Article type: Overview

Wavelet Methods **Article ID**

Piotr Fryzlewicz

London School of Economics

Keywords

wavelets, nonparametric smoothing, thresholding, Unbalanced Haar wavelets, Haar-Fisz transforms

Abstract

This overview article motivates the use of wavelets in statistics, and introduces the basic mathematics behind the construction of wavelets. Topics covered include the continuous and discrete wavelet transforms, multiresolution analysis and the non-decimated wavelet transform. We describe the basic mechanics of nonparametric function estimation via wavelets, emphasising the concepts of sparsity and thresholding. A simple proof of the mean-square consistency of the wavelet estimator is also included. The article ends with two special topics: function estimation with Unbalanced Haar wavelets, and variance stabilisation via the Haar-Fisz transformation.

Wavelets are mathematical functions which, when plotted, resemble "little waves": that is, they are compactly or almost-compactly supported, and they integrate to zero. This is in contrast to "big waves" – sines and cosines in Fourier analysis, which also oscillate, but the amplitude of their oscillation never changes.

Wavelets are useful for decomposing data into "wavelet coefficients", which can then be processed in a way which depends on the aim of the analysis. One possibly advantageous feature of this decomposition is that in some set-ups, the decomposition will be *sparse*, i.e. most of the coefficients will be close to zero, with only a few coefficients carrying most of the information about the data. One can imagine obvious uses of this fact, e.g. in image compression. The decomposition is particularly informative, fast and easy to invert if it is performed using wavelets at a range of *scales* and *locations*. The role of scale is similar to the role of frequency in Fourier analysis. However, the concept of location is unique to wavelets: as mentioned above, they are localised around a particular point of the domain, unlike Fourier functions.

This article provides a self-contained introduction to the applications of wavelets in statistics and attempts to justify the extreme popularity which they have enjoyed in the literature over the past 15 years.

Motivation

Possibly the main statistical application of wavelets is in nonparametric function estimation, also known as "signal denoising" or "smoothing". As a motivating example, consider the simulated noisy signal in the top-left plot of Figure 1. Our objective is to remove the noise and get as close as possible to revealing the true structure of the signal. For many readers, it will be apparent that the signal is composed of at least 5 different pieces. It is interesting to investigate whether some state-of-the-art smoothing techniques can reveal more than this.

The black line in the top-right plot of Figure 1 is the result of applying the "adaptive weights smoothing" technique of Polzehl and Spokoiny (2000). The reconstruction is good but it misses some of the dips in the signal (the true signal is plotted in red). The function used to produce the reconstruction was `aws` from the R package `aws` (version 1.6-1, published 12 October 2009), called with its default parameter values.

The black line in the bottom-left plot of Figure 1 is the result of smoothing the signal using the "taut string" methodology of Davies and Kovac (2001). Again, the reconstruction is good, but it over-detects the number of jumps in the signal. The function used to produce the reconstruction was `pmreg` from the R package `ftnonpar` (version 0.1-83, published 28 July 2008), called with its default parameter values.

Finally, the black line in the bottom-right plot of Figure 1 is a reconstruction which uses nonlinear wavelet shrinkage with Unbalanced Haar wavelets. The reconstruction is probably as good as it can be, in that it correctly detects all jumps in this extremely noisy signal. This methodology will be described in more detailed in the section **Unbalanced Haar wavelets and function estimation** later on. The function used to produce the reconstruction was `uh` from the R package `unbalhaar` (version 1.0, published 27 July 2006), called with its default parameter values.

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 discovery in the early eighties, 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). Monographs on statistical applications of wavelets include Härdle *et al.* (1998), Vidakovic (1999) and Nason (2008). Some of the material in this section is based on Vidakovic (1999). We also note the recent review article by Antoniadis (2007).

Formally, let $\psi_{a,b}(x)$, $a \in R \setminus \{0\}$, $b \in R$ be a family of functions being translations and dilations of a single function $\psi(x) \in L_2(R)$,

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

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

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

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.

Continuous wavelet transform

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

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

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

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

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

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

Examples of wavelets

Haar wavelets

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

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

which implies

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

for $a > 0$, $b \in R$. A wavelet ψ is said to have 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, $\text{CWT}_f(a, b) = 0$, and thus piecewise-constant functions will be "sparsely" represented by Haar wavelets, with their continuous Haar transform corresponding to those intervals taking the value of zero.

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 2 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 come back to this "sparsity" property of wavelets in the section **Wavelets for nonparametric function estimation**.

Discrete wavelet transform

$\text{CWT}_f(a, b)$ is a function of two real variables, thus being a redundant transform. To minimise the transform, the values of a and b can be discretised 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 $a = 2^{-j}$, $b = k2^{-j}$, $j, k \in Z$, will produce a basis for L_2. Moreover, under mild conditions on the wavelet function ψ, the resulting basis

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

will be orthonormal. In the remainder of this article, 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).

Multiresolution analysis

Statisticians are often faced with discretely-sampled signals and therefore 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 $L_2(R)$, i.e. a sequence $\{V_j\}_{j\in Z}$ of closed subspaces of $L_2(R)$ such that

- $\{\phi(x-k)\}_{k\in Z}$ is an orthonormal basis for V_0;
- $\ldots \subset V_{-1} \subset V_0 \subset V_1 \subset \ldots \subset L_2(R)$;
- $f \in V_j \iff f(2\cdot) \in V_{j+1}$;
- $\bigcap_j V_j = \{0\}, \overline{\bigcup_j V_j} = L_2(R)$.

The set $\{\sqrt{2}\phi(2x-k)\}_{k\in Z}$ 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

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

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 $W_j = \mathrm{span}\{\psi_{j,k} : k \in Z\}$, 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 $L_2(R)$),

$$L_2(R) = V_0 \oplus \left(\bigoplus_{i=0}^{\infty} W_i\right) = V_{j_0} \oplus \left(\bigoplus_{i=j_0}^{\infty} W_i\right), \quad \forall j_0. \tag{7}$$

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.

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 and 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 via 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: $\{\sqrt{2}\phi(2x-k)\}_{k\in Z}$, $\{\phi(x-k),\psi(x-l)\}_{k,l\in Z}$. 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 **Non-decimated Wavelet Transform** 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, $\phi^H(x) = I(0 \le x \le 1)$ 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 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 and 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 (Sweldens, 1996) 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.

Non-decimated Wavelet Transform

An often 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 and Silverman, 1995; Coifman and 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 the previous section. 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.

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 3. 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 3. Note that Figure 3 was produced in the R package `wavethresh` by Guy Nason.

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 orthonormal itself);
- 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.

Some research effort has been spent trying to find sparse multiscale representations of more complex objects such as 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 at

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

Applications of wavelets

Wavelets and their extensions have been applied in a multitude of areas, such as signal and image processing, data compression, communication, 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 to compress the CIA fingerprint database. Multiscale subdivision schemes, related to wavelets, were employed in some 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 **Wavelets for nonparametric function estimation**, we describe how wavelets have been applied in this important area of statistics.

Wavelets for nonparametric function estimation

In nonparametric function estimation, the basic 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 $E(\epsilon_i) = 0$, 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 and 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 ($f(i/n)$, ϵ_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

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

respectively. The threshold introduced in Donoho and 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 paper, Donoho and 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 was 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 and 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 means of selecting the threshold. Abramovich and 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 and Silverman (1997) consider level-dependent universal thresholding for correlated Gaussian noise. Averkamp and Houdré (2003) extend the approach of Donoho and 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 and Nason (2004) develop various thresholding procedures using complex-valued wavelets. Johnstone and Silverman (2005) propose an empirical Bayes approach to the threshold selection problem. Cai and 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 and Fryzlewicz (2006) propose a simple universal-type thresholding procedure where the threshold values are modelled parametrically across scales.

Coifman and 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").

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

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

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

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

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

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. This is because, essentially, Haar cofficients are local differences which annihilate constants, i.e. transform them to zero.

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

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

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

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

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

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

By Markov's inequality,

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

This gives

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

This finally gives

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

which proves the mean-square consistency of the Haar wavelet estimator at the nearly-parametric rate of $O(n^{-1} \log^2\, n)$.

Noise-free reconstruction property

Other than attaining the nearly-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. 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 fall below λ with a high probability. But if f is constant, then all $d_{j,k}$'s are i.i.d. $N(0, 1)$. The noise-free reconstruction property is implied by the following fact:

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

if and only if $a \geq 2$. Thus, the universal threshold $\sqrt{2 \log\, n}$ is asymptotically the lowest threshold satisfying the noise-free reconstruction property.

Unbalanced Haar wavelets and function estimation

Haar wavelets differ from other wavelet families in that nonparametric function estimation (as described in the previous section) using this wavelet family will result in a piecewise constant estimate. The analyst might be tempted to interpret this estimate as indicating intervals where the underlying true signal is well approximated by a constant, as well as indicating the likely locations of "jumps" in the signal.

However, this would not normally be an entirely accurate interpretation. One must bear in mind that Haar wavelets contain a "jump" exactly in the middle of their support, which, combined with the dyadic structure of the Haar wavelet transform, implies that the estimator will be more likely to display jumps at "dyadic" locations, i.e. $1/2, 1/4, 3/4, \ldots$, irrespective of the locations of the jumps in the true underlying signal.

To provide a more accurate description of the true piecewise constant structure of the underlying signal (and thus facilitate the interpretation of the estimate), Fryzlewicz (2007) proposes estimation using so-called "Unbalanced" Haar (UH) wavelets, which contain a jump not necessarily in the middle of their support. The location of the jumps in the UH wavelets can be chosen adaptively to match the likely structure of the signal at hand.

We first give a description of the construction of the UH wavelet vectors. Suppose that our domain is indexed by $i = 1, \ldots, n$, and that $n \geq 2$. We first construct a vector $\psi^{0,1}$, which is constant and positive for $i = 1, \ldots, b^{0,1}$, and constant and negative for $i = b^{0,1} + 1, \ldots, n$. The breakpoint $b^{0,1} < n$ is to be chosen by the analyst. The positive and negative values taken by $\psi^{0,1}$ are chosen in such a way that (a) the elements of $\psi^{0,1}$ sum to zero, and (b) the squared elements of $\psi^{0,1}$ sum to one.

We then recursively repeat this construction on the two parts of the domain determined by $\psi^{0,1}$: that is, provided that $b^{0,1} \geq 2$, we construct (in a similar fashion) a vector $\psi^{1,1}$ supported on $i = 1, \ldots, b^{0,1}$, with a breakpoint $b^{1,1}$. Also, provided that $n - b^{0,1} \geq 2$, we construct a vector $\psi^{1,2}$ supported on $i = b^{0,1} + 1, \ldots, n$ with a breakpoint $b^{1,2}$. The recursion then continues in the same manner for as long as feasible, with each vector $\psi^{j,k}$ having at most two "children" vectors $\psi^{j+1,2k-1}$ and $\psi^{j+1,2k}$. For each vector $\psi^{j,k}$, their start, breakpoint and end indices are denoted by $s^{j,k}$, $b^{j,k}$ and $e^{j,k}$, respectively. Additionally, we define a vector $\psi^{-1,1}$ with elements $\psi^{-1,1}(l) = n^{-1/2} I(1 \leq l \leq n)$, where $I(\cdot)$ is the indicator function. Note that to shorten notation, we do not explicitly emphasise the dependence of $\psi^{j,k}$ on $(s^{j,k}, b^{j,k}, e^{j,k})$. As in the classical wavelet theory, the indices j, k are scale and location parameters, respectively. Small (large) values of j can be thought of as corresponding to "coarse" ("fine") scales.

Example. We consider an example of a set of UH vectors for $n = 6$. The rows of the matrix $\mathbf{W}$ defined below contain (from top to bottom) vectors $\psi^{-1,1}$, $\psi^{0,1}$, $\psi^{1,2}$, $\psi^{2,3}$, $\psi^{2,4}$ and $\psi^{3,7}$ determined by the following set of breakpoints: $(b^{0,1}, b^{1,2}, b^{2,3}, b^{2,4}, b^{3,7}) = (1, 3, 2, 5, 4)$.

$$\mathbf{W} = \begin{pmatrix} 6^{-1/2} & 6^{-1/2} & 6^{-1/2} & 6^{-1/2} & 6^{-1/2} & 6^{-1/2} \\ \{5/6\}^{1/2} & -30^{-1/2} & -30^{-1/2} & -30^{-1/2} & -30^{-1/2} & -30^{-1/2} \\ 0 & \{3/10\}^{1/2} & \{3/10\}^{1/2} & -\{2/15\}^{1/2} & -\{2/15\}^{1/2} & -\{2/15\}^{1/2} \\ 0 & 2^{-1/2} & -2^{-1/2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 6^{-1/2} & 6^{-1/2} & -\{2/3\}^{1/2} \\ 0 & 0 & 0 & 2^{-1/2} & -2^{-1/2} & 0 \end{pmatrix}$$

In the above example, it is not possible to create further vectors $\psi^{j,k}$. There are $n = 6$ of them, and they are orthonormal. Thus, they form an orthonormal basis of R^6. This is not a coincidence: the following general results holds.

Proposition 1 *The collection of vectors $\{\psi^{j,k}\}_{j,k}$ is an orthonormal basis of R^n.*

Nonparametric function estimation using UH wavelets proceeds as follows: we first choose an appropriate basis, then take the transform of the noisy data with respect to

this basis, threshold the coefficients, and take the inverse transform. In other words, the only difference between classical and Unbalanced Haar wavelets is the basis selection step: in the classical Haar wavelets, the basis is fixed.

One basis selection procedure described in Fryzlewicz (2007) is the following greedy *forward stagewise* procedure, related to the *matching pursuit* algorithm of Mallat and Zhang (1993) and to *binary segmentation* of Sen and Srivastava (1975). We first define the *UH mother vector* $\psi_{s,b,e}$ with elements defined by

$$\psi_{s,b,e}(l) = \left\{ \frac{1}{b-s+1} - \frac{1}{e-s+1} \right\}^{1/2} I(s \le l \le b) - \left\{ \frac{1}{e-b} - \frac{1}{e-s+1} \right\}^{1/2} I(b+1 \le l \le e).$$

- Breakpoint $b^{0,1}$ is chosen such that the inner product $\langle \mathbf{X}, \psi_{1,b^{0,1},n} \rangle$ between the data $\mathbf{X}$ and $\psi_{1,b^{0,1},n}$ is maximised in absolute value.
- Similarly, $b^{j+1,l} := \text{argmax}_b |\langle \mathbf{X}, \psi_{s^{j+1,l},b,e^{j+1,l}} \rangle|$, where $l = 2k-1, 2k$.

Under a mild assumption on the permitted degree of "unbalancedness" of the thus-constructed UH basis, the computational complexity of the above procedure is $O(n \log\, n)$. The motivation for this basis selection procedure can be outlined as follows: it is known that wavelet thresholding is the most successful when the representation of the signal in the wavelet domain is *sparse*. In our set-up, this would require that only a few UH coefficients were "large" in magnitude, whilst most were "small" and thus carried mainly noise. Typically, when performing transforms with the standard Haar basis, it is often observed that large Haar coefficients are mostly concentrated at coarser scales. The above basis selection procedure makes this "concentration of power" even more extreme: it attempts to concentrate as much as possible of the signal power at coarser scales, in the hope of further improving the sparsity of representation.

Following this line of thought, Fryzlewicz (2007) also proposes an alternative, *backward stagewise* basis selection algorithm, which proceeds from the finest to coarsest scale, attempting to concentrate as *little* power as possible at fine scales, which produces a similar effect: concentrates the bulk of the power of the signal at coarse scales. This strategy can be termed a *generous*, as opposed to "greedy", algorithm. Such a *bottom-up* UH basis selection algorithm is a natural starting point for the meaningful extension of the UH idea to smoother wavelet bases and to image data.

We note that the UH estimate displayed in Figure 1 was computed using the forward UH basis selection procedure as described above. UH estimation is implemented in the R package `unbalhaar`.

Haar-Fisz transforms: variance stabilisation via wavelets

The previous sections describe how wavelets can be useful in nonparametric function estimation. In this section, we show how wavelets can be used to stabilise noise variance in nonparametric regression set-ups where the variance of the noise depends on

the local mean level of the signal. We deliberately change the notation to avoid confusion with the previous sections and consider

$$X_t = \alpha(t/n) + \varepsilon_t, \quad t = 1, \ldots, n, \tag{14}$$

where the X_t's are modelled as independent, and Var(ε_t) depends on $\alpha(t/n)$. Examples of such set-ups include

- *Poisson intensity estimation.* In Poisson intensity estimation, X_t are modelled as independent Pois$\{\alpha(t/n)\}$ variables, which implies that ε_t are centered Poisson. The mean and variance of X_t are linked via the relationship Var$(X_t) = h\{E(X_t)\}$ with $h(u) = u$.
- *Nonparametric volatility estimation.* Nonparametric volatility estimation techniques are widely used in the finance industry. In this set-up, the X_t's represent squared log-returns on a financial instrument and are modelled as independent and distributed as $X_t = \alpha(t/n)\, Z_t^2$, where $E(Z_t^2) = 1$. Note that $\varepsilon_t = \alpha(t/n)(Z_t^2 - 1)$. Thus, the model is multiplicative and the variance function $h(u)$ is proportional to u^2.
- *Spectral density estimation.* In spectral density estimation based on the periodogram, the X_t's represent periodogram ordinates and are assumed to be asymptotically independent and asymptotically distributed as $\alpha(t/n)\, Z_t^2$, where $\alpha(t/n)$ represents the spectral density at frequency t/n, and Z_t^2 are Exp(1) random variables. This again makes the set-up multiplicative and, asymptotically, the variance function takes the form $h(u) = u^2$.

In the above and similar set-ups, variance stabilisation is often desirable as many statistical estimation and testing techniques work best when the data at hand are homogeneous, or even Gaussian. The classical tool for variance stabilisation is the well-known Box-Cox transform, which in the case of Poisson intensity estimation would simply square-root the data (up to constants), and in the case of nonparametric volatility or spectral density estimation – log the data.

Alternatively, variance stabilisation can be performed in the wavelet domain, leading to the so-called *Haar-Fisz* and, more generally, *wavelet-Fisz* transform, first introduced in the context of Poisson intensity estimation by Fryzlewicz and Nason (2004) and described more fully in a general context by Fryzlewicz (2008). One advantage of wavelet-Fisz transforms over Box-Cox transforms is that the former lead to data with approximately Gaussian (in particular: more symmetric) noise, unlike the latter.

In the set-up (14), a simple Haar-Fisz transform would proceed as follows.

1. Take the usual Haar wavelet transform of X_t, to obtain the detail coefficients $d_{j,k}$ and the smooth coefficients $c_{j,k}$.
2. Modify the smooth coefficients at scales $j = 1, \ldots, J-1$ to transform them into local means of the data, i.e. form $c^*_{j,k} = 2^{(j-J)/2} c_{j,k}$.

3. Note that $\mathrm{Var}(d_{j,k})$ is approximately equal to $h(\alpha_{j,k})$, where $\alpha_{j,k}$ denotes the local mean of the true function $\alpha(\cdot)$ computed over the same support as the corresponding coefficients $d_{j,k}$ and $c_{j,k}$. Further, note that $\alpha_{j,k}$ can be pre-estimated by $c^*_{j,k}$.

4. Thus, to stabilise the variance of $d_{j,k}$, form the Haar-Fisz stabilised coefficients

$$d^*_{j,k} = \frac{d_{j,k}}{h^{1/2}(c^*_{j,k})}.$$

   This can be viewed as a kind of "studentization" in the wavelet domain.

5. Take the inverse Haar transform of the transformed coefficients $d^*_{j,k}$. The variance of the data is now stabilised.

We note that the transform is easily invertible.

As an example, consider the series of squared logged daily returns on the Dow Jones Industrial Average index, observed on 2048 consecutive trading days ending 10 March 2010. The series, together with its log and Haar-Fisz transforms, is plotted in Figure 4. It is clear that both transforms stabilise the variance of the original series very well. However, the log transform leads to a distribution with a downward skew and spikes which obscure the picture. By contrast, the Haar-Fisz transform leads to a series with symmetric and "almost-Gaussian" noise, and brings out the overall shape of the signal more clearly.

Conclusion

Wavelets not only had revolutionised engineering and statistics, but also had given rise to, and emphasised the importance of, other concepts which subsequently took these disciplines by storm. For example, the concept of sparsity, popularised by wavelets, gained further popularity in the late nineties and 2000s thanks to the many attempts at solving, theoretically and practically, the problem of high-dimensional variable selection (such as LASSO or the Dantzig selector), as well as to modern signal recovery techniques such as compressed sensing.

We strongly feel that the potential of wavelets in modern statistics has not yet been fully explored. In the ongoing data revolution, statisticians are routinely challenged by having to manipulate ever more massive and complex datasets, which also defy classical assumptions, such as stationarity. Wavelets, with their in-built high computational speed, and the ability to sparsify (and thereby reduce complexity) appear ideal for handling large datasets. Another attractive aspect is their localisation in space and frequency, which means that they are naturally suited to provide a good descriptive framework for phenomena whose characteristics evolve over time or space.

References

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

[2] Antoniadis, A. and Fryzlewicz, P. (2006). Parametric modelling of thresholds across scales in wavelet regression. *Biometrika*, **93**, 465–471.

[3] Antoniadis, A. (2007). Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, **1**, 16–55.

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

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

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

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

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

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

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

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

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

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

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

[15] Fryzlewicz, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *J. Amer. Stat. Assoc.*, **102**, 1318–1327.

[16] Fryzlewicz, P. (2008). Data-driven wavelet-Fisz methodology for nonparametric function estimation. *Elec. J. Stat.*, **2**, 863–896.

[17] Fryzlewicz, P. and Nason, G.P. (2004). A Haar-Fisz algorithm for Poisson intensity estimation. *J. Comp. Graph. Stat.*, **13**, 621-638.

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

[19] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications*. New York: Springer.

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

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

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

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

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

[25] Mallat, S. and Zhang, Z. (1993). Matching pursuit with time-frequency dictionaries. *IEEE Trans. Sig. Proc.*, **41**, 3397–3415.

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

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

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

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

[30] Nason, G.P. (2008). *Wavelet methods in statistics with R*. New York: Springer.

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

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

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

[34] Sen, A. and Srivastava, M.S. (1975). On tests for detecting change in mean, *Ann. Stat.*, **3**, 98–108.

[35] Sweldens, W. (1996). The lifting scheme: A custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmon. Anal.*, **3**, 186–200.

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

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

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

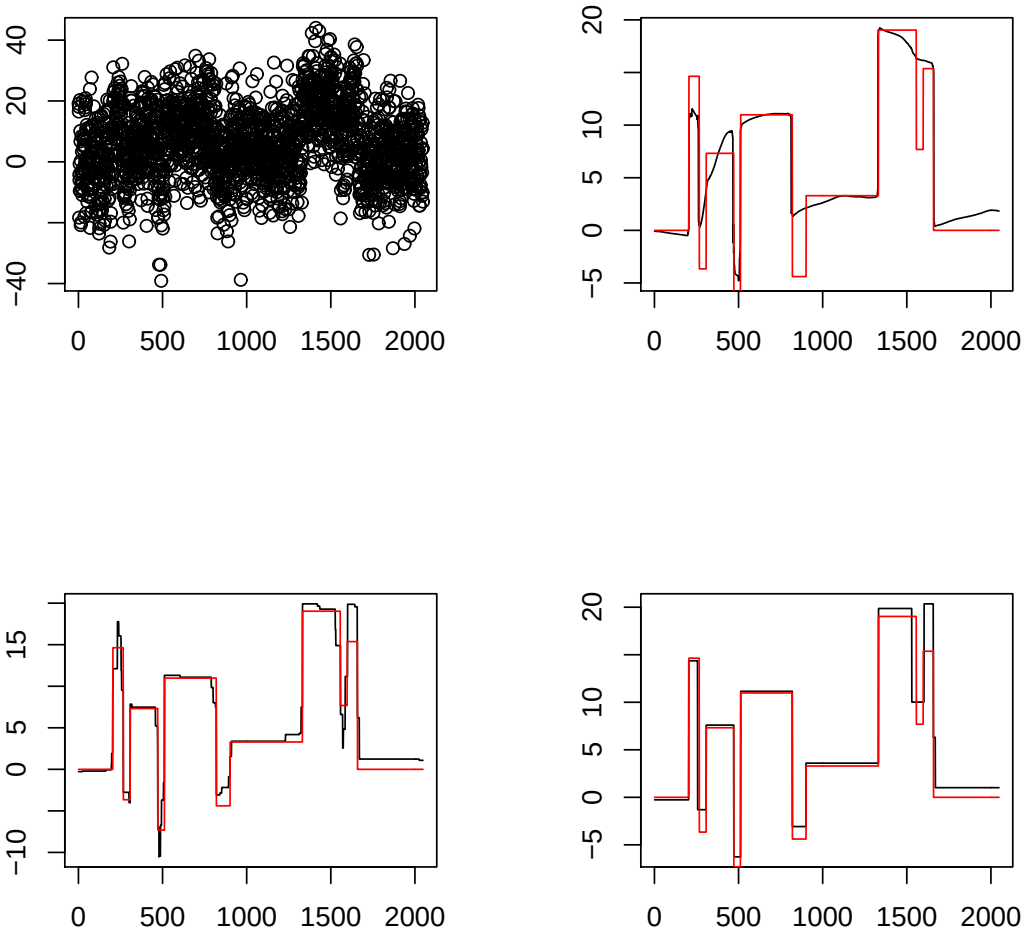

Figure 1: Top-left: simulated noisy signal. Top-right: reconstruction via Adaptive Weights Smoothing (black) and true signal (red; also in other plots). Bottom-left: reconstruction via Taut String methodology. Bottom-right: reconstruction via Unbalanced Haar wavelets.

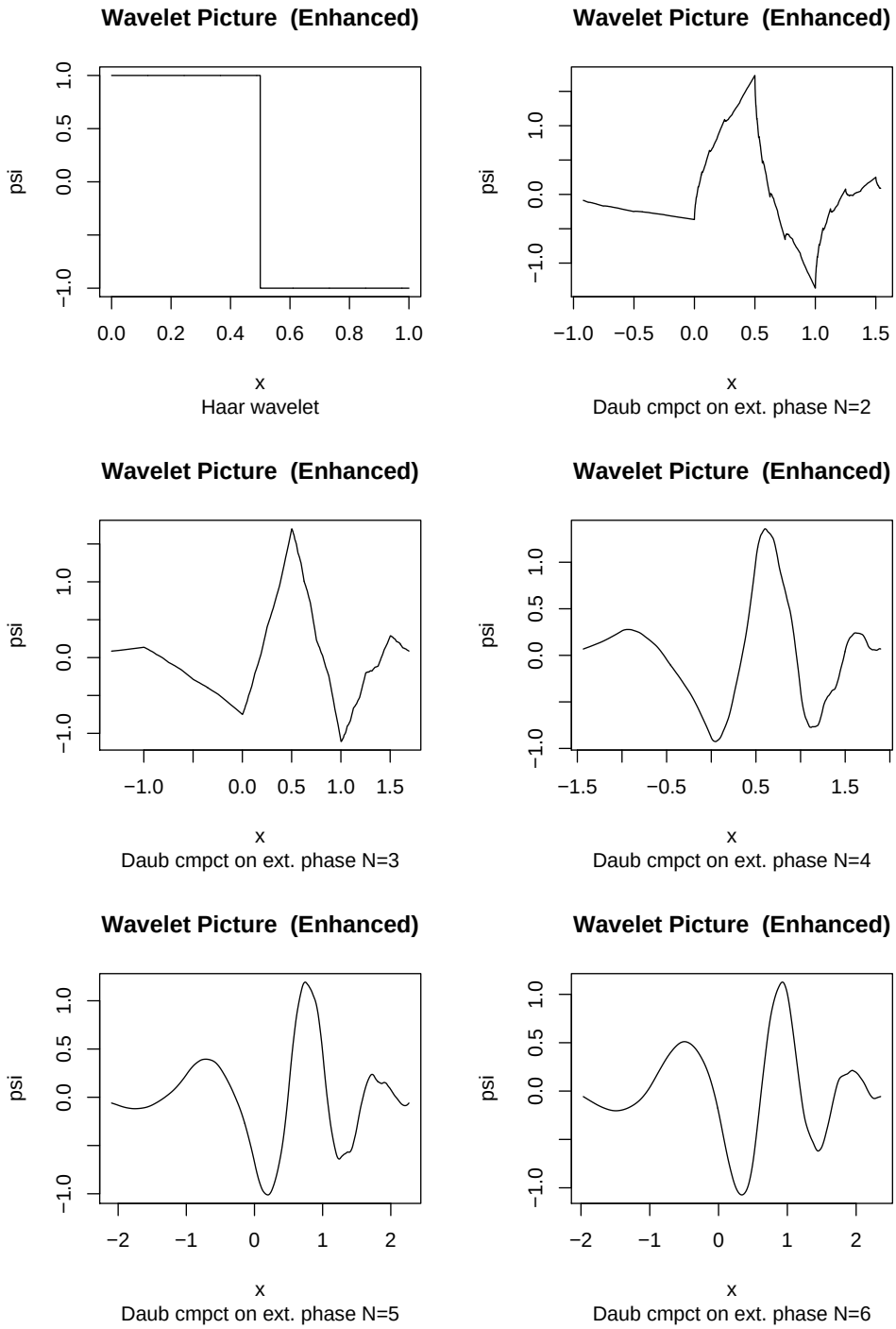

Figure 2: Daubechies' Extremal Phase wavelets with different numbers $n = N - 1$ of vanishing moments.

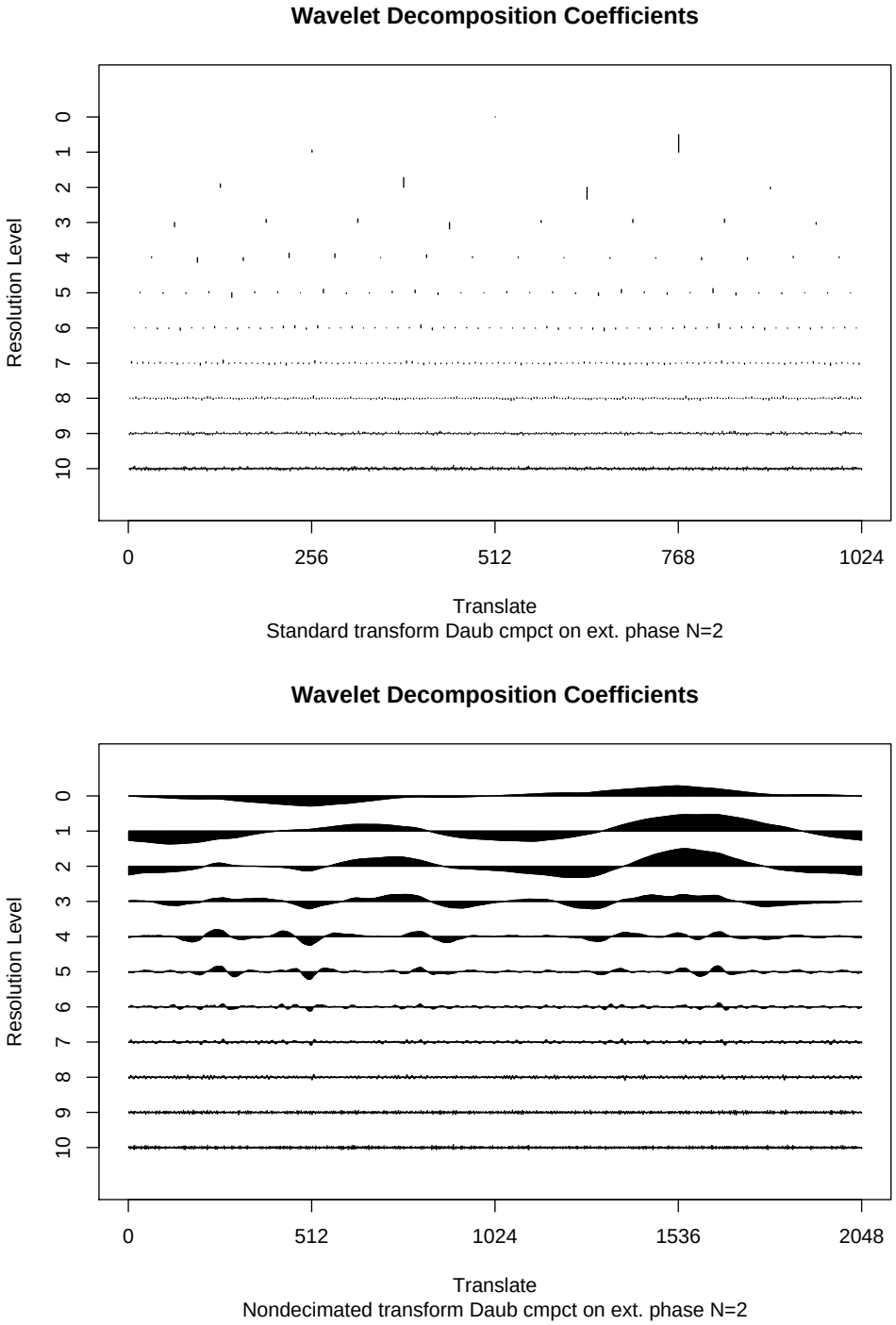

Figure 3: Top: DWT of noisy vector of Figure 1. Bottom: its NDWT. Both using "DaubExPhase 2" wavelets.

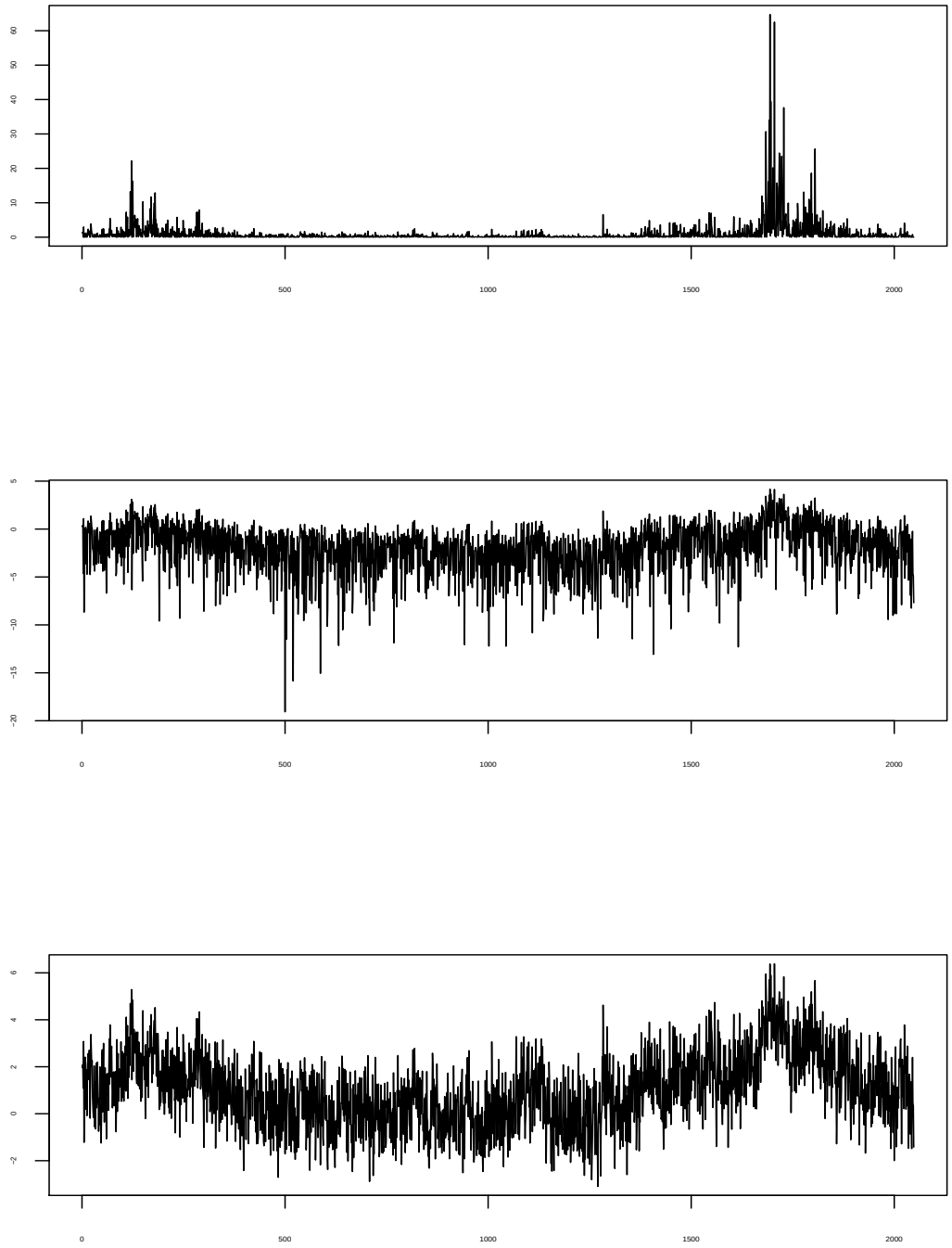

Figure 4: Top: squared daily log-returns (normalised so that overall variance is one) of the Dow Jones Industrial Average index on 2048 trading days ending 10 March 2010. Middle plot: logged. Bottom plot: Haar-Fisz'ed.