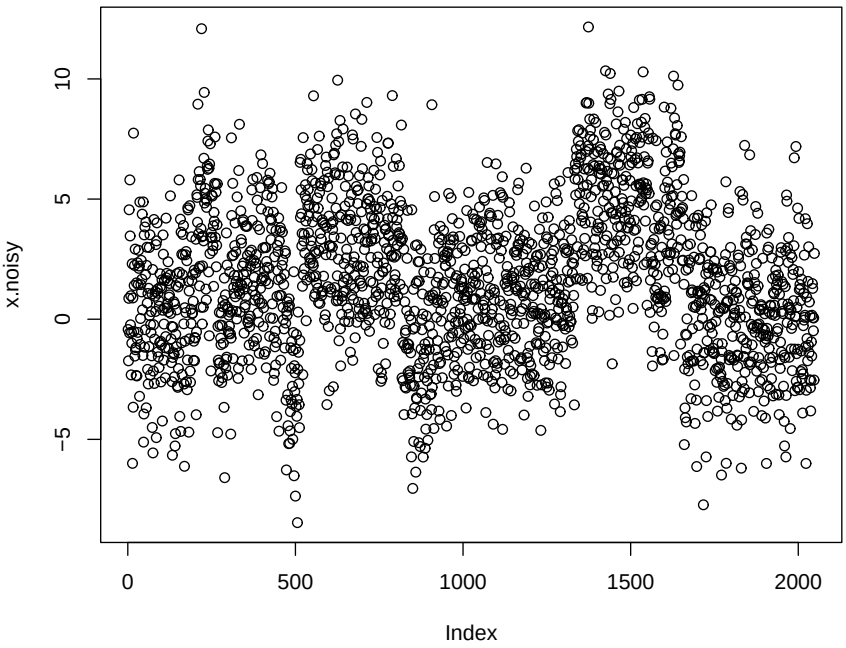

Figure 1: Noisy signal.

dilations of a single function $\psi(x) \in L_2(\mathbb{R})$,

$$\psi_{a,b}(x) = |a|^{-1/2}\psi\left(\frac{x-b}{a}\right).$$

Note that $\|\psi_{a,b}(x)\|_2$ does not depend on (a, b) (typically $\|\psi_{a,b}(x)\|_2 = 1$). The function $\psi(x)$ is called *the wavelet function* or *the mother wavelet*. It is assumed to satisfy the admissibility condition

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} d\omega < \infty, \tag{1}$$

where $\Psi(\omega)$ is the Fourier transform of $\psi(x)$. Condition (1) implies, in particular, that

$$0 = \Psi(0) = \int \psi(x)dx. \tag{2}$$

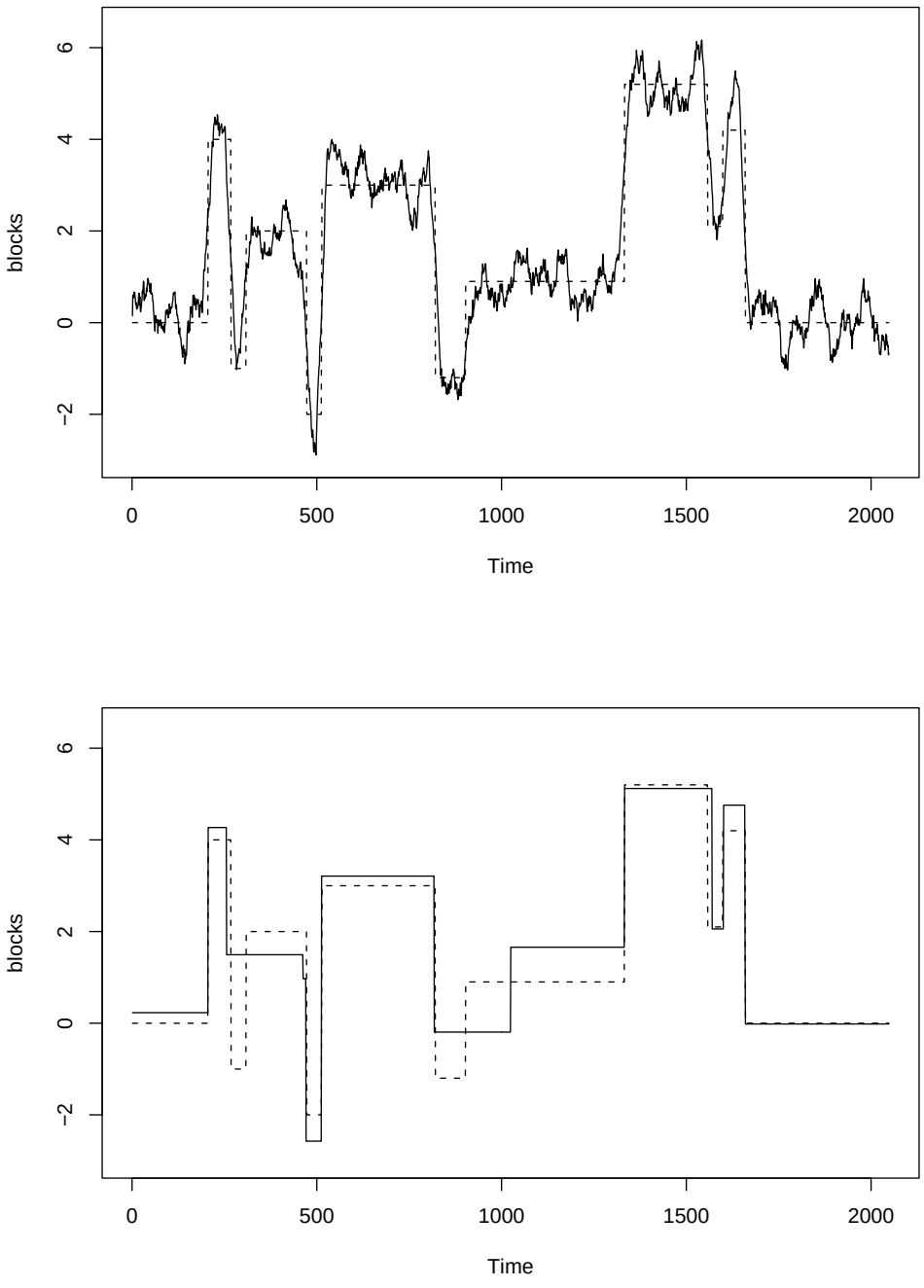

Figure 2: Top: the noisy signal of Figure 1 smoothed with a box kernel with an optimally selected bandwidth (solid) and the true signal (dashed). Bottom: the noisy signal smoothed using the taut string method (solid) and the true signal (dashed).

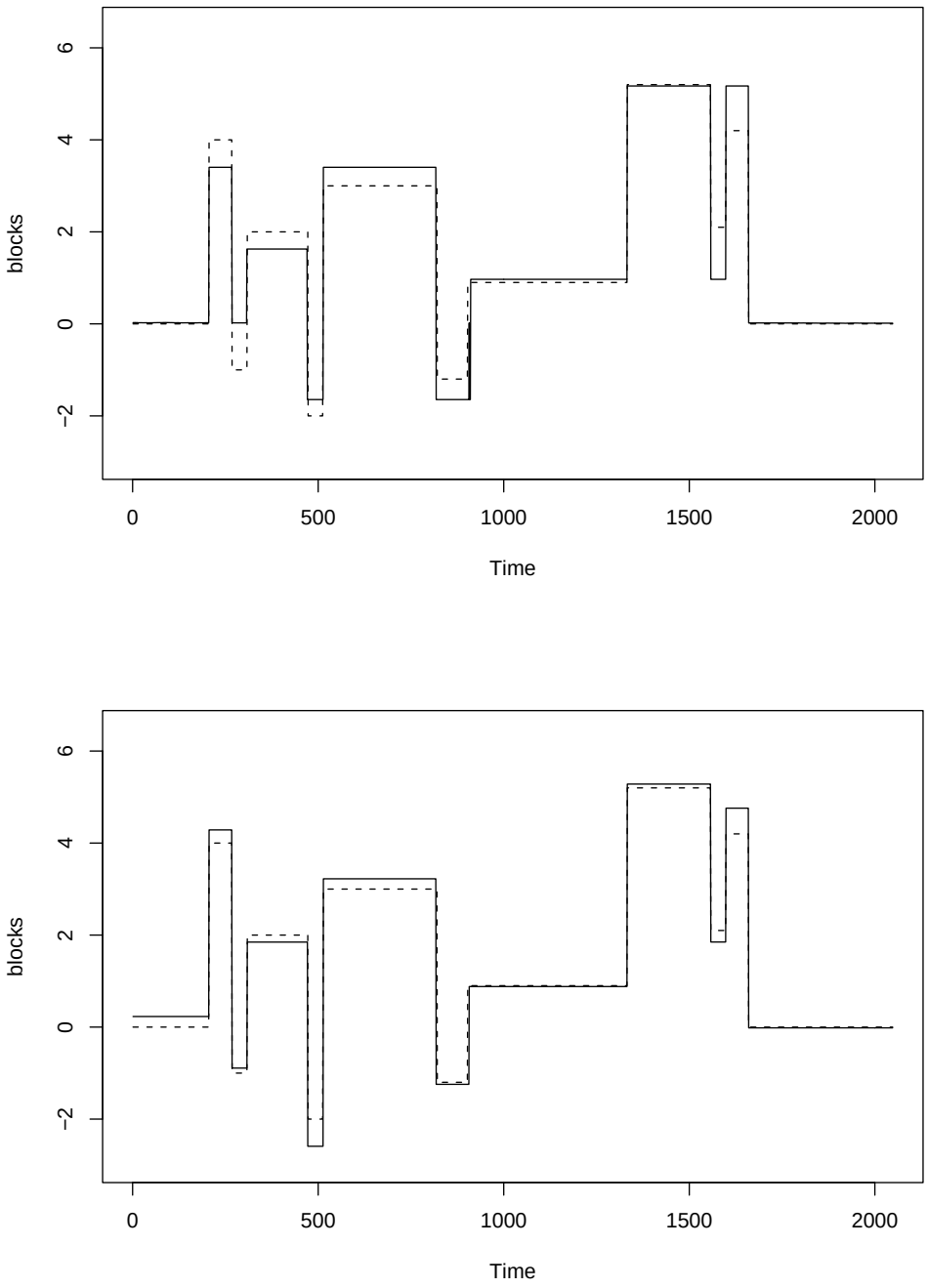

Figure 3: Top: the noisy signal smoothed using "adaptive weights smoothing" (solid) and the true signal (dashed). Bottom: the noisy signal smoothed via Haar wavelets (with a little twist!) and the true signal (dashed).

Condition (1) means that $\psi(x)$ should be localised in frequency. On the other hand, condition (2) means that $\psi(x)$ is localised in time, and also oscillatory. Hence the name "wavelet". The parameter b is the location parameter, and a is the scale parameter. It can be thought of as a reciprocal of frequency.

2.1 Continuous wavelet transform

For any function $f \in L_2$, its continuous wavelet transform is defined as a function of two variables,

$$\mathrm{CWT}_f(a,b) = \langle f, \psi_{a,b} \rangle = \int f(x)\overline{\psi_{a,b}(x)}dx.$$

If condition (1) is satisfied, then the following inverse formula ("resolution of identity") holds

$$f(x) = C_\psi^{-1} \int_{\mathbb{R}^2} \mathrm{CWT}_f(a,b)\psi_{a,b}(x)a^{-2}dadb.$$

The parameter a is often restricted to be positive (as it can be viewed as the "inverse" of frequency). If this is the case, then condition (1) becomes $C_\psi = \int_0^\infty \frac{|\Psi(\omega)|^2}{\omega}d\omega < \infty$, and the resolution of identity becomes

$$f(x) = C_\psi^{-1} \int_{-\infty}^{\infty}\int_0^\infty \mathrm{CWT}_f(a,b)\psi_{a,b}(x)a^{-2}dadb.$$

2.2 Examples of wavelets

2.2.1 Haar wavelets

The best-known example of wavelets are Haar wavelets introduced by Haar (1910) (but not called by this name back then). They are given by

$$\psi^H(x) = \mathbb{I}(0 \le x < 1/2) - \mathbb{I}(1/2 \le x \le 1),$$

which implies

$$\psi^H_{a,b}(x) = a^{-1/2}\{\mathbb{I}(b \leq x < a/2 + b) - \mathbb{I}(a/2 + b \leq x \leq a + b)\}$$

for $a > 0$, $b \in \mathbb{R}$.

We say that the wavelet ψ has n vanishing moments if

$$\int_{-\infty}^{\infty} x^k \psi(x) dx = 0 \quad \text{for} \quad k \in \{0, 1, \ldots, n\}.$$

It is easy to see that ψ^H has 0 vanishing moments. Thus, if f is constant on the interval $[b, a + b]$, then, for Haar wavelets, $\text{CWT}_f(a, b) = 0$.

We will be coming back to this example throughout the course.

2.2.2 Compactly supported Daubechies' wavelets

Daubechies (1992, Chapter 6) identifies the *Extremal Phase* family of wavelet systems: a collection of wavelet systems with compactly supported wavelet functions, possessing different degrees of smoothness and numbers of vanishing moments. This family of systems is indexed by the number of vanishing moments and the Haar system is its zeroth member. A review of this and other families of wavelets, including Daubechies' *Least Asymmetric* family can be found in Vidakovic (1999), Sections 3.4 and 3.5.

Figure 4 shows graphs of Daubechies' Extremal Phase wavelets with $n = 0, 1, 2, 3, 4, 5$ vanishing moments. Note that the higher the number of vanishing moments, the longer the support and the higher the degree of smoothness. Except for Haar wavelets, explicit formulae for other Daubechies' wavelets are not available in the time domain.

Suppose now that over the support of $\psi_{a,b}$, f is a polynomial of degree less than or equal to the number of vanishing moments of $\psi(x)$. Then the corresponding $\text{CWT}_f(a, b) = 0$. We shall be coming back to this "sparsity" property of wavelets.

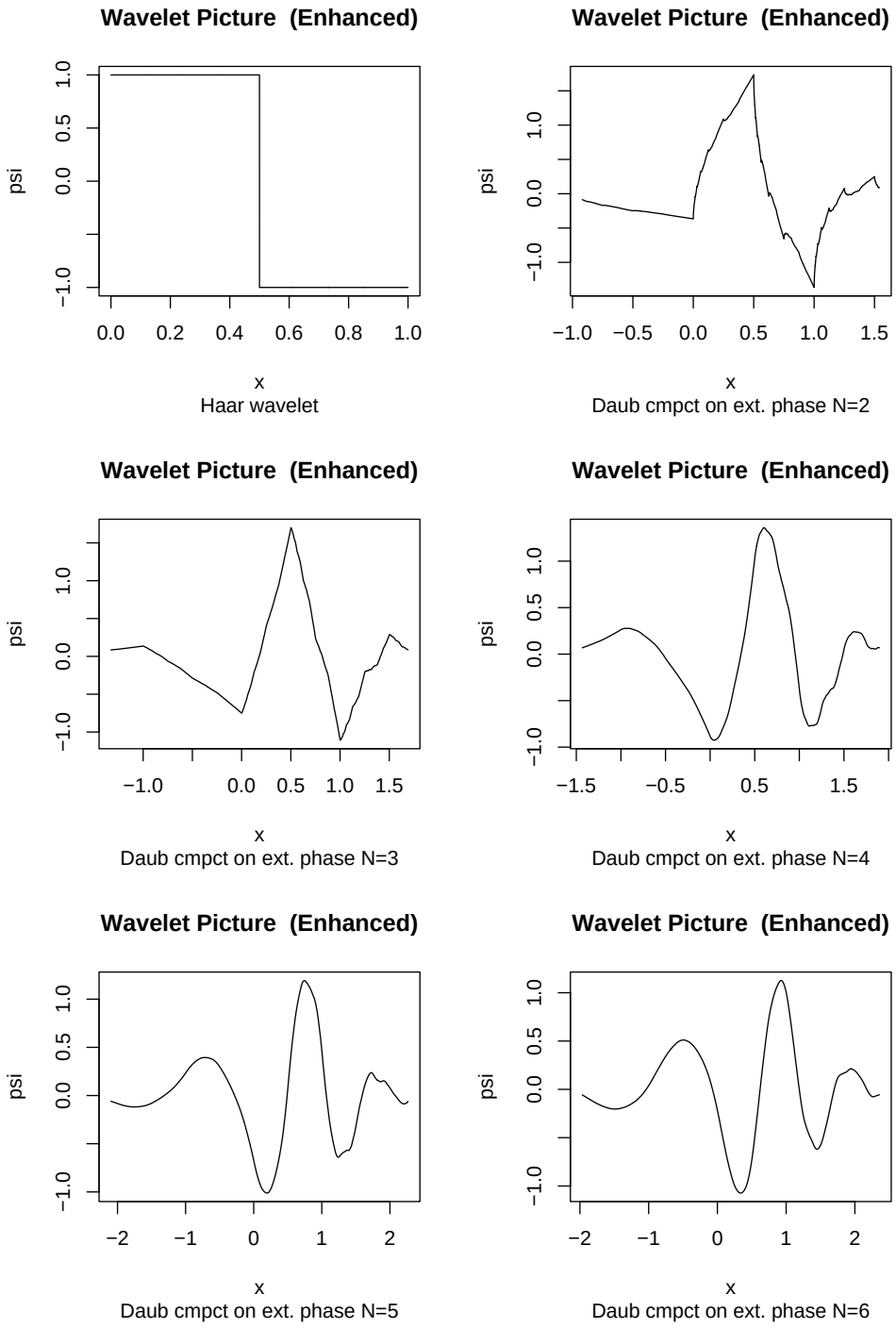

Figure 4: Daubechies' Extremal Phase wavelets with different numbers $n = N - 1$ of vanishing moments.

2.3 Discrete wavelet transform

$\mathrm{CWT}_f(a, b)$ is a function of two real variables so clearly it is a redundant transform. To minimise the transform we might attempt to discretise the values of a and b so that the invertibility of the transform is still retained. Such discretisation cannot be coarser than the so-called *critical sampling*, or otherwise information will be lost. The critical sampling defined by $a = 2^{-j}$, $b = k2^{-j}$, $j, k \in \mathbb{Z}$, will produce a minimal basis for $\mathbb{L}_2$. Moreover, under mild conditions on the wavelet function ψ, the resulting basis

$$\{\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), j, k \in \mathbb{Z}\} \tag{3}$$

will be orthonormal. From now on, we will only be looking at wavelets for which it is the case. All the wavelet functions mentioned so far satisfy this condition.

Other discretisation choices are possible but the above is particularly convenient as it enables a fast implementation of the Discrete Wavelet Transform: a fast decomposition of function or vectors with respect to the above basis (3). An elegant framework for this is the *multiresolution analysis* introduced by Mallat (1989).

2.3.1 Multiresolution analysis

In statistics, we are often faced with discretely-sampled signals and therefore we need to be able to perform wavelet decomposition of vectors, rather than continuous functions as above. The multiresolution analysis framework is commonly used to define discrete wavelet filters. The starting point is a scaling function ϕ and a multiresolution analysis of $\mathbb{L}_2(\mathbb{R})$, i.e. a sequence $\{V_j\}_{j\in\mathbb{Z}}$ of closed subspaces of $\mathbb{L}_2(\mathbb{R})$ such that

- $\{\phi(x - k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_0;
- $\ldots \subset V_{-1} \subset V_0 \subset V_1 \subset \ldots \subset \mathbb{L}_2(\mathbb{R})$;

- $f \in V_j \iff f(2\cdot) \in V_{j+1}$;
- $\bigcap_j V_j = \{0\}$, $\overline{\bigcup_j V_j} = \mathbb{L}_2(\mathbb{R})$.

The set $\{\sqrt{2}\phi(2x-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_1 since the map $f \mapsto \sqrt{2}f(2\cdot)$ is an isometry from V_0 onto V_1. The function ϕ is in V_1 so it must have an expansion

$$\phi(x) = \sqrt{2}\sum_k h_k \phi(2x-k), \quad \{h_k\}_k \in l_2, \quad x \in \mathbb{R}. \tag{4}$$

Once we have the scaling function ϕ, we use it to define the wavelet function (also called the *mother wavelet*) ψ. We define the latter in such a way that $\{\psi(x-k)\}_k$ is an orthonormal basis for the space W_0, being the orthogonal complement of V_0 in V_1:

$$V_1 = V_0 \oplus W_0. \tag{5}$$

Defining $W_j = \text{span}\{\psi_{j,k} : k \in \mathbb{Z}\}$, we obtain that W_j is the orthogonal complement of V_j in V_{j+1}. We can write

$$V_{j+1} = V_j \oplus W_j = \ldots = V_0 \oplus \left(\bigoplus_{i=0}^{j} W_i\right), \tag{6}$$

or, taking the limit (recall that $\bigcup_j V_j$ is dense in $\mathbb{L}_2(\mathbb{R})$),

$$\mathbb{L}_2(\mathbb{R}) = V_0 \oplus \left(\bigoplus_{i=0}^{\infty} W_i\right) = V_{j_0} \oplus \left(\bigoplus_{i=j_0}^{\infty} W_i\right), \quad \forall j_0. \tag{7}$$

There are precise procedures for finding ψ once ϕ is known (see Daubechies, 1992, Section 5.1). One possibility (Daubechies, 1992, Theorem 5.1.1) is to set

$$\psi(x) = \sqrt{2}\sum_k h_{1-k}(-1)^k \phi(2x-k). \tag{8}$$

It can be shown that the appropriate orthogonality conditions are satisfied.

2.3.2 Algorithm for the Discrete Wavelet Transform

The nested structure of the multiresolution analysis can be exploited to construct a fast decomposition-reconstruction algorithm for discrete data, analogous to the Fast Fourier Transform of Cooley & Tukey (1965). The algorithm, called the *Discrete Wavelet Transform* (Mallat, 1989) produces a vector of wavelet coefficients of the input vector at dyadic scales and locations. The transformation is linear and orthonormal but is not performed by matrix multiplication to save time and memory.

We first describe a single *reconstruction* step, used in computing the inverse Discrete Wavelet Transform (DWT). The following two sets are orthonormal bases for V_1: $\{\sqrt{2}\phi(2x-k)\}_{k\in\mathbb{Z}}$, $\{\phi(x-k), \psi(x-l)\}_{k,l\in\mathbb{Z}}$. Using (4) and (8), we obtain for any $f \in V_1$

$$\begin{aligned} f(x) &= \sum_k c_{0,k}\phi(x-k) + \sum_k d_{0,k}\psi(x-k) \\ &= \sum_l \left(\sum_k h_l c_{0,k} + \sum_k h_{1-l}(-1)^l d_{0,k} \right) \sqrt{2}\phi(2x-2k-l) \\ &= \sum_{l'} \left(\sum_k h_{l'-2k} c_{0,k} + \sum_k h_{1-l'+2k}(-1)^{l'} d_{0,k} \right) \sqrt{2}\phi(2x-l'). \end{aligned}$$

Writing the expansion w.r.t. the other basis as $f(x) = \sum_{l'} c_{1,l'}\sqrt{2}\phi(2x-l')$ and equating the coefficients, we obtain

$$c_{1,l'} = \sum_k h_{l'-2k} c_{0,k} + \sum_k h_{1-l'+2k}(-1)^{l'} d_{0,k}, \tag{9}$$

which completes the reconstruction part: the coarser scale coefficients $\{c_{0,k}\}, \{d_{0,k}\}$ are used to obtain the finer scale coefficients $\{c_{1,k}\}$.

The *decomposition* step (used in the DWT) is equally straightforward: we have

$$\begin{aligned} c_{0,k} &= \int_{-\infty}^{\infty} f(x)\phi(x-k)dx \\ &= \int_{-\infty}^{\infty} f(x)\sum_l h_l\sqrt{2}\phi(2x-2k-l)dx \\ &= \sum_l h_l c_{1,2k+l} = \sum_l c_{1,l}h_{l-2k}. \end{aligned} \tag{10}$$

Similarly,

$$d_{0,k} = \sum_l (-1)^{l-2k} h_{1-l+2k} c_{1,l}. \tag{11}$$

The same mechanism works for each scale: $\{c_{j,k}\}$ gives $\{c_{j-1,k}\}$ and $\{d_{j-1,k}\}$ for all j. On the other hand, $\{c_{j,k}\}$ can be reconstructed using $\{c_{j-1,k}\}$ and $\{d_{j-1,k}\}$ for all j. To start this "pyramid" algorithm, we only need to compute the scaling coefficients $c_{j,k}$ at the finest scale of interest, say $j = J$. Indeed, when performing wavelet decomposition of finite sequences, it is commonly assumed that our input vector $\mathbf{f} = \{f_n\}_{n=0}^{2^J-1}$ is a vector of scaling coefficients of a function f, i.e. $f_n = c_{J,n} = \langle f, \phi_{J,n}\rangle$, where $\phi_{j,k} = 2^{j/2}\phi(2^j x - k)$. The DWT of $\mathbf{f}$ is given by

$$\mathrm{DWT}(\mathbf{f}) = (c_{0,0}, d_{0,0}, d_{1,0}, d_{1,1}, d_{2,0}, \ldots, d_{2,3}, \ldots, d_{J-1,0}, \ldots, d_{J-1,2^{J-1}-1}). \tag{12}$$

Informally speaking, the wavelet coefficients $d_{j,k}$ contain information on the local oscillatory behaviour of $\mathbf{f}$ at scale j and location $2^{J-j}k$, whereas the coefficient $c_{0,0}$ contains information on the global "mean level" of $\mathbf{f}$. A few remarks are in order.

Decimation. Define

$$\begin{aligned} c^*_{0,k} &= \sum_l c_{1,l}h_{l-k} \\ d^*_{0,k} &= \sum_l (-1)^{l-k} h_{1-l+k} c_{1,l}, \end{aligned}$$

so that $c^*_{0,k}$ is a convolution of $c_{1,k}$ with h_k, and $d^*_{0,k}$ is a convolution of $c_{1,k}$ with $(-1)^k h_{1-k}$. We have $c_{0,k} = c^*_{0,2k}$ and $d_{0,k} = d^*_{0,2k}$: coarser scale coefficients are *decimated* convolutions of finer scale coefficients with fixed (scale-independent) filters. This is in contrast to the *Non-decimated Wavelet Transform* where no decimation is performed, yielding a shift-invariant (but redundant) transform: see Section 2.4 for details.

High-pass and low-pass filters. We define $g_k = (-1)^k h_{1-k}$. Due to its effect in the frequency domain, g_k (h_k) is often referred to as a *high-pass (low-pass) filter* in the wavelet literature. This motivates the commonly used name for the wavelet and scaling coefficients: they are often referred to as *detail* and *smooth* coefficients, respectively.

Example of the DWT. By simple algebra, $\phi^H(x) = \mathbb{I}(0 \leq x \leq 1)$ generates the Haar wavelet ψ^H, with a low-pass filter h_k s.t. $h_0 = h_1 = 1/\sqrt{2}$, $h_k = 0$ otherwise, and a high-pass filter g_k s.t. $g_0 = -g_1 = 1/\sqrt{2}$, $g_k = 0$ otherwise. We shall now decompose a four-element vector

$$(c_{2,0}, c_{2,1}, c_{2,2}, c_{2,3}) = (1, 1, 2, 3)$$

using the DWT with Haar wavelets. By (10) and (11), we obtain

$$\begin{aligned}
c_{1,0} &= 1/\sqrt{2} \times 1 + 1/\sqrt{2} \times 1 = \sqrt{2} \\
c_{1,1} &= 1/\sqrt{2} \times 2 + 1/\sqrt{2} \times 3 = 5/\sqrt{2} \\
d_{1,0} &= 1/\sqrt{2} \times 1 - 1/\sqrt{2} \times 1 = 0 \\
d_{1,1} &= 1/\sqrt{2} \times 2 - 1/\sqrt{2} \times 3 = -1/\sqrt{2}.
\end{aligned}$$

Continuing at the next coarser scale, we obtain

$$\begin{aligned}
c_{0,0} &= 1/\sqrt{2} \times \sqrt{2} + 1/\sqrt{2} \times 5/\sqrt{2} = 7/2 \\
d_{0,0} &= 1/\sqrt{2} \times \sqrt{2} - 1/\sqrt{2} \times 5/\sqrt{2} = -3/2.
\end{aligned}$$

The original vector $(c_{2,0}, c_{2,1}, c_{2,2}, c_{2,3})$ can now be easily reconstructed from $(c_{0,0}, d_{0,0}, d_{1,0}, d_{1,1})$, (i.e. from the smooth coefficient at the coarsest scale and the detail coefficients at all scales) using the inverse DWT. As the DWT is orthonormal, the inverse DWT uses exactly the same filters as the DWT.

Note that the high-pass filter annihilates constants (recall that Haar wavelets have vanishing moments up to degree 0). Wavelets with higher numbers of vanishing moments are capable of annihilating polynomials of higher degrees.

Boundary issue. With wavelet filters longer than Haar, there often arises the problem of what action to perform when the support of the filter extends beyond the support of the input vector. Several solutions have been proposed, including symmetric reflection of the input vector at the boundaries, polynomial extrapolation, periodising the vector, padding it out with zeros, etc. See Nason & Silverman (1994) for an overview. Cohen *et al.* (1993) introduced *wavelets on the interval*, i.e. wavelet bases for functions defined on an interval as opposed to the whole real line. They also proposed a corresponding fast wavelet transform which uses filters adapted to the finite support situation. The lifting scheme offers a natural way of dealing with the boundary problem.

Computational speed. $O(n)$ operations are needed for the DWT which uses a compactly-supported wavelet, where n is the size of the input sequence. This is an advantage over the Fast Fourier Transform, which requires $O(n \log(n))$ operations.

2.4 Non-decimated Wavelet Transform

An undesirable property of the DWT is that it is not translation-invariant, and that at any given scale, it only provides information about the input vector at certain (dyadic) locations. Using the toy example above, the coefficient $c_{1,0}$ uses $c_{2,0}$ and $c_{2,1}$, while the coefficient $c_{1,1}$ uses $c_{2,2}$ and $c_{2,3}$, but there is no coefficient which would use, say, $c_{2,1}$ and

$c_{2,2}$. Motivated by this, Pesquet *et al.* (1996) introduce a Non-decimated DWT (NDWT) which remedies this problem by computing wavelet coefficients at all possible locations at all scales (see also Nason & Silverman, 1995; Coifman & Donoho, 1995). Continuing the example of the previous section, the NDWT of $(c_{2,0}, c_{2,1}, c_{2,2}, c_{2,3}) = (1, 1, 2, 3)$ which uses Haar wavelets is performed as follows. We begin with

$$\begin{aligned} c_{1,0} &= (1/\sqrt{2}, 1/\sqrt{2}) \cdot (c_{2,0}, c_{2,1}) \\ c_{1,1} &= (1/\sqrt{2}, 1/\sqrt{2}) \cdot (c_{2,1}, c_{2,2}) \\ c_{1,2} &= (1/\sqrt{2}, 1/\sqrt{2}) \cdot (c_{2,2}, c_{2,3}) \\ c_{1,3} &= (1/\sqrt{2}, 1/\sqrt{2}) \cdot (c_{2,3}, c_{2,0}), \end{aligned}$$

where the "$\cdot$" denotes the dot product. The detail coefficients $d_{1,k}$ are obtained similarly by replacing the low-pass filter with the high-pass one. Note that we implicitly assume "periodic" boundary conditions in the above (see the remark on the "boundary issue" in Section 2.3.2). Before we proceed to the next stage, we insert zeros between each two elements of the wavelet filters. Thus, we have

$$\begin{aligned} c_{0,0} &= (1/\sqrt{2}, 0, 1/\sqrt{2}, 0) \cdot (c_{1,0}, c_{1,1}, c_{1,2}, c_{1,3}) \\ c_{0,1} &= (1/\sqrt{2}, 0, 1/\sqrt{2}, 0) \cdot (c_{1,1}, c_{1,2}, c_{1,3}, c_{1,0}) \\ c_{0,2} &= (1/\sqrt{2}, 0, 1/\sqrt{2}, 0) \cdot (c_{1,2}, c_{1,3}, c_{1,0}, c_{1,1}) \\ c_{0,3} &= (1/\sqrt{2}, 0, 1/\sqrt{2}, 0) \cdot (c_{1,3}, c_{1,0}, c_{1,1}, c_{1,2}), \end{aligned}$$

and similarly for the detail coefficients. The insertion of zeros is necessary since decimation is not performed. Were we to compute the NDWT at yet another scale, we would use the filter $(1/\sqrt{2}, 0, 0, 0, 1/\sqrt{2}, 0, 0, 0)$ for the smooth and $(1/\sqrt{2}, 0, 0, 0, -1/\sqrt{2}, 0, 0, 0)$ for the detail. The computational speed of the NDWT is $O(n \log(n))$, where n is the length of the input vector.

2.5 Visualisation of discrete and non-decimated wavelet transforms

Typically, the result of the DWT is depicted as a binary tree whose main node is the coefficient $d_{0,0}$ (scale 0, location 0), its "children" are the coefficients $d_{1,0}$ and $d_{1,1}$, and so on. The DWT of the noisy vector of Figure 1 (using "DaubExPhase 2" wavelets) is shown in the top plot of Figure 5. The numbers along the y-axis denote scale ($j = 0$ is the coarsest scale; $j = 10 = \log_2(2048) - 1$ is the finest scale).

Contrary to the DWT where there are 2^j coefficients at each scale j, the NDWT always has n coefficients at each scale. Thus it is natural to display them as in the bottom plot of Figure 5.

2.6 Recent (and less recent) extensions of wavelets

Since the late eighties, several extensions and modifications of wavelets have been proposed. For more details and **references** on the following topics, see Vidakovic (1999), Chapter 5:

- multivariate version of the DWT
- biorthogonal wavelets (two mutually orthogonal wavelet bases neither of which is itself orthonormal)
- multiwavelets (which use translations and dilations of more that one wavelet function)
- complex-valued wavelets
- wavelet packets (over-complete collections of linear combinations of wavelets; work by applying both low- and high-pass filters to both smooth and detail coefficients; can be rapidly searched for the "best basis" representation)
- lifting scheme: alternative construction of wavelets for irregularly spaced data.

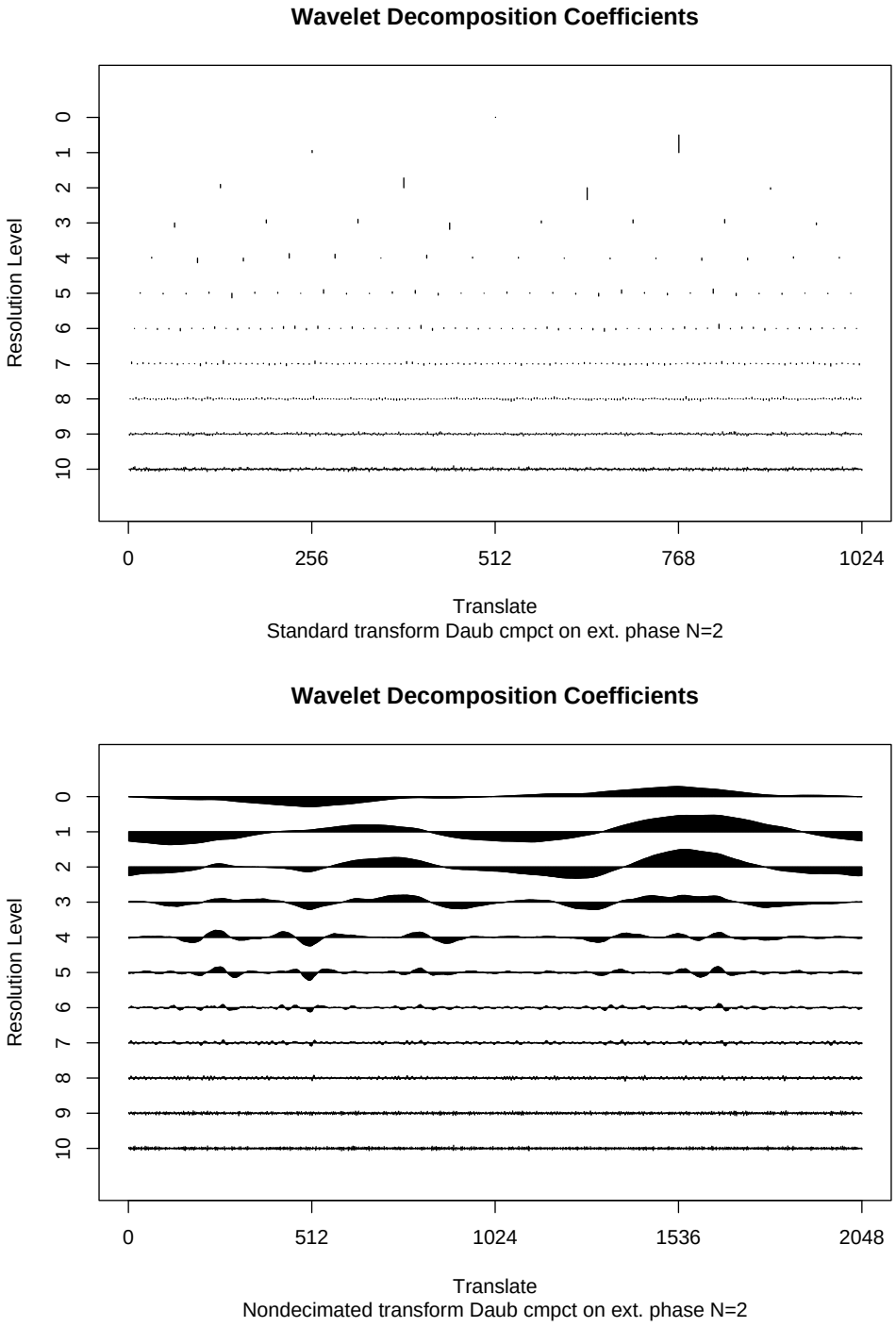

Figure 5: Top: DWT of noisy vector of Figure 1. Bottom: its NDWT. Both using "DaubEx-Phase 2" wavelets.

Recently, more and more research effort has been spent trying to find sparse multiscale representations of images. Here challenges are different from 1D because the types of singularities encountered in images are different. Those efforts have resulted in *ridgelets*, *curvelets*, *wedgelets*, *beamlets* and possibly other 'lets'.

A readable introduction to this topic can be found here:

`http://www-stat.stanford.edu/~donoho/Lectures/CBMS/CBMSLect.html`

2.7 Applications of wavelets

Wavelets and their extensions have been applied in a multitude of areas, such as signal and image processing, data compression, computer graphics, astronomy, quantum mechanics and turbulence: for a discussion of these and other areas of application see the monographs of Ruskai (1992) and Jaffard *et al.* (2001). An important field of application is numerical analysis, extensively covered in Cohen (2003). One can venture to say that wavelets are indeed one of those fortunate mathematical concepts that have almost become "household objects": for example, they were used in the JPEG2000 compression algorithm and also to compress the CIA fingerprint database. Multiscale subdivision schemes, related to wavelets, were employed in some recent animated movies such as "A Bug's Life".

Following Vidakovic (1999), who gives a comprehensive overview of wavelet applications in statistics, we list some of the most important areas of statistics where wavelets have been successfully applied:

- time series analysis,
- non-parametric function estimation,
- density estimation,
- deconvolution and inverse problems,

- statistical turbulence.

In Section 3, we describe how wavelets have been applied in nonparametric function estimation.

3 Wavelets for nonparametric function estimation

The setup is

$$y_i = f(i/n) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $f(i/n)$ is unknown and needs to be estimated, and the noise ϵ_i is iid with $\mathbb{E}(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma^2$.

For irregular (e.g. discontinuous) functions, linear (e.g. kernel) smoothing performs inadequately, and non-linear smoothing methods are needed. In a seminal paper, Donoho & Johnstone (1994) introduce the principle of a non-linear smoothing method called *wavelet thresholding*. First, the signal is transformed via the DWT to obtain $d_{j,k} = \theta_{j,k} + \epsilon_{j,k}$, where $d_{j,k}$, $(\theta_{j,k}, \epsilon_{j,k})$ is the DWT of y_i $(f(i/n), \epsilon_i)$. Then, $d_{j,k}$ are shrunk towards zero (with the threshold chosen in an appropriate manner), and finally the inverse DWT is taken to obtain an estimate of f. The rationale behind this principle is twofold:

- As DWT is orthonormal, i.i.d. Gaussian noise in the time domain transforms into i.i.d. Gaussian noise in the wavelet domain;
- Due to the vanishing moments property, wavelet coefficients $\theta_{j,k}$ corresponding to the locations where the signal is smooth will be close to zero. On the other hand, those (hopefully few) corresponding to discontinuities or other irregularities will be significantly different from zero: the signal will be represented *sparsely* in the wavelet domain. Therefore, we can expect that an appropriately chosen threshold will be able to accurately separate signal from noise.

Two thresholding rules have been particularly commonly used and well-studied. For a given threshold λ, *hard* and *soft* thresholding shrink $d_{j,k}$ to

$$\begin{aligned} d^h_{j,k} &= d_{j,k}\mathbb{I}(|d_{j,k}| > \lambda) \\ d^s_{j,k} &= \text{sgn}(d_{j,k})(|d_{j,k}| - \lambda)_+, \end{aligned}$$

respectively. The threshold introduced in Donoho & Johnstone (1994) was the so-called *universal threshold*, $\lambda = \sigma\sqrt{2\log(n)}$. The authors show that the MSE of the soft thresholding estimator with the universal threshold is close (within a logarithmic factor) to the ideal risk one can achieve by "keeping" or "killing" the wavelet coefficients $d_{j,k}$ using knowledge of the underlying signal. At the same time, the universal threshold is an efficient noise suppressor as described in Section 4.2 of their paper.

In another ground-breaking paper, Donoho & Johnstone (1995) consider a non-linear wavelet estimator with soft thresholding where the threshold selection procedure is based on Stein's shrinkage method for estimating the mean of multivariate normal variables. They consider the behaviour of the estimator over a range of so-called Besov spaces (Triebel, 1983), which form an extremely rich collection of functions with various degrees of smoothness (for certain values of the space parameters, Besov spaces can be shown to contain other better known function spaces such as Hölder or Sobolev spaces or the space of functions with bounded variation). The authors demonstrate that their estimator is *simultaneously nearly minimax* over a range of Besov balls, i.e. without knowing the regularity of the function, it nearly achieves the optimal rate of convergence which could be achieved if the regularity were known.

In most papers on the theory of non-linear wavelet estimation, it is assumed that the standard deviation σ of the noise is known. In practice, it needs to be estimated from the data. For Gaussian data, the method recommended by several authors (see e.g. Johnstone & Silverman, 1997) computes the scaled Median Absolute Deviation (MAD) on the sequence

of wavelet coefficients at the finest resolution level, thereby ensuring robustness.

More recently, other thresholding rules have been proposed. Nason (1996) uses cross-validation as a way of selecting the threshold. Abramovich & Benjamini (1996) set up wavelet thresholding as a multiple hypothesis testing problem and propose an approach based on the so-called *false discovery rate*. Johnstone & Silverman (1997) consider level-dependent universal thresholding for correlated Gaussian noise. Averkamp & Houdré (2003) extend the approach of Donoho & Johnstone (1994) to other noise distributions such as exponential, mixture of normals or compactly supported distributions. Vanreas *et al.* (2002) consider stable wavelet transforms for denoising data observed on non-equispaced grids. Barber & Nason (2004) develop various thresholding procedures using complex-valued wavelets. Johnstone & Silverman (2005) propose an empirical Bayes approach to the threshold selection problem. Cai & Silverman (2001), amongst others, consider *block thresholding*: they propose a thresholding procedure whereby wavelet coefficients are considered in overlapping blocks and the action performed on the coefficients in the middle of the block depends upon the data in the whole block. Antoniadis & Fryzlewicz (2006) propose a simple universal-type thresholding procedure where the threshold values are modelled parametrically across scales.

Coifman & Donoho (1995) introduce *translation invariant denoising*: the full NDWT transform of the data is taken, then the universal threshold is applied to all resulting wavelet coefficients, and then an inverse NDWT transform yields an estimate of the signal. As the NDWT is redundant, there are many possible ways of generating an inverse NDWT transform: the one proposed by the authors is equivalent to taking the average over all possible DWT's contained in the NDWT, corresponding to all possible circular shifts of the data set (hence the name "translation invariant").

3.1 Simple example: Haar wavelets + piecewise constant regression function

In this section, we show how to prove mean-square consistency of a hard-thresholding universal estimator of a piecewise-constant regression function contaminated with independent Gaussian $N(0,1)$ noise. The number of jumps in the function f is unknown but finite (bounded by M). As before, $d_{j,k}$, $\theta_{j,k}$ and $\epsilon_{j,k}$ are the Haar wavelet coefficients of y_i, $f(i/n)$ and ϵ_i, respectively. The range of (j,k) is $j = 0, \ldots, J-1 := \log_2 n - 1$; $k = 1, \ldots, 2^j$. The only smooth coefficient is indexed by $(j,k) = (-1,1)$. The wavelet noise coefficients $\epsilon_{j,k}$ are iid $N(0,1)$ because the Haar transform is orthonormal.

Except $(j,k) = (-1,1)$ where we leave the coefficient intact, we estimate $\theta_{j,k}$ by

$$\hat{\theta}_{j,k} = d_{j,k}\mathbb{I}(|d_{j,k}| > \lambda),$$

where $\lambda = \sqrt{2 \log n}$, ie λ is the universal threshold. Then the estimate $\hat{f}(i/n)$ is constructed by applying the inverse Haar transform to $\hat{\theta}_{j,k}$. We are interested in the mean-square error

$$\mathrm{MSE}(\hat{f}, f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(f(i/n) - \hat{f}(i/n))^2. \tag{13}$$

Lemma 3.1 (Parseval inequality) *Let W be an orthonormal matrix, x a column vector, and $y = Wx$. Then $x^T x = y^T y$.*

Proof. As W is orthonormal, we have $W^{-1} = W^T$. Thus $y^T y = x^T W^T W x = x^T x$. □

Applying this to (13), we obtain

$$\mathrm{MSE}(\hat{f}, f) = \frac{1}{n}\sum_{j,k} \mathbb{E}(\hat{\theta}_{j,k} - \theta_{j,k})^2.$$

Since f is piecewise constant, at most M coefficients $\theta_{j,k}$ at each scale j are non-zero. The rest of them (corresponding to the intervals where f is constant), are zero. Remember that Haar cofficients are local differences which "annihilate" constants!

Let us first look at the case $\theta_{j,k} = 0$, so that $d_{j,k}$ is distributed as $N(0,1)$. We have

$$\begin{aligned}\mathbb{E}(\hat{\theta}_{j,k} - \theta_{j,k})^2 &= \mathbb{E}\hat{\theta}_{j,k}^2 = \mathbb{E}d_{j,k}^2\mathbb{I}(|d_{j,k}| > \lambda) = \sqrt{2/\pi}\int_\lambda^\infty x^2\exp(-x^2/2)dx \\ &= \sqrt{2/\pi}\lambda\exp(-\lambda^2/2) + 2(1-\Phi(\lambda)),\end{aligned}$$

where Φ is the cdf of the standard normal. By a "standard result",

$$1 - \Phi(\lambda) \leq \phi(\lambda)/\lambda,$$

where ϕ is the pdf of the standard normal. Thus

$$\mathbb{E}(\hat{\theta}_{j,k} - \theta_{j,k})^2 \leq \sqrt{2/\pi}\exp(-\lambda^2/2)(\lambda + \lambda^{-1}) = O\left(\frac{\log^{1/2} n}{n}\right).$$

We now move to the case $\theta_{j,k} \neq 0$ and without loss of generality, we assume $\theta_{j,k} > 0$.

$$\begin{aligned}\mathbb{E}(\hat{\theta}_{j,k} - \theta_{j,k})^2 &= \mathbb{E}(d_{j,k}\mathbb{I}(|d_{j,k}| > \lambda) - \theta_{j,k})^2 \\ &= \mathbb{E}(d_{j,k}\mathbb{I}(|d_{j,k}| > \lambda) - \theta_{j,k}\mathbb{I}(|d_{j,k}| > \lambda) + \theta_{j,k}\mathbb{I}(|d_{j,k}| > \lambda) - \theta_{j,k})^2 \\ &\leq 2\mathbb{E}(d_{j,k}\mathbb{I}(|d_{j,k}| > \lambda) - \theta_{j,k}\mathbb{I}(|d_{j,k}| > \lambda))^2 + 2\mathbb{E}(\theta_{j,k}\mathbb{I}(|d_{j,k}| > \lambda) - \theta_{j,k})^2 \\ &\leq 2\mathrm{var}(d_{j,k}) + 2\theta_{j,k}^2\mathbb{P}(|d_{j,k}| \leq \lambda) \leq 2 + 2\theta_{j,k}^2\mathbb{P}(d_{j,k} \leq \lambda) \\ &= 2 + 2\theta_{j,k}^2\mathbb{P}(\lambda + \theta_{j,k} - d_{j,k} \geq \theta_{j,k}).\end{aligned}$$

By Markov's inequality,

$$\mathbb{P}(\lambda + \theta_{j,k} - d_{j,k} \geq \theta_{j,k}) \leq \mathbb{E}(\lambda + \theta_{j,k} - d_{j,k})^2/\theta_{j,k}^2.$$

This gives

$$\begin{aligned}\mathbb{E}(\hat{\theta}_{j,k} - \theta_{j,k})^2 &\leq 2 + 2\mathbb{E}(\lambda + \theta_{j,k} - d_{j,k})^2 \\ &\leq 2 + 4(\lambda^2 + \mathrm{var}(d_{j,k})) = 4\lambda^2 + 6 = O(\log\ n).\end{aligned}$$

This finally gives

$$\begin{aligned}\mathrm{MSE}(\hat{f}, f) &= \frac{1}{n}\sum_{j,k}\mathbb{E}(\hat{\theta}_{j,k} - \theta_{j,k})^2 \\ &\leq O(1/n^2) \quad \text{[smooth coefficient]} \\ &+ 1/n \times n \times O\left(\frac{\log^{1/2} n}{n}\right) \quad \text{[coefficients with } \theta_{j,k} = 0\text{]} \\ &+ 1/n \times J \times M \times O(\log\ n) \quad \text{[coefficients with } \theta_{j,k} \neq 0\text{]} \\ &= O(n^{-1}\log^2\ n).\end{aligned}$$

3.2 Noise-free reconstruction property

Other than attaining the near-parametric MSE rate above, the universal threshold also enjoys the "noise-free reconstruction" property: if the true signal f is constant, then the estimate $\hat{f}$ is also constant and equal to the sample mean of the data, with high probability. For $\hat{f}$ to be constant, we need all $\hat{\theta}_{j,k}$'s to be zero with a high probability. This happens if all $d_{j,k}$'s exceed λ with a high probability. But if f is constant, then all $d_{j,k}$'s are (independent and) distributed as $N(0,1)$. The noise-free reconstruction property is implied by the following fact:

$$\lim_{n\to\infty}\mathbb{P}\left(\max_{j,k}|d_{j,k}| > \sqrt{a\log\ n}\right) = 0,$$

if and only if $a \geq 2$. So the universal threshold is asymptotically the "lowest" threshold satisfying the noise-free reconstruction property.

References

[1] Abramovich, F. & Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Comput. Statist. Data Anal.*, **22**, 351–361.

[2] Antoniadis, A. & Fryzlewicz, P. (2006). Parametric modelling of thresholds across scales in wavelet regression. *Biometrika*, to appear.

[3] Averkamp, R. & Houdré, C. (2003). Wavelet thresholding for non-necessarily Gaussian noise: idealism. *Ann. Stat.*, **31**, 110–151.

[4] Barber, S. & Nason, G.P. (2004). Real nonparametric regression using complex wavelets. *J. Roy. Statist. Soc. Ser. B*, **66**, 927–939.

[5] Cai, T. & Silverman, B.W. (2001). Incorporating information on neighbouring coefficients into wavelet estimation. *Sankhyā Ser. B*, **63**, 127–148.

[6] Cohen, A., Daubechies, I. & Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, **1**, 54–81.

[7] Cohen, A. (2003). *Numerical Analysis of Wavelet Methods*. Studies in Mathematics and Its Applications, vol. 32. Elsevier.

[8] Coifman, R.R. & Donoho, D.L. (1995). Translation-invariant de-noising. *Technical Report, Statistics Department, Stanford University, USA*.

[9] Cooley, J.W. and Tukey, O.W. (1965). An algorithm for the machine calculation of complex Fourier series. *Math. Comput.*, **19**, 297–301.

[10] Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia, Pa.: SIAM.

[11] Davies, P.L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution (with discussion). *Ann. Statist.*, **29**, 1–65.

[12] Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

[13] Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Stat. Assoc.*, **90**, 1200–1224.

[14] Haar, A. (1910). Zur Theorie der orthogonalen Funktionensysteme. *Math. Ann.*, **69**, 331–371.

[15] Jaffard, S., Meyer, Y. & Ryan, R.D. (2001). *Wavelets: Tools for Science & Technology.* Philadelphia, Pa.: SIAM.

[16] Johnstone, I.M. & Silverman, B.W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B*, **59**, 319–351.

[17] Johnstone, I.M. & Silverman, B.W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Stat.*, **33**, 1700–1752.

[18] Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, **11**, 674–693.

[19] Mallat, S. (1998). *A Wavelet Tour of Signal Processing.* Academic Press.

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

[21] Morlet, J., Arens, G., Fourgeau, E. & Giard, D. (1982). Wave propagation and sampling theory. *Geophysics*, **47**, 203–236.

[22] Nason, G.P. (1996). Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc. Ser. B*, **58**, 463–479.

[23] Nason, G.P. & Silverman, B.W. (1994). The discrete wavelet transform in S. *J. Comput. Graph. Statist.*, **3**, 163–191.

[24] Nason, G.P. & Silverman, B.W. (1995). The stationary wavelet transform and some statistical applications. *Pages 281–300 of:* Antoniadis, A. & Oppenheim, G. (eds), *Lecture Notes in Statistics, vol. 103*. Springer-Verlag.

[25] Pesquet, J.C., Krim, H. & Carfantan, H. (1996). Time-invariant orthonormal wavelet representations. *IEEE Trans. Sig. Proc.*, **44**, 1964–1970.

[26] Polzehl, J. and Spokoiny, V. (2000). Adaptive weights smoothing with applications to image restoration. *J. Roy. Stat. Soc. B*, **62**, 335–354.

[27] Ruskai, M.B. (ed). (1992). *Wavelets and Their Applications*. Jones and Bartlett books in mathematics. Jones and Bartlett.

[28] Triebel, H. (1983). *Theory of Function Spaces*. Basel: Birkhäuser Verlag.

[29] Vanreas, E., Jansen, M. & Bultheel, A. (2002). Stabilized wavelet transforms for non-equispaced data smoothing. *Signal Procssing*, **82**, 1979–1990.

[30] Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. New York: John Wiley & Sons.