

# Wavelets

by

**Alan J. Izenman**

## Introduction

Wavelet-based techniques are very popular in signal processing and image analysis, and especially for nonparametric function estimation, including density estimation. Indeed, a *wavelet-based density estimator* (Donoho, Johnstone, Kerkyacharian, and Picard, 1996; Hall, McKay, and Turlach, 1996) is a special type of orthogonal series density estimator.

*Wavelets* (*ondelettes*, in French) form a natural extension of Fourier analysis, which has had an enormous effect on science, engineering, and mathematics. The primary tool in Fourier analysis is the *Fourier transform*, where certain functions of time can be represented by a weighted sum of sine and cosine waves, and the weights (called *Fourier coefficients*) are functions of frequency; the weights change the amplitudes (i.e., heights) of the waves, and, by translating the waves, we can also change their phases.

One of the main defects of a Fourier transform is that it takes a function of time (the *signal*) and turns it into a function of frequency, while burying all the information about time in the phases of the transformed function. As a result, we say that Fourier basis functions are *localized in frequency*, but not in time. For example, the Fourier transform of a piece of music gives complete information about which notes (i.e., frequencies) make up the score, but not the time-order in which those notes are to be played. Similarly, if several instruments are playing together, we can recognize which they are by the frequency spectrum derived from the Fourier analysis of their combined sound signal; but we cannot discover which instruments play together with each other and when in the score they do this, or which instruments play separately from the others and when. In principle, we can recover information about the order of the musical notes from the phases; however, in practice, accurate recovery of the time order may be quite difficult and subject to distortion.

The wavelet transform, on the other hand, is *localized in frequency and time*: it takes a function of time (the signal) and produces a function of both frequency (via dilations) and time (via translations). This feature gives wavelets a big advantage over Fourier analysis, especially when the signal to be analyzed is nonstationary. Using an approach called “multiresolution analysis,” the amount of localization can be adjusted automatically: the signal is studied, first, at a low resolution which produces a global view of the function, at then at increasingly-higher resolutions, which produce the fine details.

A considerable literature has emerged detailing the various properties of

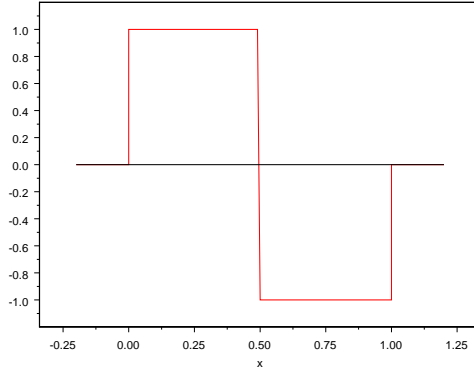


Figure 1: *Haar Wavelet*  $\psi(x) = \psi_{00}(x)$ .

wavelet-based methods. Wavelets have proved to be very flexible for the non-parametric estimation of curves and densities under different conditions of smoothness of the underlying function. In terms of MISE, wavelet-based estimators have been shown to have better properties than kernel methods when discontinuities or functions with sudden peaks are present.

## The Haar Wavelet

There are many different types of wavelets. We begin our discussion with a very simple example, the *Haar wavelet* (Haar, 1910). Haar proposed the following piecewise-constant function,

$$\psi(x) = I_{[0, \frac{1}{2})}(x) - I_{[\frac{1}{2}, 1]}(x), \quad (1)$$

which takes values 1 on  $[0, \frac{1}{2})$ ,  $-1$  on  $[\frac{1}{2}, 1]$ , and zero otherwise. and integrates to zero:

$$\int_0^1 \psi(x) dx = 0. \quad (2)$$

This particular  $\psi$ -function is called the *Haar mother wavelet* (or *Haar base function*). It has *compact support*, being zero everywhere outside the finite interval  $[0, 1]$ . See Figure 1 for a graph of  $\psi(x)$ .

Next, we transform  $\psi$  by using translation, dilation, and normalization operations:

$$\psi_{jk}(x) = 2^{j/2} \psi\left(\frac{x - 2^{-j}k}{2^{-j}}\right) = 2^{j/2} \psi(2^j x - k), \quad j, k \in Z, \quad (3)$$

where  $Z = \{0, \pm 1, \pm 2, \dots\}$  is the set of (positive and negative) integers. The transformations of  $\psi$  can be viewed as follows:

**Translation:** The graph of  $\psi(x)$  is shifted by a factor  $2^{-j}k$  along the horizontal axis (to the right if  $k > 0$  and to the left if  $k < 0$ ); this gives the ‘position’ of the wavelet.

**Dilation:** For  $j > 0$  ( $j < 0$ ), the graph of  $\psi(x - 2^{-j}k)$  is shrunk (expanded) along the horizontal axis by a “dilation” (or “scale”) factor  $2^{-j}$ ; this gives the ‘width’ of the wavelet.

**Normalization:** The graph of  $\psi(2^j x - k)$  is then stretched along the vertical axis by multiplying it by the factor  $2^{j/2}$ .

These three operations on  $\psi$  yield the graph of  $\psi_{jk}$ .

The wavelet,  $\psi_{jk}$ , can be visualized by fixing the *resolution level*  $j \in Z$  and letting  $k$  run over all the (positive and negative) integers, including zero. For fixed  $j \in Z$ ,  $\psi_{jk}$  is defined over the dyadic intervals

$$I_{jk} = [2^{-j}k, 2^{-j}(k+1)), \quad k \in Z, \quad (4)$$

each of which has common length  $2^{-j}$ . In the S+WAVELETS software package (Bruce and Gao, 1996),  $j$  is called the **level** index and  $k$  the **shift** index. See Figure 2 for graphs of  $\psi_{jk}$  for different  $j$  and  $k$ , where the rows display  $\psi_{1k}$ ,  $\psi_{2k}$ , and  $\psi_{3k}$ ,  $k = 0, 1, 2$ , corresponding to resolution levels  $j = 1, 2, 3$ .

It is not difficult to see that the set  $\{\psi_{jk}, j, k \in Z\}$  yields an orthogonal basis; this follows because

$$\int \psi_{jk}(x)\psi_{j'k'}(x)dx = \delta_{jj'}\delta_{kk'}, \quad (5)$$

which equals one only if  $j = j'$  and  $k = k'$  (where  $\delta_{ij}$  denotes the Dirac delta function); otherwise, the integral equals zero. Furthermore, we see that the  $L_2$ -norms of  $\psi_{jk}$  are the same for all  $j, k \in Z$ :  $\|\psi_{jk}\|^2 = \int |\psi_{jk}(x)|^2 dx = \|\psi\|^2 = 1$ .

This *orthonormal wavelet basis*,  $\{\psi_{jk}, j, k \in Z\}$ , can also be expressed in an alternative form. Let  $j_1 \in Z$  be a fixed integer, often referred to as the *critical resolution level*. Then, the basis is equivalent to

$$\{\phi_{j_1,k}, \psi_{jk}, j \geq j_1, k \in Z\}, \quad (6)$$

where the set  $\{\phi_{j_1,k}(x), k \in Z\}$  spans the same subspace as  $\{\psi_{jk}(x), j < j_1, k \in Z\}$ . Just as we defined the Haar wavelet basis function,  $\psi_{jk}$ , we now define the functions  $\{\phi_{j_1,k}\}$  as

$$\phi_{j_1,k}(x) = 2^{j_1/2} \phi\left(\frac{x - 2^{-j_1}k}{2^{-j_1}}\right) = 2^{j_1/2} \phi(2^{j_1}x - k), \quad k \in Z. \quad (7)$$

Here,  $\phi$  is called the *scaling function* (and sometimes *father wavelet*). For a given resolution level  $j_1 \in Z$ ,  $\phi_{j_1,k}$  enjoys a fixed shape, while changing the

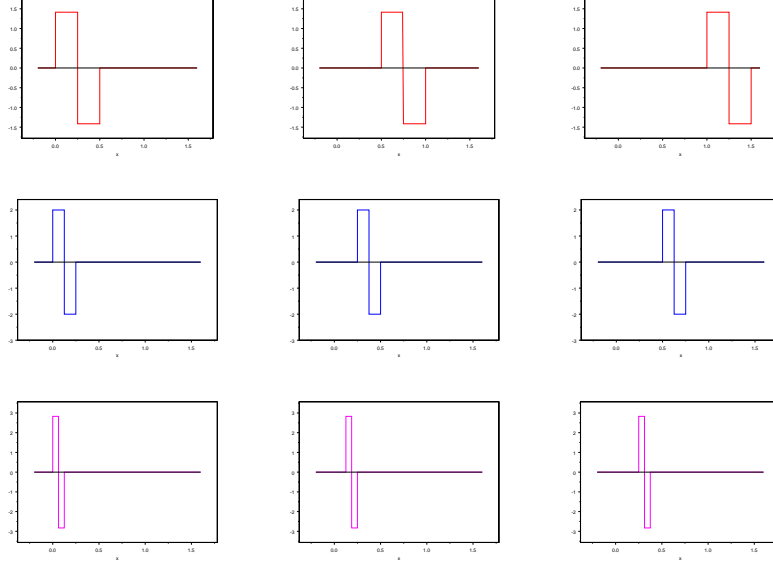


Figure 2: Haar wavelets,  $\psi_{jk}$ , for resolution level  $j = 1, 2, 3$  (rows) and  $k = 0, 1, 2$  (columns). Top row:  $\psi_{10}, \psi_{11}, \psi_{12}$ ; middle row:  $\psi_{20}, \psi_{21}, \psi_{22}$ ; bottom row:  $\psi_{30}, \psi_{31}, \psi_{32}$ . As  $k$  gets larger (smaller), the wavelet shifts to the right (left), while as  $j$  gets larger (smaller), the support width for the wavelet becomes finer (coarser).

value of  $k$  shifts the horizontal position of  $\phi_{j_1, k}$ . For the Haar wavelet basis,  $\phi$  is just the indicator function of the interval  $[0, 1]$ :

$$\phi(x) = I_{[0,1]}(x). \quad (8)$$

Note that for the Haar basis,  $[\psi(x)]^2 = \phi(x)$ . For a specific resolution  $j_1$ , the scaling function,  $\phi_{j_1, k}$ , forms a rectangle of height  $2^{j_1/2}$  over the interval

$$I_{j_1, k} = [2^{-j_1}k, 2^{-j_1}(k+1)]. \quad (9)$$

Each of the intervals  $I_{j_1, k}, k \in \mathbb{Z}$ , has common length  $2^{-j_1}$ .

The set  $\{\phi_{j_1, k}(x), k \in \mathbb{Z}\}$ , where  $\phi$  is the Haar scaling function, forms an orthogonal basis on  $L_2(\mathfrak{R})$  because

$$\int \phi_{j_1, k}(x)\phi_{j_1, k'}(x)dx = \delta_{k, k'}, \quad k, k' \in \mathbb{Z}, \quad (10)$$

which equals one only if  $k = k'$ ; otherwise, it equals zero. Furthermore,

$$\int \psi_{jk}(x)\phi_{j_1, k'}(x)dx = 0, \quad j \geq j_1, \quad k, k' \in \mathbb{Z}, \quad (11)$$

showing that, for all  $k \in Z$ , the *Haar wavelets*,  $\phi_{j_1,k}$  and  $\psi_{jk}$ ,  $j \geq j_1$ , form a pair of mutually-orthonormal wavelet functions.

## Multiresolution Analysis

Not every pair,  $(\phi_{j_1,k}, \psi_{jk})$ , of mutually-orthonormal wavelets constitutes a basis for  $L_2(\mathfrak{R})$ . In fact, the choice of  $\phi$  and  $\psi$  is crucial to the success of the wavelet technique. Although the Haar parent functions,  $\phi$  and  $\psi$ , provide us with a set of real-valued, compactly-supported, orthonormal wavelets, the Haar wavelet is not continuous, which makes it unappealing as a basis for approximating smooth functions.

It was not until the advent of *multiresolution analysis* that a firm mathematical foundation for wavelets could be constructed. Multiresolution analysis was formulated in 1986 by Stéphane Mallet and Yves Meyer as a way of understanding wavelet bases and as a recipe for generating new examples of such bases. This groundbreaking work (Mallat, 1989) showed that the problems of wavelet construction were essentially the same as those encountered by those working in image processing (pyramid algorithms), signal processing (subband coding), and digital speech processing (quadrature mirror filters). Prior to 1986, finding a smooth wavelet basis and then proving it to be orthonormal were not easy tasks. The multiresolution approach was introduced to explain in a mathematical way the difference in the information to be extracted from two resolutions, one twice as high as the other.

The goal is to provide a general framework for constructing an  $L_2$ -function  $\psi$  on  $\mathfrak{R}$ , so that the set  $\{\psi_{jk}(x) : j, k \in \mathfrak{R}\}$  will be an orthonormal basis for  $\mathfrak{R}$ .

### The Scaling Function $\phi(x)$

We start with a sequence of closed subspaces,  $\{V_j, j \in K\}$ , of  $L_2(\mathfrak{R})$ , where  $V_j$  forms the  $j$ th level of resolution for approximating a function  $f \in L_2(\mathfrak{R})$ . The subspaces are often referred to as *approximation spaces*. As  $j$  gets larger (smaller), the resolution becomes finer (coarser). These subspaces have to satisfy the following conditions:

1.  $\{0\} \subset \dots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \dots \subset L_2(\mathfrak{R})$
2.  $\overline{\cup_{j \in Z} V_j} = L_2(\mathfrak{R})$ .
3.  $\cap_{j \in Z} V_j = \{0\}$ .
4.  $f(x) \in V_j \Leftrightarrow f(2^j x) \in V_0, j \in Z$ .
5.  $f(x) \in V_0 \Rightarrow f(x - k) \in V_0$ , for all  $k \in Z$ .
6. There exists a *scaling function*  $\phi \in V_0$  such that  $\{\phi_{0k}(x) = \phi(x - k) : k \in Z\}$  forms an orthonormal basis for  $V_0$ .

In words, these conditions are that: (1) the subspaces form a nested sequence, (2) the union of the subspaces is dense in the whole space, (3) the only element common to all of the subspaces is the trivial function 0, (4) all the subspaces are scaled versions of the approximation space  $V_0$  having resolution 0, (5)  $V_0$  is invariant under integer translations, and (6) there exists a scaling function  $\phi$  which is orthogonal to its own integer translates. Note that

$$V_0 = \left\{ f \in L_2(\mathfrak{R}) : f(x) = \sum_{k \in Z} \alpha_{0k} \phi(x - k) \right\}. \quad (12)$$

Because  $\phi \in V_0 \subset V_1$ , it follows that  $\phi$  can be written as a linear combination of elements of  $V_1$ :

$$\phi(x) = \sum_{k \in Z} c_k \phi_{1,k}(x) = \sqrt{2} \sum_{k \in Z} c_k \phi(2x - k), \quad (13)$$

where

$$c_k = \sqrt{2} \int \phi(x) \phi(2x - k) dx \quad (14)$$

and  $\int \phi(x) dx = 1$ . The coefficients  $\{c_k\}$ , which satisfy  $\sum_k |c_k|^2 < \infty$ , are referred to collectively as a *scaling filter* (or, in signal processing, a *low-pass filter*). Conditions (4), (5), and (6) ensure that the set  $\{\phi_{jk} = 2^{j/2} \phi(2^j x - k) : k \in Z\}$  forms an orthonormal basis for  $V_j$ , for all  $j \in Z$ . The orthogonal projection of  $f \in L_2(\mathfrak{R})$  into  $V_j$  is

$$P_j f(x) = \sum_{k \in Z} \alpha_{jk} \phi_{jk}(x), \quad \alpha_{jk} = \int \phi_{jk}(x) f(x) dx, \quad (15)$$

which can be viewed as a  $2^{-j}$ -resolution approximation to  $f$ .

For example, the multiresolution analysis associated with the Haar basis is given by the sequence of subspaces,

$$V_j = \{f \in L_2(\mathfrak{R}) : \text{for all } k \in Z, f|_{I_{j1,k}} \text{ is constant}\}, \quad j \in Z, \quad (16)$$

where  $f|_{I_{j1,k}}$  refers to the function  $f$  restricted to the interval  $I_{j1,k}$ . We can take the scaling function  $\phi$  in this case to be the indicator function of the interval  $[0, 1]$ . For this (Haar) scaling function, it is easy to show that the only nonzero coefficients,  $\{c_k\}$ , are those corresponding to  $k = 0, 1$ , whence,  $c_0 = c_1 = 1/\sqrt{2}$ , and  $c_0 + c_1 = \sqrt{2}$ . From (13), we have that  $\phi(x) = \phi(2x) + \phi(2x - 1)$ .

### The Wavelet Function $\psi(x)$

Because  $V_j \subset V_{j+1}$ , the orthogonal complement of  $V_j$  in  $V_{j+1}$  is the closed set  $W_j$  of all points  $x \in V_{j+1}$  such that  $x \perp V_j$ . The direct sum,  $V_j \oplus W_j$ , is a closed subspace, which, by the projection theorem (see, e.g., Halmos, 1957, Section 12), equals  $V_{j+1}$ ; that is,

$$V_j \oplus W_j = V_{j+1}, \quad j \in Z. \quad (17)$$

Thus, any  $L_2$ -function in  $V_{j+1}$  can be written as the sum of two orthogonal parts, one part of which is in  $V_j$  and the other part is in  $W_j$ . Note that, for  $j \in Z$ ,  $W_j \subset V_{j+1} \perp W_{j+1}$ , so that, in general, we have that,  $W_j \perp W_{j'}$  for  $j \neq j'$ . This gives us a sequence,  $\{W_j, j \in Z\}$ , of closed mutually-orthogonal subspaces of  $L_2(\mathfrak{R})$ . By expanding (17), it follows that, for  $j < j_1$ ,

$$V_{j+1} = V_{j_1} \oplus \{W_{j_1} \oplus W_{j_1+1} \oplus \cdots \oplus W_j\}, \quad (18)$$

where all these subspaces are mutually orthogonal; see Figure 3 for a schematic of the decomposition of  $V_{j+1}$ , where we have taken  $j_1 = j - 3$ . Taking the limit as  $j \rightarrow \infty$  and using condition (2) above, we have that

$$L_2(\mathfrak{R}) = V_{j_1} \oplus \bigoplus_{m=0}^{\infty} W_{j_1+m}. \quad (19)$$

We next repeat this argument in the opposite direction. Thus,  $V_{j_1} = V_{j_1-1} \oplus W_{j_1-1} = V_{j_1-m} \oplus \{W_{j_1-m} \oplus W_{j_1-m+1} \oplus \cdots \oplus W_{j_1-1}\}$ . Now, let  $m \rightarrow \infty$ . Taking note of condition (3) above, we have that  $V_{j_1-m} \rightarrow \{0\}$ , whence,  $V_{j_1} = \bigoplus_{m=-\infty}^{-1} W_{j_1+m}$ . Substituting this result into (19), we have

$$\bigoplus_{m \in Z} W_j = L_2(\mathfrak{R}), \quad (20)$$

so that  $L_2(\mathfrak{R})$  is the direct sum of mutually-orthogonal subspaces.

Thus, there exists a *wavelet function*  $\psi \in W_0$  with the following properties:

1.  $\{\psi(x - k) : k \in Z\}$  is an orthonormal basis for  $W_0$ .
2. For any fixed  $j \in Z$ ,  $\{\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k) : k \in Z\}$  is an orthonormal basis for  $W_j$ .
3.  $\{\psi_{jk}(x) : j, k \in Z\}$  is an orthonormal basis for  $L_2(\mathfrak{R})$ .

The function  $\psi$  is usually called a *mother wavelet* and the derived set  $\{\psi_{jk}(x)\}$  is a *wavelet orthonormal basis on  $\mathfrak{R}$* . The  $\{W_j\}$  are called *wavelet subspaces*.

Because  $\psi \in W_0 \subset V_1$ , the wavelet function  $\psi$  can be written as a linear combination of elements of  $V_1$ :

$$\psi(x) = \sum_k d_k \phi_{1,k}(x) = \sqrt{2} \sum_k d_k \phi(2x - k), \quad (21)$$

where the coefficients  $d_k = (-1)^k c_{1-k}$  are known collectively as a *wavelet filter* (or, in signal processing, a *high-pass filter*) and the summations converge in the  $L_2$  sense. The *filter coefficients*  $\{c_k\}$  define the properties of the wavelet.

For the Haar wavelet function,  $d_0 = c_1 = 1/\sqrt{2}$  and  $d_1 = -c_0 = -1/\sqrt{2}$ , and, hence,  $d_0 + d_1 = 0$ . From (21), we have that  $\psi(x) = \phi(2x) - \phi(2x - 1)$ .

A wavelet function  $\psi$  can, therefore, be customized in a number of different ways. Two popular strategies are as follows:

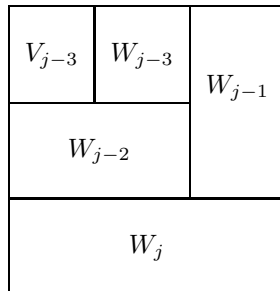


Figure 3: *Decomposition of  $V_{j+1}$  into mutually-orthogonal subspaces.*

1. Choose a scaling function  $\phi \in V_0$ , where  $\int \phi(x)dx \neq 0$ ; generate the associated multiresolution analysis  $\{V_j, j \in \mathbb{Z}\}$ ; use (14) to compute the scaling filter  $\{c_k\}$ ; and then use (21) to construct the wavelet function  $\psi$ .
2. Start by defining a scaling filter  $\{c_k\}$ , where  $\sum_k |c_k|^2 < \infty$ ; use (13) to construct a scaling function  $\phi$ ; and then use (21) to construct  $\psi$ .

## Other Wavelet Bases

Although many other wavelet bases have since been proposed using the multiresolution framework, none of them has a closed-form analytic formula for compactly-supported  $\phi$  and  $\psi$ , other than the Haar basis. These wavelets are instead the products of iterative computer algorithms.

The first types of orthonormal wavelets were constructed by the Swedish mathematician J.O. Strömberg in 1982 and Yves Meyer in 1985 to have infinite support. Stromberg's orthonormal wavelets had an arbitrarily large number of continuous derivatives and exponential decay, but were not very smooth.

Unaware of Strömberg's work, Meyer set out to prove that orthogonal wavelets could not exist. Instead, he was able to construct the exact type of wavelet basis that he thought did not exist. Meyer derived his wavelet function  $\psi$  through its Fourier transform,  $\widehat{\psi}(\omega)$ , which, as a function of frequency  $\omega$ , had compact support (Meyer, 1985). The associated scaling function  $\phi$  is also defined by its Fourier transform. No explicit formulas of  $\phi$  and  $\psi$  are available. For the explicit Fourier transforms of  $\phi$  and  $\psi$ , see Daubechies (1992, pp. 117 and 137). The Meyer wavelets form an orthonormal basis with infinite support, exponential decay, and an infinite number of *vanishing moments* (i.e.,  $\int x^\ell \psi(x)dx = 0$ , for any integer  $\ell > 0$ ).

Ingrid Daubechies (1988, 1992) was the first to construct orthonormal bases for compactly-supported wavelets  $\psi$ . The trick was to begin with a scaling filter having only a finite number of nonzero coefficients  $\{c_k\}$  which satisfy the

above conditions. As with the Meyer wavelet construction, we define the Fourier transform  $\widehat{\phi}(\omega)$  of such a scaling function (in this case,  $\widehat{\phi}(\omega)$  is a trigonometric polynomial in frequency  $\omega$ ), and then invert it to obtain a  $\phi$  (and, by extension, a  $\psi$ ) which is compactly-supported and has excellent *time-frequency localization* properties. There are no explicit formulas for compactly-supported  $\phi$  and  $\psi$ . Instead, a fast “cascade” algorithm with arbitrarily-high precision is used to compute approximate values of the  $\phi$  and  $\psi$  functions and graph them; see Daubechies (1992, Section 6.5) for details.

Frequency localization can be viewed in terms of how the wavelet function  $\psi$  decays. If  $\psi$  decays over the high frequencies, we have a much smoother function; the smoother the function, the more rapid the decay. Decay of  $\psi$  over the low frequencies means that  $\psi$  has  $L$  vanishing moments (i.e.,  $\int x^\ell \psi(x) dx = 0$ ,  $\ell = 0, 1, 2, \dots, L$ ). Note that the Haar wavelet has  $L = 1$  vanishing moment.

Daubechies constructed two families of orthonormal wavelet bases, referred to as *extremal-phase* and *least-asymmetric* families, which combine compact support with various degrees of smoothness and numbers of vanishing moments.

*Coiflets.* (Daubechies, 1993)

## Wavelet Expansions

The wavelet series representation of  $p \in L_2(\mathfrak{R})$  is given by

$$p(x) = \sum_{k \in \mathbb{Z}} \alpha_{j_1, k} \phi_{j_1, k}(x) + \sum_{j \geq j_1} \sum_{k \in \mathbb{Z}} \beta_{jk} \psi_{jk}(x), \quad (22)$$

where the *wavelet coefficients* are defined by

$$\alpha_{j_1, k} = \int \phi_{j_1, k}(x) p(x) dx = \mathbf{E}_X \{ \phi_{j_1, k}(X) \}, \quad (23)$$

$$\beta_{jk} = \int \psi_{jk}(x) p(x) dx = \mathbf{E}_X \{ \psi_{jk}(X) \}. \quad (24)$$

The first term of (22) is the orthogonal projection of  $p$  onto the “coarsest” approximation space  $V_{j_1}$  having resolution  $j_1$ , while the second term provides details to enhance the approximation. The  $\{\alpha_{j_1, k}\}$  coefficients are often referred to as the “smooth” coefficients while the  $\{\beta_{jk}\}$  are the “detail” coefficients. The value of each coefficient indicates how much the corresponding wavelet function contributes to the expansion (22).

For example, the wavelet coefficients of the Haar basis are:

$$\alpha_{j_1, k} = 2^{j_1/2} \int_{2^{-j_1} k}^{2^{-j_1} (k+1)} p(x) dx, \quad (25)$$

$$\beta_{jk} = 2^{j/2} \left\{ \int_{2^{-j} k}^{2^{-j} (k+\frac{1}{2})} p(x) dx - \int_{2^{-j} (k+\frac{1}{2})}^{2^{-j} (k+1)} p(x) dx \right\}, \quad (26)$$

so that  $\alpha_{j_1, k}$  is  $2^{j_1/2}$  times the probability that  $X$  falls into the interval

$$[2^{-j_1}k, 2^{-j_1}(k+1))$$

and  $\beta_{jk}$  is  $2^{j/2}$  times the difference between the probability that  $X$  falls into the interval  $[2^{-j}k, 2^{-j}(k+\frac{1}{2}))$  and the probability that  $X$  falls into the adjacent interval  $[2^{-j}(k+\frac{1}{2}), 2^{-j}(k+1))$ . In this case, therefore,  $0 \leq \alpha_{j_1, k} \leq 2^{j_1/2}$  and  $-2^{j/2} \leq \beta_{jk} \leq 2^{j/2}$ .

A useful way of visualizing the wavelet representation of a density  $p$  is provided by the following analogy (Abramovich, Bailey, and Sapatinas, 2000). Suppose we wish to photograph a particular section of  $p$  using a camera equipped with a zoom lens and two filters, a *low-pass filter*,  $\phi$ , and a *high-pass filter*,  $\psi$ . First, we apply the  $\phi$ -filter with the lowest lens magnification (i.e., resolution level  $j_1$ ) and, by translating  $\phi$ , we take a panoramic sequence of photos covering the particular section of the function  $p$ . Next, we change filters to the  $\psi$ -filter, but retain the same magnification (resolution level  $j_1$ ); by translating  $\psi$ , we take a similar panoramic sequence of photos. Then, with  $\psi$  still operating as the filter, we zoom in, doubling the magnification (to resolution level  $j_1+1$ ); by translating  $\psi$  at that resolution, we take a similar panoramic sequence of photos to cover the same section of  $p$ , but now we need to take twice as many photos as we used at resolution  $j_1$ . This zooming process using  $\psi$  is repeated over and over again, each time doubling the magnification used (i.e., incrementing the resolution from  $j_1+\ell$  to  $j_1+\ell+1$ ) and, hence, doubling the number of photos taken to cover the same section of  $p$ . The entire collection of photos enables us to display the desired section of  $p$  at a whole range of different magnifications.

If we now set

$$S_{j_1}(x) = \sum_{k \in \mathbb{Z}} \alpha_{j_1, k} \phi_{j_1, k}(x), \quad D_j(x) = \sum_{k \in \mathbb{Z}} \beta_{jk} \psi_{jk}(x), \quad (27)$$

then, (22) becomes

$$p(x) = S_{j_1}(x) + \sum_{j \geq j_1} D_j(x). \quad (28)$$

So,  $S_{j_1}(x)$  corresponds to the smooth part of the density at resolution level  $j_1$  and  $D_j(x)$  corresponds to a detailed part of the density at resolution level  $j$  ( $j \geq j_1$ ).

## Wavelet Density Estimation and Thresholding

Assume we have  $n$  i.i.d. observations,  $X_1, X_2, \dots, X_n$ , from  $p$ . In some wavelet algorithms (e.g., Mallet's fast discrete wavelet transform and inverse discrete wavelet transform algorithms), it is assumed that  $n$  is a power of 2 (i.e.,  $n = 2^J$ , for some positive integer  $J$ ); if this is not the case, then modified versions of those algorithms have to be used.

We estimate the wavelet coefficients,  $\alpha_{j_1,k}$  in (23) and  $\beta_{jk}$  in (24), using the “method of moments” estimates in a similar manner as for orthogonal series estimators by the *empirical wavelet coefficients*,

$$\hat{\alpha}_{j_1,k} = n^{-1} \sum_{i=1}^n \phi_{j_1,k}(X_i), \quad \hat{\beta}_{jk} = n^{-1} \sum_{i=1}^n \psi_{jk}(X_i), \quad j \geq j_1, \quad (29)$$

respectively. Computation of  $\hat{\alpha}_{j_1,k}$  and  $\hat{\beta}_{jk}$ ,  $j \geq j_1$ , using (29) is typically slow, and more efficient ways of computing them are available (see, e.g., Herrick, Nason, and Silverman, 2001). Because the number of observations,  $n$ , is finite, we can estimate only a finite number of coefficients. Thus, these empirical coefficients are computed for resolution levels in the range  $j_1 \leq j \leq J$ , where the *finest resolution level*  $J$  is chosen large enough to expose any interesting features of the density.

For practical application, the *empirical wavelet detail coefficients*,  $\{\hat{\beta}_{jk}\}$ , are modified by “thresholding” (or “denoising”) them (Donoho, Johnstone, Kerkyacharian, and Picard, 1995). The idea is to apply the wavelet transform to the signal and then drop (or shrink towards zero) every estimated detail coefficient whose magnitude does not exceed a given size. In this way, the act of thresholding the estimated coefficients removes the effect of white noise on the coefficient estimates, while at the same time producing a parsimonious wavelet expansion. We do not threshold the smooth wavelet coefficients  $\{\alpha_{j_1,k}\}$ .

Let  $\lambda_j > 0$  denote the *threshold* for the  $j$ th detail coefficient estimate,  $\hat{\beta}_{jk}$ ,  $j \geq j_1$ . We identify two methods of thresholding. A *soft threshold* estimator,

$$\tilde{\beta}_{jk}^{\text{soft}} = \text{sign}(\hat{\beta}_{jk}) \max\{0, |\hat{\beta}_{jk}| - \lambda_j\}, \quad (30)$$

shrinks the coefficient estimate if it is greater than the threshold value to reduce the effect of noise; otherwise, it sets it equal to zero. This method is popularly known as “shrink or kill.” A *hard threshold* estimator,

$$\tilde{\beta}_{jk}^{\text{hard}} = \hat{\beta}_{jk} I_{\{|\hat{\beta}_{jk}| > \lambda_j\}}. \quad (31)$$

preserves a coefficient estimate without shrinking if it is greater than the threshold; otherwise, it sets it equal to zero. This is the “keep or kill” method. Figure 4 shows the effects on a linear function of soft and hard thresholding. The soft threshold estimator is continuous, while the hard threshold estimator is not. Hard thresholding tends to keep track of peaks and discontinuities in the density and so accentuates the variance of the density estimate; soft thresholding, on the other hand, accentuates the bias.

The probability density  $p$  can be estimated by a *linear wavelet estimator*,

$$\hat{p}_{j_1}(x) = \sum_{k \in Z} \hat{\alpha}_{j_1,k} \phi_{j_1,k}(x), \quad j_1 \in Z, \quad (32)$$

or by a *nonlinear wavelet estimator*,

$$\hat{p}(x) = \hat{S}_{j_1}(x) + \sum_{j=j_1}^J \hat{D}_j(x), \quad j_1 \in Z, \quad (33)$$

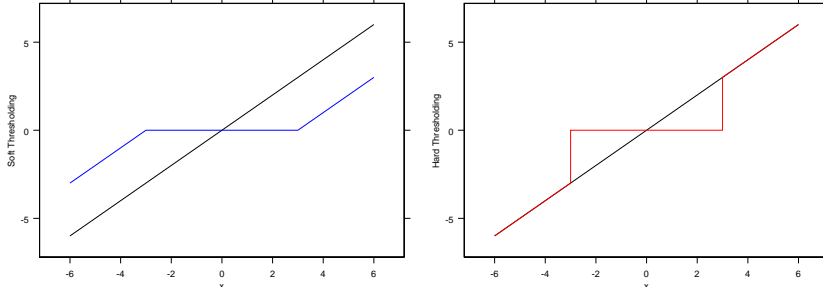


Figure 4: Comparison of soft and hard thresholding of a linear function (black line) with threshold  $\lambda = 3$ . Left panel: soft thresholding (blue line). Right panel: hard thresholding (red line).

where

$$\widehat{S}_{j_1}(x) = \sum_{k \in Z} \widehat{\alpha}_{j_1, k} \phi_{j_1, k}(x), \quad \widehat{D}_j(x) = \sum_{k \in Z} \widetilde{\beta}_{jk} \psi_{jk}(x), \quad (34)$$

and  $\widetilde{\beta}_{jk}$  refers to either a soft or hard thresholded version of  $\widehat{\beta}_{jk}$ . The truncation index  $J$  in (33) operates as a choice of fine resolution for the density estimator; it is recommended (Donoho and Johnstone, 1995) to be

$$J = J(n) \propto \left\{ n^{-1} \log_2 \left( \frac{n}{\log_e n} \right) \right\}^{1/2}. \quad (35)$$

The first term  $\widehat{S}_{j_1}(x)$  in (33) is similar in spirit to an orthogonal series density estimator, where low-resolution terms are used in the series expansion, while the second term  $\sum_{j=j_1}^J \widehat{D}_j(x)$  adds detail to  $\widehat{S}_{j_1}(x)$  based upon nonlinear shrinkage of the high-resolution wavelet coefficients.

If  $|\widehat{\beta}_{jk}|$  exceeds its hard threshold, then the corresponding term in the estimated wavelet expansion (20) is retained; otherwise, the term is dropped from the estimate. The choice of threshold is an important part of this method and a number of possible approaches to this problem have been proposed. Donoho, Johnstone, Kerkyacharian, and Picard (1995) recommend that the threshold,  $\lambda_j$ , of individual coefficients be taken to be

$$\lambda_j = \lambda_j(n) \propto \sqrt{j/n}, \quad j_1 \leq j \leq J. \quad (36)$$

The result is that the vast majority of the thresholded wavelet coefficients are reduced to zero.

Improvements in convergence results are available in which blocks of coefficients are thresholded together rather than individually, and where a simultaneous “retain-drop” decision is made on all coefficients within each block (Hall, Kerkyacharian, and Picard, 1998).

## Gaussian Approximation

Donoho, Johnstone, Kerkyacharian, and Picard (1995, 1996) suggest another method for using wavelets in density estimation. The steps in the algorithm are as follows:

1. Construct a histogram from the i.i.d. observations  $X_1, X_2, \dots, X_n$  based upon  $B$  equally-spaced bins. Denote by  $X_{(b)} = b$  the midpoint of the  $b$ th bin and by  $N_b$  the number of  $X_i$  which fall into the  $b$ th bin,  $b = 1, 2, \dots, B$ .
2. If  $X_i$  falls into the  $b$ th bin with probability  $p_b$ , then  $N_b \sim \text{Bin}(n, p_b)$ . For large  $n$  and small  $p_b$ , we can approximate the binomial  $N_b$  by a Poisson  $N_b$  with variance  $p(x)/nh_n$ , where  $h_n$  is the bin width. To remove the dependence of the variance on  $p(x)$ , we apply a square-root transformation (Anscombe, 1948) to get  $Y_b = 2(N_b + 0.375)^{1/2}$ ,  $b = 1, 2, \dots, B$ .
3. For large  $n$ , the ordered sequence of points,  $Y_b$ ,  $b = 1, 2, \dots, B$ , can now be regarded as approximately i.i.d. observations taken on a Gaussian “signal” recorded at  $B$  equispaced time points (see, e.g., Cramer, 1946, p. 250). Specifically,  $Y_b$  is modelled as  $Y_b = f_b + \epsilon_b$ , where  $f_b = f(b)$  is an unknown discrete signal and  $\epsilon_b \sim \mathcal{N}(0, \sigma^2)$  is an unobservable i.i.d. Gaussian error,  $b = 1, 2, \dots, B$ .
4. Apply the *discrete wavelet transform (DWT)* to the  $Y_1, Y_2, \dots, Y_B$ . The DWT computes the vector,  $\mathbf{w} = (w_1, w_2, \dots, w_B)^T$ , of coefficient estimates for the wavelet decomposition of the discrete “signal”  $\mathbf{f} = (f_1, f_2, \dots, f_B)^T$  by the matrix equation  $\mathbf{w} = \mathbf{W}\mathbf{f}$ , where  $\mathbf{W}$  is an  $(N \times N)$  orthogonal matrix whose elements are the wavelet basis functions evaluated at the  $B$  equally-spaced points  $1, 2, \dots, B$ . The vector  $\mathbf{w}$  contains all the wavelet coefficient estimates,  $\{\hat{\alpha}_{j_1, k}\}$  and  $\{\hat{\beta}_{jk}\}$ . The matrix multiplication is carried out using a fast “pyramid” algorithm, which, if  $B$  is a power of 2, can be completed in  $O(B)$  operations. (Note that this is faster than the  $O(B \log_e B)$ -rate of the fast Fourier transform.)
5. Shrink the estimated detail coefficients towards zero by applying soft or hard thresholding. The threshold recommended for density estimation is  $\lambda \propto 2B^{-1/2} \log_e B$ .
6. Apply the *inverse discrete wavelet transform (IDWT)*, a fast reconstruction algorithm),  $\hat{\mathbf{f}} = \mathbf{W}^T \mathbf{w}$ , to provide an estimate  $\tilde{p}$  of the density function  $p$  evaluated at the equispaced points  $1, 2, \dots, B$ . This computational step can also be completed in  $O(B)$  operations.
7. Rescale the resulting estimate  $\tilde{p}$  by transforming  $1, 2, \dots, B$  back into  $X_{(b)}$  and interpolate  $\tilde{p}$  for each bin to calculate the estimated density  $\hat{p}$ .

This Gaussian approximation method of wavelet density estimation has been implemented in the MATLAB Wavelet Toolbox.

## Two-Dimensional Wavelets

Two-dimensional wavelets are used for analyzing digitized images and for data compression.

The simplest way of constructing a 2-D orthonormal wavelet basis for  $L_2(\mathbb{R}^2)$  is to treat the two variables,  $X_1$  and  $X_2$ , as if they were independent; that is, by constructing 1-D wavelet bases for each, and then forming the 2-D function of their product. If we set the 1-D wavelet basis to be  $\psi_{jk}(x) = 2^{j/2}\psi(2^j - k)$ , then,

$$\Psi_{jk;j'k'}(x_1, x_2) = \psi_{jk}(x_1)\psi_{j'k'}(x_2) \quad (37)$$

is a 2-D wavelet and the set  $\{\Psi_{jk;j'k'}; j, j', k, k' \in Z\}$  is an orthonormal basis for  $L_2(\mathbb{R}^2)$ .

A better way of constructing a 2-D wavelet starts with two 1-D multiresolution analyses and computes their product. Define the following sequence of closed spaces,  $\mathbf{V}_j, j \in Z$ :

$$\mathbf{V}_0 = V_0 \otimes V_0 = \overline{\text{span}\{F(x, y) = f(x)g(y); f, g \in V_0\}}, \quad (38)$$

$$F(x, y) \in \mathbf{V}_j \Leftrightarrow F(2^j x, 2^j y) \in \mathbf{V}_0. \quad (39)$$

The  $\{\mathbf{V}_j, j \in Z\}$  form a multiresolution set of nested subspaces of  $L_2(\mathbb{R}^2)$  which satisfy

1.  $\{0\} \subset \cdots \subset \mathbf{V}_{-2} \subset \mathbf{V}_{-1} \subset \mathbf{V}_0 \subset \mathbf{V}_1 \subset \mathbf{V}_2 \subset \cdots \subset L_2(\mathbb{R}^2)$ ,
2.  $\bigcap_{j \in Z} \mathbf{V}_j = \{0\}$ ,
3.  $\overline{\bigcup_{j \in Z} \mathbf{V}_j} = L_2(\mathbb{R}^2)$ .

Because the  $\phi(x - k), k \in Z$ , functions form an orthonormal basis for  $V_0$ , the product functions

$$\Phi_{0;k,k'}(x, y) = \phi(x - k)\phi(y - k'), \quad k, k' \in Z, \quad (40)$$

form an orthonormal basis for  $\mathbf{V}_0$ . Similarly, the

$$\begin{aligned} \Phi_{j;k,k'} &= \phi_{jk}(x)\phi_{jk'}(y) \\ &= 2^j \Phi(2^j x - k, 2^j y - k'), \quad k, k' \in Z, \end{aligned} \quad (41)$$

form an orthonormal basis for  $\mathbf{V}_j$ .

Next, we define  $\mathbf{W}_j$ , for each  $j \in Z$ , to be the orthogonal complement of  $\mathbf{V}_j$  to  $\mathbf{V}_{j+1}$ . That is,

$$\begin{aligned} \mathbf{V}_{j+1} &= V_{j+1} \otimes V_{j+1} \\ &= (V_j \oplus W_j) \otimes (V_j \oplus W_j) \\ &= (V_j \otimes V_j) \oplus [(W_j \otimes V_j) \oplus (V_j \otimes W_j) \oplus (W_j \otimes W_j)] \\ &= \mathbf{V}_j \oplus \mathbf{W}_j. \end{aligned} \quad (42)$$

Thus,  $\mathbf{W}_j$  consists of the direct sum of three product spaces,  $W_j \otimes V_j$ ,  $V_j \otimes W_j$ , and  $W_j \otimes W_j$ , with orthonormal bases  $\psi_{jk}(x)\phi_{jk'}(y)$ ,  $\phi_{jk}(x)\psi_{jk'}(y)$ , and  $\psi_{jk}(x)\psi_{jk'}(y)$ , respectively. This yields three wavelets,

$$\begin{aligned}\Psi_{j;k,k'}^h(x,y) &= 2^j \phi(2^j x - k) \psi(2^j y - k'), \\ \Psi_{j;k,k'}^v(x,y) &= 2^j \psi(2^j x - k) \phi(2^j y - k'), \\ \Psi_{j;k,k'}^d(x,y) &= 2^j \psi(2^j x - k) \psi(2^j y - k'),\end{aligned}$$

where the superscripts  $h$ ,  $v$ , and  $d$  stand for the ‘horizontal,’ ‘vertical,’ and ‘diagonal’ orientations, respectively, present in a 2-D image. So, an orthonormal basis for  $\mathbf{W}_j$  is given by

$$\{\Psi_{j;k,k'}^\lambda; k, k' \in Z, \lambda = h, v, \text{ or } d\}, \quad (43)$$

and an orthonormal basis for  $\overline{\oplus_{j \in Z} \mathbf{W}_j} = L_2(\mathbb{R}^2)$  is given by

$$\{\Psi_{j;\mathbf{k}}^\lambda; \mathbf{k} \in Z^2, \lambda = h, v, \text{ or } d\}. \quad (44)$$

If the 1-D functions  $\phi$  and  $\psi$  have compact support, then so do  $\Phi$  and  $\Psi^h$ ,  $\Psi^v$ , and  $\Psi^d$ .

We now use this representation of 2-D wavelets to represent a smooth bivariate density function  $p(x, y) \in L_2(\mathbb{R}^2)$  as a sum of the four types of wavelets:

$$p(x, y) \sim S_{j_1}(x, y) + \sum_{j=j_1}^J D_j^h(x, y) + \sum_{j=j_1}^J D_j^v(x, y) + \sum_{j=j_1}^J D_j^d(x, y), \quad (45)$$

where

$$S_{j_1}(x, y) = \sum_{k, k'} \alpha_{j_1; k, k'} \Phi_{j_1; k, k'}(x, y), \quad (46)$$

$$D_j^h(x, y) = \sum_{k, k'} \beta_{j; k, k'}^h \Psi_{j; k, k'}^h(x, y), \quad (47)$$

$$D_j^v(x, y) = \sum_{k, k'} \beta_{j; k, k'}^v \Psi_{j; k, k'}^v(x, y), \quad (48)$$

$$D_j^d(x, y) = \sum_{k, k'} \beta_{j; k, k'}^d \Psi_{j; k, k'}^d(x, y). \quad (49)$$

In (45),  $S_{j_1}$  gives the smooth part of the density at resolution level  $j_1$ , while  $D_j^h$ ,  $D_j^v$ , and  $D_j^d$  give the horizontal, vertical, and diagonal detail parts of the density, respectively, each at the  $j$ th resolution level. The 2-D wavelet coefficients in (46)–(49) are, approximately,

$$\begin{aligned}\alpha_{j_1; k, k'} &\sim \int \int \Phi_{j_1; k, k'}(x, y) p(x, y) dx dy = \mathbb{E}_{X, Y} \{\Phi_{j_1; k, k'}(X, Y)\} \\ \beta_{j; k, k'}^h &\sim \int \int \Psi_{j; k, k'}^h(x, y) p(x, y) dx dy = \mathbb{E}_{X, Y} \{\Psi_{j; k, k'}^h(X, Y)\}\end{aligned}$$

$$\beta_{j;k,k'}^v \sim \int \int \Psi_{j;k,k'}^v(x,y)p(x,y)dxdy = E_{X,Y}\{\Psi_{j_1;k,k'}^v(X,Y)\}$$

$$\beta_{j;k,k'}^d \sim \int \int \Psi_{j;k,k'}^d(x,y)p(x,y)dxdy = E_{X,Y}\{\Psi_{j_1;k,k'}^d(X,Y)\}$$

The three detail coefficients,  $\beta^h, \beta^v$ , and  $\beta^d$ , carry horizontal, vertical, and diagonal information on the respective orientations of a 2-D image. The first term on the rhs of (45) provides resolution level- $j_1$  coefficients, while the other three terms provide the sums over resolution levels  $j_1$  to  $J$  of the three types of detail coefficients (i.e., horizontal, vertical, and diagonal detail) and basis functions.

## Bibliographical Notes

There is an enormous literature on wavelets, including many fine books. An introductory book we highly recommend is Hubbard (1996). More technical books we recommend are Daubechies (1992), Walter and Shen (2001), Vidakovic (200?). An advanced collection of articles on wavelets is Silverman and Vassilicos (2000), which includes several excellent case studies on the use of wavelets in turbulence, vision, image processing, economic and financial data, and vibrations and acoustics. An excellent book which emphasises the application of wavelets to chemometrics is Chau, Liang, Gao, and Shao (2004). There are also a number of review articles on wavelets. The ones we found to be most useful are Antoniadis (1997), Vidakovic and Mueller (1994), Liò (2003).

## References

- Abramovich, F., Bailey, T.C., and Sapatinas, T. (2000). Wavelet analysis and its statistical applications, *The Statistician*, **49**, 1–29.
- Anscombe, F. (1948). The transformation of poisson, binomial, and negative-binomial data, *Biometrika*, **35**, 246–254.
- Antoniadis, A. (1997). Wavelets in statistics: a review, unpublished report, University Joseph Fourier, Grenoble, France.
- Bruce, A. and Gao, H.-Y. (1996). *Applied Wavelet Analysis with S-Plus*, New York: Springer.
- Chau, F.T., Liang, Y.-Z., Gao, J., and Shao, Z.-G. (2004). *Chemometrics: From Basics to Wavelet Transform*, New York: Wiley.
- Cramer, H. (1946). *Mathematical Methods of Statistics*, Princeton, NJ: Princeton University Press.
- Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets, *Communications of Pure and Applied Mathematics*, **41**, 909–996.

- Daubechies, I. (1992). *Ten Lectures on Wavelets*, Philadelphia, PA: Society for Industrial and Applied Mathematics.
- Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Pickard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion), *Journal of the Royal Statistical Society, Series B*, **57**, 301–369.
- Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding, *The Annals of Statistics*, **24**, 508–539.
- Hall, P., McKay, I., and Turlach, B.A. (1996). Performance of wavelet methods for functions with many discontinuities, *The Annals of Statistics*, **24**, 2462–2476.
- Halmos, P.R. (1957). *Introduction to Hilbert Space and the Theory of Spectral Multiplicity*, New York: Chelsea Publishing Co.
- Herrick, D.R.M., Nason, G.P., and Silverman, B.W. (2001). Some new methods for wavelet density estimation, *Sankhya–Series A*, **63**, 394–411.
- Hubbard, B.B. (1996). *The World According to Wavelets: The Story of a Mathematical Technique in the Making*, Wellesley, MA: A.K. Peters, Ltd.
- Liò, P. (2003). Wavelets in bioinformatics and computational biology: state of art and perspectives, *Bioinformatics Review*, **19**, 2–9.
- Mallat, S. (1989). Multiresolution approximation and wavelets, *Transactions of the American Mathematical Society*, **315**, 69–88.
- Meyer, Y. (1985). Principe d’incertitude, bases hilbertiennes et algèbres d’opérateurs, *Seminar Bourbaki*, 1985–1986, No. 662.
- Silverman, B.W. and Vassilicos, J.C. (2000) (eds.). *Wavelets: The Key to Intermittent Information?*, Oxford University Press.
- Stromberg, J.O. (1982). A modified Franklin system and higher order spline systems on  $\mathbb{R}^n$  as unconditional bases for Hardy spaces, *Conference in Honor of A. Zygmund, Volume II* (W. Beckner et al., eds.), Wadsworth Mathematics Series, pp. 475–493.
- Vidakovic, B. (200?). *Statistical Modeling by Wavelets*. New York: Wiley.
- Vidakovic, B. and Mueller, P. (1994). Wavelets for kids: a tutorial introduction, unpublished report, Duke University.
- Walter, G.G. and Shen, X. (2001). *Wavelets and Other Orthogonal Systems*, New York: Chapman & Hall/CRC.