THE ROYAL SOCIETY

Wavelets in statistics: beyond the standard assumptions

By Bernard W. Silverman

Department of Mathematics, University of Bristol,
University Walk, Bristol BS8 1TW, UK

The original application of wavelets in statistics was to the estimation of a curve given observations of the curve plus white noise at 2^J regularly spaced points. The rationale for the use of wavelet methods in this context is reviewed briefly. Various extensions of the standard statistical methodology are discussed. These include curve estimation in the presence of correlated and non-stationary noise, the estimation of (0–1) functions, the handling of irregularly spaced data and data with heavy-tailed noise, and deformable templates in image and shape analysis. Important tools are a Bayesian approach, where a suitable prior is placed on the wavelet expansion, encapsulating the notion that most of the wavelet coefficients are zero; the use of the non-decimated, or translation-invariant, wavelet transform; and a fast algorithm for finding all the within-level covariances within the table of wavelet coefficients of a sequence with arbitrary band-limited covariance structure. Practical applications drawn from neurophysiology, meteorology and palaeopathology are presented. Finally, some directions for possible future research are outlined.

Keywords: Bayesian non-parametric modelling; deformable templates; meteorology; neurophysiology; non-parametric regression; non-stationarity

1. Introduction

(*a*) *The standard assumptions*

The early papers on the use of wavelets in statistics concentrated on the standard non-parametric regression problem of estimating a function g from observations

$$Y_i = g(t_i) + \epsilon_i, \quad i = 1, \dots, n, \tag{1.1}$$

where $n = 2^J$ for some J, $t_i = i/n$ and ϵ_i are independent identically distributed normal variables with zero mean and variance σ^2. The basic method used was the discrete wavelet transform (DWT); for convenience of reference, the notation we shall use is set out in the appendix. The DWT of a sequence x will be written $\mathcal{W}x$.

In the problem (1.1), let (d_{jk}) be the DWT of the sequence $g(t_i)$. Consider the DWT $(\eta_{jk}) = \mathcal{W}Y$ of the observed data. Since $\mathcal{W}$ is an orthogonal transform, we will have

$$\eta_{jk} = d_{jk} + \epsilon_{jk}, \tag{1.2}$$

where the ϵ_{jk} are, again, independent identically distributed normal random variables with zero mean.

On the face of it, the structure of (1.2) is the same as that of (1.1), and the DWT has not done anything to help us. However, the wavelet transform has the property that large classes of functions g likely to crop up in practice have *economical* wavelet expansions, in the sense that g is well approximated by a function, most of whose wavelet coefficients are zero. These do not just include functions that are smooth in a conventional sense, but also those that have discontinuities of value or of gradient, and those with varying frequency behaviour. This is another manifestation of the ability of wavelets to handle intermittent behaviour, and is demonstrated by mathematical results such as those discussed by Donoho *et al.* (1995).

The notion that most of the d_{jk} are zero, or may be taken to be zero, gives intuitive justification for a *thresholding rule*; if $|\eta_{jk}| \leqslant \tau$ for some threshold τ, then we set $\hat{d}_{jk} = 0$, on the understanding that η_{jk} is pure noise. For larger $|\eta_{jk}|$, the estimate is either η_{jk} itself, or a value shrunk towards zero in some way that depends only on the value of $|\eta_{jk}|$. Several papers (Donoho & Johnstone 1994, 1995, 1998, and references therein) show that estimators of this kind have excellent asymptotic adaptivity properties, and an example of the kind of result that can be derived is given in § 2 below. However, for finite sample sizes, the basic recipe of thresholding the DWT can be improved and extended.

(*b*) *Prior information and modelling uncertainty*

Before moving away from standard assumptions, we discuss a way in which the ideas behind thresholding can be developed further. The properties of wavelet expansions make it natural to model the unknown function g by placing a prior distribution on its wavelet coefficients. We focus on one particular prior model and posterior estimate (for alternative approaches, see, for example, Chipman *et al.* (1997) and Clyde *et al.* (1998)).

The approach we consider in detail is that of Abramovich *et al.* (1998), who consider a Bayesian formulation within which the wavelet coefficients are independent with

$$d_{jk} \sim (1 - \pi_j)\delta_0 + \pi_j N(0, \tau_j^2), \tag{1.3}$$

a mixture of an atom of probability at zero and a normal distribution. The mixing probability π_j and the variance of the non-zero part of the distribution are allowed to depend on the level j of the coefficient in the transform. Different choices of these hyperparameters correspond to different behaviours of the functions drawn from the prior, and, in principle, these properties can be used to choose the properties of the functions.

In practice, it is often more straightforward to have an automatic choice of the hyperparameters, and this is provided by Johnstone & Silverman (1998) who use a marginal maximum likelihood formulation. Under the prior (1.3), the marginal distribution of the wavelet coefficients η_{jk} is a mixture of a $N(0, \sigma^2)$ and a $N(0, \sigma^2 + \tau_j^2)$. The likelihood of the hyperparameters can then be maximized, most conveniently using an EM algorithm.

The posterior distribution of the individual wavelet coefficients is a mixture of an atom at zero and a general normal distribution. The traditional summary of the posterior distribution is the posterior mean, but, in this case, the posterior *median* has attractive properties. Abramovich *et al.* (1998) show that it yields a thresholding rule, in that the posterior median of d_{jk} is only non-zero if the absolute value of the corresponding coefficient η_{jk} of the data exceeds some threshold. Generally, the posterior median will have a sparse wavelet expansion, and this is in accordance with the construction of the prior. Also, the posterior median corresponds to the minimum of an L^1 loss, which is more appropriate for discontinuous and irregular functions than the squared error loss that leads to the posterior mean.

2. Correlated and non-stationary noise

One important move away from the standard assumptions is to consider data that have correlated noise. This issue was discussed in detail by Johnstone & Silverman (1997). For many smoothing methods correlated noise can present difficulties, but, in the wavelet case, the extension is straightforward.

(*a*) *Level-dependent thresholding*

Provided the noise process is stationary, one effect of correlated noise is to yield an array of wavelet coefficients with variances that depend on the level j of the transform. This leads naturally to *level-dependent* thresholding, using for each coefficient a threshold that is proportional to its standard deviation. The variances are constant within levels. Therefore, at least at the higher levels, it is possible to estimate the standard deviation separately at each level, implicitly estimating both the standard deviation of the noise and the relevant aspects of its correlation structure. The usual estimator, in the wavelet context, of the noise standard deviation is

$$(\text{median of } |\eta_{jk}| \text{ on level } j)/0.6745, \tag{2.1}$$

where the constant 0.6745 is the upper quartile of the standard normal distribution. The motivation for the estimate (2.1) is again based on the properties of wavelets: in the wavelet domain, we can assume that the signal is sparse and so only the upper few $|\eta_{jk}|$ will contain signal as well as noise.

(*b*) *Adaptivity results*

Johnstone & Silverman (1997) derive theoretical justification for the idea of using a method that uses thresholds proportional to standard deviations. Suppose that X has an n-dimensional multivariate normal distribution with mean vector θ and general variance matrix V, with $V_{ii} = \sigma_i$. Let $\hat{\theta}$ be a suitable estimator obtained by thresholding the X_i one by one, using thresholds proportional to standard deviations. Under very mild conditions, the mean square error of $\hat{\theta}$ is within a factor of $(1 + 2\log n)$ of an ideal but unattainable estimator, where the optimal choice for each θ is made of 'keeping' or 'killing' each X_i in constructing the estimate; furthermore, no other estimator based on the data can improve, in order of magnitude, on this behaviour.

In our setting, the vector X consists of the wavelet coefficients of the data. Risk calculations for the 'ideal' estimator show that, for both short- and long-range-dependent noise, the level-dependent thresholding method applied in wavelet regression gives optimally adaptive behaviour. For a wide range of smoothness classes of functions g, the estimate's behaviour is close to that of the best possible estimator for each particular smoothness class. The smoothness classes include those that allow for intermittently irregular behaviour of the function; thus, the level-dependent thresholded wavelet estimator automatically adjusts to the regularity (or intermittent irregularity) of the unknown function being estimated.

3. The non-decimated wavelet transform

(*a*) *The transform and the average basis reconstruction*

An important development in the statistical context has been the routine use of the non-decimated wavelet transform (NDWT), also called the stationary or translation-invariant wavelet transform (see, for example, Shensa 1992; Nason & Silverman 1995; Lang *et al*. 1996; Walden & Contreras Cristan 1998). Conceptually, the NDWT is obtained by modifying the Mallat DWT algorithm as described in the appendix: at each level, as illustrated in figure 7, no decimation takes place, but instead the filters $\mathcal{H}$ and $\mathcal{G}$ are repeatedly padded out with alternate zeros to double their length. The effect (roughly speaking, depending on boundary conditions) is to yield an overdetermined transform with n coefficients at each of $\log_2 n$ levels. The transform contains the standard DWT for every possible choice of time origin.

Since the NDWT is no longer (1–1), it does not have a unique inverse, but the DWT algorithm is easily modified to yield the *average basis* inverse (Coifman & Donoho 1995), which gives the average of the DWT reconstructions over all choices of time origin. Both the NDWT and the average basis reconstruction are $O(n \log n)$ algorithms.

(*b*) *Using the NDWT for curve estimation*

In virtually every case that has been investigated in detail, the performance of wavelet regression methods is improved by the use of the NDWT. The standard paradigm is to perform the NDWT, and then to process each level of the resulting coefficients using the particular method of choice, regarding the coefficients as independent. Then the average basis reconstruction method is used to give the curve estimate. This procedure corresponds to carrying out a separate smoothing for each choice of time origin and then averaging the results, but the time and storage used only increase by an $O(\log n)$ factor.

In the Bayesian case, it should be noted that the prior model for the wavelet expansion will not be exactly consistent between different positions of the origin. If a random function has independent wavelet coefficients with a mixture distribution like (1.3) with respect to one position of the origin, then it will not, in general, have such a representation if the origin is changed. The prior distributions cannot be generated from a single underlying prior model for the curve, and so, strictly speaking, the method corresponds to a separate modelling of the prior information at each position of the origin.

Johnstone & Silverman (1998) investigate the use of the NDWT in conjunction with the marginal maximum likelihood approach discussed above. Even in the case where the original noise is independent and identically distributed, the n coefficients of the NDWT of the data at each level are not actually independent. In the case of correlated noise, even the standard wavelet coefficients are not, in general, actually independent. However, good results are obtained by proceeding *as if* the NDWT coefficients are independent, maximizing an as-if-independent likelihood function to

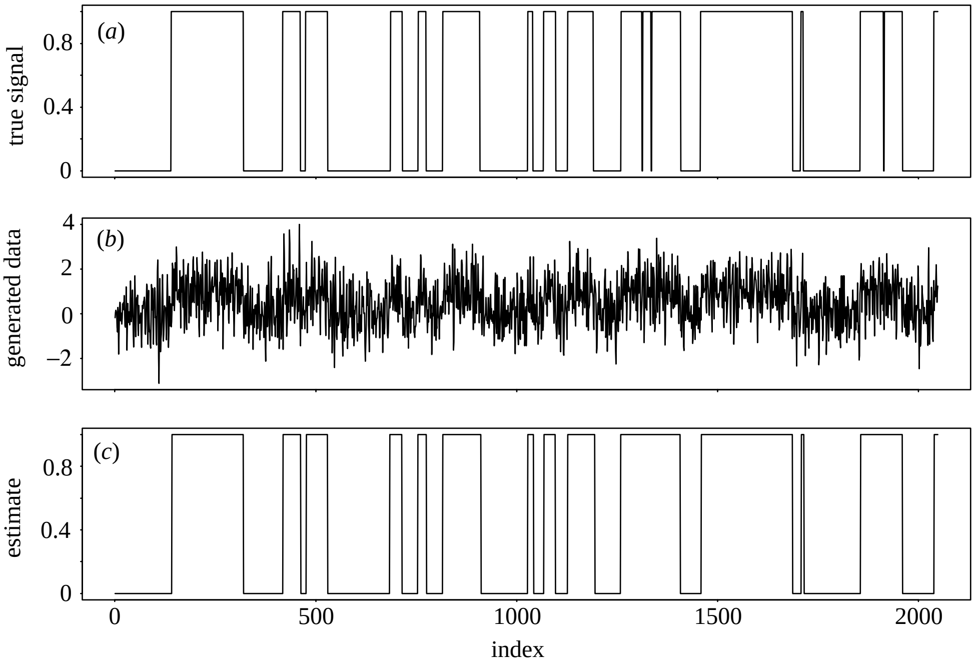

Figure 1. (*a*) The 'true' ion channel signal for time points 1–2048. (*b*) The corresponding section of generated data (on a smaller vertical scale). (*c*) The estimate obtained by the translation-invariant marginal maximum likelihood method. Reproduced from Silverman (1998) with the permission of the author.

find the mixture hyperparameters at each level, and then using an average basis reconstruction of the individual posterior medians of the wavelet coefficients. In a simulation study the improvement over the fixed basis marginal maximum likelihood method is substantial, typically around 40% in mean square error terms.

(*c*) *A neurophysiological example*

An important problem in neurophysiology is the measurement of the very small currents that pass through the single membrane channels that control movement in and out of cells. A key statistical issue is the reconstruction of a (0–1) signal from very noisy, and correlated, data. The two levels correspond to periods within which the membrane channel is closed or open.

In connection with their encouragement (Eisenberg 1994) of the application of modern signal processing techniques to this problem, Eisenberg and Levis have supplied a generated dataset intended to represent most of the relevant challenges in such single-channel data. This generated example differs in kind from the usual kind of simulated data, in that its underlying model is carefully selected by practitioners directly involved in routine collection and analysis of real data. The reason for using a generated dataset rather than an actual dataset obtained in practice is that in the case of a 'real' dataset, the 'truth' is not known, and so it is impossible to quantify the performance of any particular method.

The top two panels of figure 1 illustrate the data we have to deal with. The top graph is the generated 'true' (0–1) signal, which is used to judge the quality of

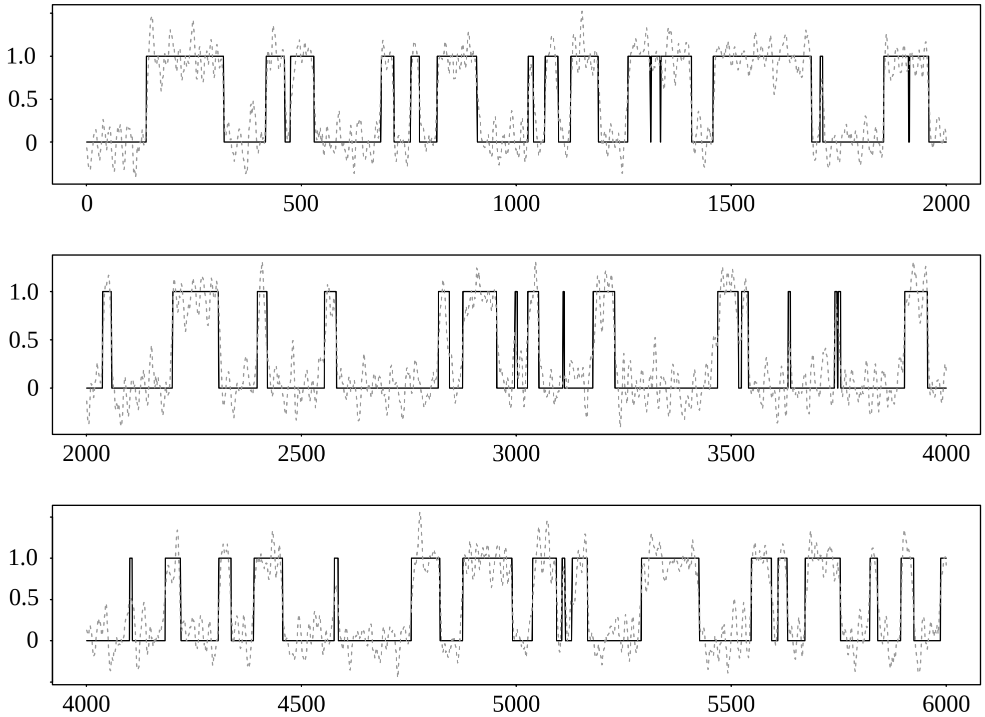

Figure 2. The true ion channel signal, and the curve obtained by applying the translation-invariant marginal maximum likelihood approach to the top three levels only. Rounding the curve off to the nearest integer gives an excellent estimate of the original signal. Reproduced from Johnstone & Silverman (1998) with the permission of the authors.

the reconstruction. Of course, the processing is done without any reference to this information. The middle panel shows the actual data. The vertical scale is much smaller than the top panel; the range of the displayed data is −3.1 to 4, and the signal to noise ratio is about $\frac{1}{3}$.

The bottom panel is obtained by using the translation-invariant marginal maximum likelihood Bayesian procedure, as described in § 3 *b* above, and then rounding off the result to the nearest of 0 and 1. Only the three highest levels of the wavelet transform are processed, and the rest are passed straight through to the rounding stage. The way in which this works out is illustrated in figure 2. The output from the wavelet smoothing step is still fairly noisy, but its variance is sufficiently small that the rounding step gives very good results.

It should be pointed out that rounding the original data gives very poor results; about 28.6% of the points are misclassified, almost as bad an error rate as setting every value to zero, which would misclassify 34.3% of points. The method illustrated in the figures, on the other hand, achieves an error rate of only 2%; for fuller details of numerical comparisons see Silverman (1998). This performance is as good as that of the special purpose method designed by Eisenberg and Levis. Close examination of figure 1 shows that the wavelet method faithfully recovers the true pattern of openings and closings, except for three very short closings, each of which is only of length two time points. The special-purpose method also misses these closings. The way in which a general purpose wavelet method can match a special purpose method that cannot be expected to work well in other contexts is particularly encouraging.

4. Dealing with irregular data

(*a*) *Coefficient-dependent thresholding*

So far we have considered the case of stationary correlated data, but the results of Johnstone & Silverman (1997) give motivation for the treatment of much more general correlation structures. Their theory gives support to the thresholding of every coefficient in the wavelet array using a threshold proportional to its standard deviation. If the original data are heteroscedastic, or have non-stationary covariance structure, then these standard deviations will not necessarily even be constant at each level of the transform.

A heteroscedastic error structure can arise in the data in several ways. For example, it may be known that the covariance structure is of a given non-stationary form, but this will not often apply in practice. Kovac & Silverman (1998) consider other possibilities. Often, the original observations are on an irregular grid of points. One can then interpolate or average locally to obtain regular gridded data in order to use a wavelet approach. Even if the noise in the original data is independent and identically distributed, the irregularity of the data grid will lead to data that are, in general, heteroscedastic and correlated in a non-stationary way.

Another context in which heteroscedastic error structures occur is in robust versions of standard regression. Even if we start with a regular grid, downweighting or eliminating outlying observations will lead to a heteroscedastic error structure or to data on an irregular grid, and this will be discussed in § 4 *c*. A final possibility considered by Kovac & Silverman (1998) is that of a number of data points that is not a power of 2. Though other methods are available, a possible approach is to interpolate to a grid of length 2^m for some m and then, as in the case of irregular data, the resulting error structure will be heteroscedastic.

(*b*) *Finding the variances of the wavelet coefficients*

In all these cases, it is important to find the variances of the wavelet coefficients of the data given the covariance matrix Σ of the original data. Kovac & Silverman (1998) provide an algorithm that yields all the variances and within-level covariances of the wavelet coefficients. If the original Σ is band limited, which it will be in the cases described above, then the algorithm will take only $O(n)$ operations, for $n = 2^J$ data points.

Using the notation for the DWT set out in the appendix, let Σ^j denote the variance matrix of c^j and $\tilde{\Sigma}^j$ that of d^j. Then $\Sigma^J = \Sigma$ by definition. From the recursive definition of the c^j and d^j it follows that, for each $j = J-1, J-2, \ldots, 0$,

$$\Sigma^j = H^{j+1}\Sigma^{j+1}(H^{j+1})^{\mathrm{T}} \tag{4.1}$$

and

$$\tilde{\Sigma}^j = G^{j+1}\Sigma^{j+1}(G^{j+1})^{\mathrm{T}}. \tag{4.2}$$

Note that this gives us not only the variances

$$\sigma_{jk} = \tilde{\Sigma}^{j}_{k,k}$$

of the individual wavelet coefficients d^j_k, but also the covariance structure of the wavelet coefficients at each level. The sparsity structure of the matrices H_{j+1} and G_{j+1} allows the calculations (4.1) and (4.2) to be carried out in $O(n_j)$ operations. Hence, the complexity of the entire algorithm, deriving the variance matrices for all j, is $O(2^J)$, and is also linear in the length of the wavelet filters and in the bandwidth of the original variance matrix (see Kovac & Silverman (1998) for details).

(*c*) *Robust estimation*

Standard wavelet regression methods do not work well in the presence of noise with outliers or with a heavy-tailed distribution. This is almost inevitable, because the methods allow for abrupt changes in the signal, and so outlying observations are likely to be interpreted as indicating such abrupt changes. Bruce *et al*. (1994) and Donoho & Yu (1998) have suggested approaches based on nonlinear multi-resolution analyses, but the algorithm of § 4 *b* allows for a simple method based on classical robustness methods. The method works whether or not the original data are equally spaced. As a first step, a wavelet estimator is obtained using the method of § 4 *a*, using a fairly high threshold. Outliers are then detected by a running median method. These are removed and the method of § 4 *a* applied again to produce the final estimate (see Kovac & Silverman (1998) for a detailed discussion).

An example is given in figure 3. The data are taken from a weather balloon and describe the radiation of the sun. The high-frequency phenomena in the estimated signal are due to outlier patches in the data; these may be caused by a rope that sometimes obscured the measuring device from the direct sunlight. It is interesting to note that the robust estimator removes the spurious high-frequency effects, but still models the abrupt change in slope in the curve near time 0.8.

5. Deformable templates

(*a*) *Images collected in palaeopathology*

We now turn to a rather different area of application, that of deformable templates in image and shape analysis. There are many problems nowadays where an observed image can be modelled as a deformed version of a standard image, or template. The assumption is that the image is a realization of the template, perhaps with additional variability that is also of interest. My own interest in this issue stems from a study of skeletons temporarily excavated from a cemetery in Humberside. Of particular interest to the palaeopathology group in the Rheumatology Department at Bristol University is the information that can be gained about patterns of osteoarthritis of the knee. Shepstone *et al*. (1999) discuss the collection and analysis of a considerable number of images of the kind shown in figure 4, using the experimental set-up shown in figure 5. Further details of the work described in this section are given by Downie *et al*. (1998).

The important features of these bones as far as the osteoarthritis study is concerned are the shape of the bone and the occurrence and position of various changes, notably eburnation (polished bone, caused by loss of cartilage) and osteophytes (bony

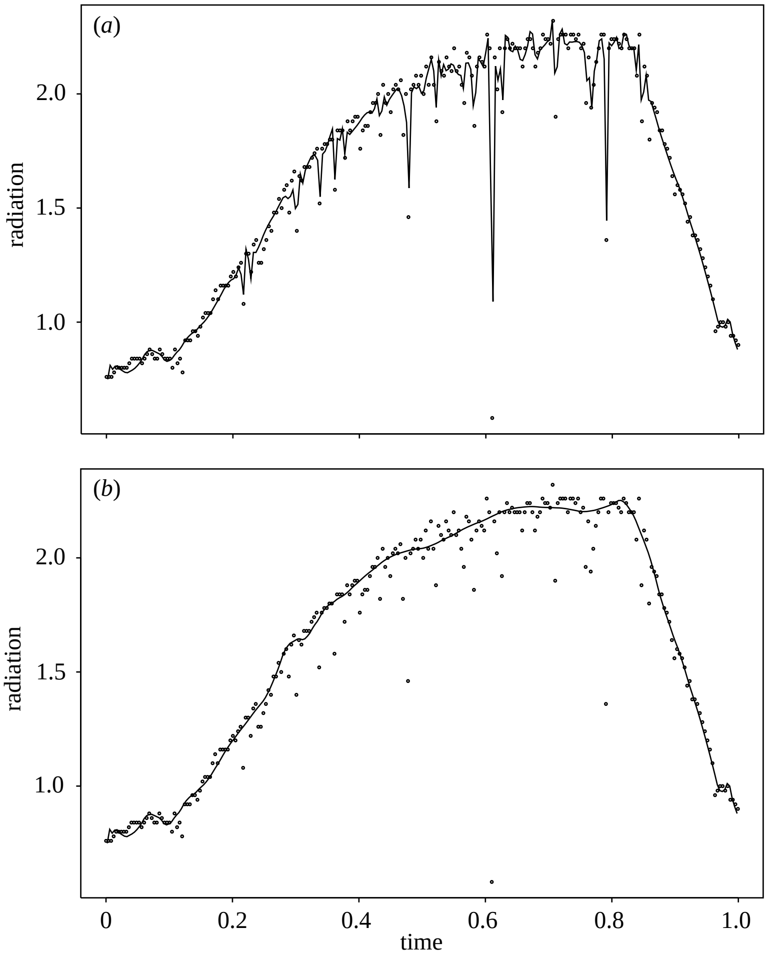

Figure 3. (*a*) Thresholding without outlier removing: balloon data with a standard wavelet regression estimator. (*b*) Thresholding after outlier removing: balloon data with a robust wavelet estimator. From Kovac & Silverman (1998) with the permission of the authors.

outgrowths). The images are digitized as pixel images and are marked up by comparison with the original bone to label the pixels corresponding to the areas of these changes. The aim of any study of the deformations is twofold: firstly, to give a standard mapping to relate positions on various bones; and secondly, to gain information about the shape of individual bones. For the first purpose, we are interested only in the *effect* of the deformation, but, for the second, the details of the deformation itself are important.

(*b*) *Models for deformations*

Deformations can be modelled as follows. Let I and T be functions on the unit square $\mathcal{U}$, representing the image and the template, respectively. In our particular application, they will be 0–1 functions. The deformation is defined by a two-dimensional *deformation function* f such that, for u in $\mathcal{U}$, $u + f(u)$ is also in $\mathcal{U}$. The

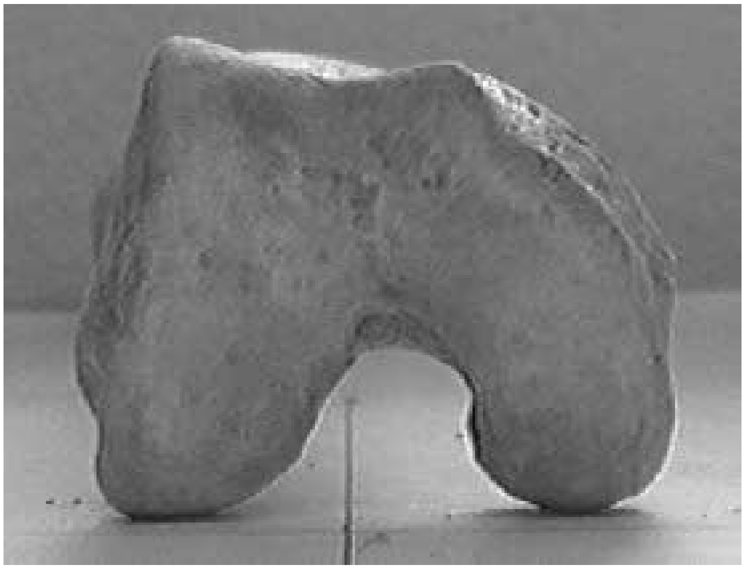

Figure 4. A typical image of the femoral condyles from the distal end of the femur. Reproduced from Downie *et al*. (1998) with the permission of the authors.

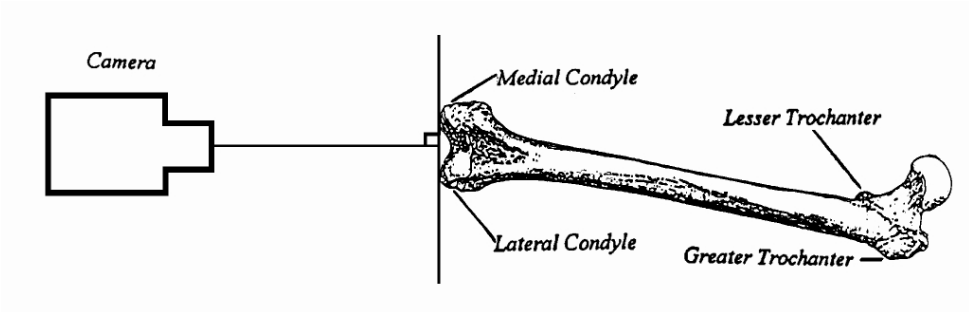

Figure 5. The experimental set-up for collecting the femur image data. Reproduced from Downie *et al*. (1998) with the permission of the authors.

aim is then to get a good fit of the image $I(u)$ to the deformed template $T(u+f(u))$ measuring discrepancy by summed squared difference over the pixels in the image.

The deformation f is a vector of two functions (f_x, f_y) on the unit square, giving the coordinates of the deformation in the x and y directions. In our work, we expand each of them as a two-dimensional wavelet series. Because it is reasonable to assume that deformations will have localized features, this may be more appropriate than the Fourier expansion used, for example, by Amit *et al*. (1991). In two dimensions the wavelet multi-resolution analysis of an array of values (Mallat 1989*b*) yields coefficients w_κ, where the index $\kappa = (j, k_1, k_2, \ell)$; this coefficient gives information about the array near position (k_1, k_2) on scale j. Three orthogonal aspects of local behaviour are modelled, indexed by ℓ in $\{1, 2, 3\}$, corresponding to horizontal, vertical and diagonal orientation.

To model the notion that the deformation has an economical wavelet expansion, a mixture prior of the kind described in § 1 *b* was used. Because the assumption of

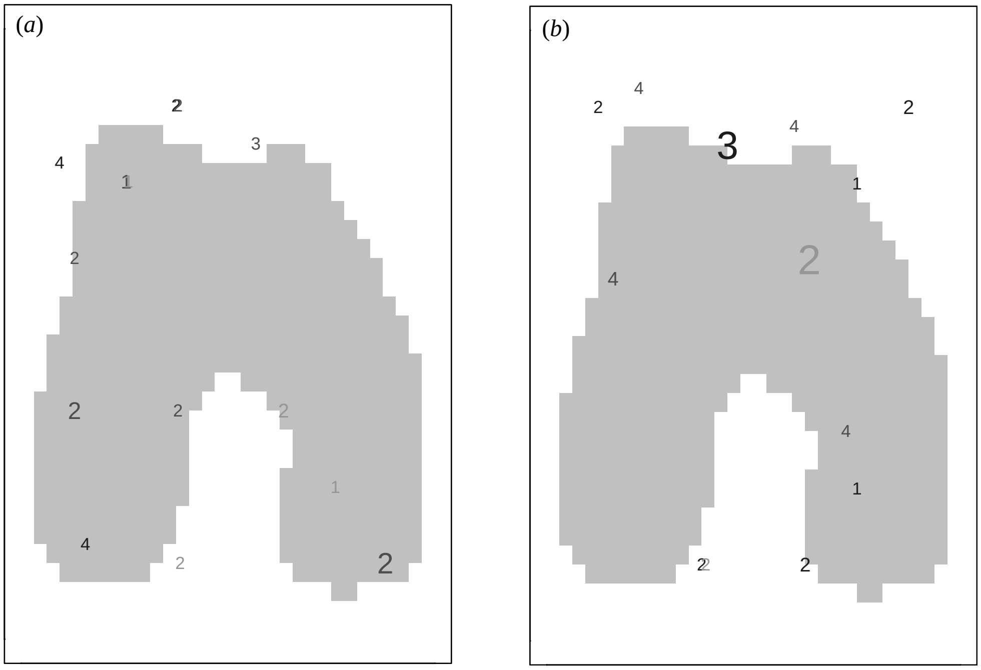

Figure 6. The wavelet coefficient positions and sizes of a typical deformation. (*a*) x wavelet coefficients (i.e. showing the x-coordinate of the deformation); (*b*) y wavelet coefficients (i.e. showing the y-coordinate of the deformation). The numbers denote the scale of the particular coefficient, and are plotted at the centre of the support of the corresponding wavelet. The printed size of each number indicates the absolute size of the wavelet coefficient. Adapted from Downie *et al.* (1998) with the permission of the authors.

normal identically distributed errors is not realistic, we prefer to consider our method as being a penalized least-squares method with a Bayesian motivation, rather than a formal Bayesian approach. A particular bone unaffected by any pathology was arbitrarily designated as the template. An iterated coordinatewise maximization is used to maximize the posterior likelihood.

(*c*) *Gaining information from the wavelet model*

Figure 6 demonstrates the information available in the wavelet expansion of a particular deformation. Only 27 of the possible 2048 coefficients are non-zero, indicating the extreme economy of representation of the deformation. For each of these coefficients, a number equal to the level j of the coefficient is plotted at the position (k_1, k_2). The size at which the number is plotted gives the absolute size of the coefficient; the orientation ℓ is indicated by colours invisible on this copy, but available on the World Wide Web version of Downie *et al.* (1998) at www.statistics.bristol.ac.uk/ ~bernard.

The figure shows that most of the non-zero coefficients are near the outline of the image, because of the localization properties of the wavelets. At the top of the image in the y component, coefficients at all resolution levels are present, indicating the presence of both broad-scale and detailed warping effects. The deformation is dominated by two coefficients, showing that the main effects are a fairly fine-scale effect at the middle of the top of the image, and a larger-scale deformation centred in the interior of the image. The full implications of this type of plot remain a subject for future research; in some contexts, the coefficients and their values will be candidates for subsequent statistical analysis, while elsewhere they will be valuable for the insight they give into the position and scale of important aspects of the deformation.

6. Discussion

Although there has been a recent explosion of interest in the use of wavelets in statistics, there are very many open questions and issues, both theoretical and practical, to be addressed before the full value of wavelets in statistics is understood and appreciated. Some other issues, not discussed here, will of course be raised in the other papers in this issue.

Even for the standard regression problem, there is still considerable progress to be made in determining the appropriate way to process the wavelet coefficients of the data to obtain the estimate. In this paper, attention has been focused on methods that treat coefficients at least as if they were independent. However, it is intuitively clear that if one coefficient in the wavelet array is non-zero, then it is more likely (in some appropriate sense) that neighbouring coefficients will be also. One way of incorporating this notion is by some form of *block thresholding*, where coefficients are considered in neighbouring blocks (see, for example, Hall *et al*. 1998; Cai & Silverman 1998). An obvious question for future consideration is how to integrate the ideas of block thresholding and related methods within the range of models and methods considered in this paper.

It is clear that a much wider class of Bayesian approaches will be useful. There are two related directions to proceed in. As in most statistical contexts where Bayesian methods are used, careful thought about priors within the DWT and NDWT contexts is needed in order to genuinely model prior knowledge of unknown functions. More particularly, the NDWT frees the modelling from the choice of origin, but one might wish to go further and move away from the powers of two in the scale domain. The *atomic decomposition* models discussed, for example, by Abramovich *et al*. (2000), may be a good starting point here.

The present paper has discussed the NDWT, but there are many extensions and generalizations of the basic DWT idea using other bases and function dictionaries to express the functions of interest. In a statistical context, some consideration has been given to the use of multiple wavelets (Downie & Silverman 1998) and ridgelets and other functions (Candès & Donoho, this issue). There is much scope for gaining a clear understanding of the contexts in which more general function dictionaries, and developments such as wavelets for irregular grids (see, for example, Daubechies *et al*., this issue) will be of statistical use. In the more traditional regression context (see, for example, Green & Silverman 1994), semi-parametric methods, which use a combination of classical linear methods and non-parametric regression, are often useful. Similarly, in the wavelet context, there may well be scope for the use of a combination of ideas from wavelets and from other regression methods, to give hybrid approaches that may combine the advantages of both.

Until now, most of the work on wavelets in statistics has concentrated on the standard regression problem. There has been some work on statistical inverse prob-

$$\begin{tikzcd}
 & \boxed{d^{J-1}} & & \boxed{d^{J-2}} & & \cdots & \boxed{d^1} & \boxed{d^0} \\
y = c^J \arrow[ur, "\mathcal{D}_0\mathcal{G}"] \arrow[rr, "\mathcal{D}_0\mathcal{H}"] & & c^{J-1} \arrow[ur, "\mathcal{D}_0\mathcal{G}"] \arrow[rr, "\mathcal{D}_0\mathcal{H}"] & & c^{J-2} & \cdots & c^1 \arrow[ur, "\mathcal{D}_0\mathcal{G}"] \arrow[r, "\mathcal{D}_0\mathcal{H}"] & \boxed{c^0}
\end{tikzcd}$$

Figure 7. The discrete wavelet transformation of a sequence y.

lems (see, for example, Abramovich & Silverman 1998) and on time-series analysis (see, for example, Walden & Contreras Cristan 1998; Nason & von Sachs, this issue). However, it is important to extend the use of wavelets to a much wider range of statistical problems. One of the major advances in statistics in recent decades has been the development and routine use of generalized linear models (see, for example, McCullagh & Nelder 1989). There has been a considerable amount of work in the application of penalized likelihood regression methods to deal non-parametrically or semi-parametrically with generalized linear model dependence (see, for example, Green & Silverman 1994, ch. 5 and 6), and it is natural to ask whether wavelet methods can make a useful contribution. One common ingredient of generalized linear model methods is iterated reweighted least squares (see, for example, Green 1984), and, in the wavelet context, any implementation of an iterated least-squares method will require algorithms like that discussed in § 4 *b* above.

Above all, the greatest need is for advances in theoretical understanding to go hand-in-hand with widespread practical application. The wide interest in wavelets demonstrated by the papers in this issue indicates that wavelets are not just an esoteric mathematical notion, but have very widespread importance in and beyond science and engineering. Of course, they are not a panacea, but as yet we have only made a small step in the process of understanding their true statistical potential.

The author gratefully acknowledges the financial support of the Engineering and Physical Sciences Research Council (UK) and the National Science Foundation (USA). The comments of Graeme Ambler, Guy Nason and Theofanis Sapatinas on previous versions were greatly appreciated.

Appendix A. The discrete wavelet transform

In order to define notation, it is useful to review the standard discrete wavelet transform algorithm of Mallat (1989*a*, *b*). For further details see any standard text on wavelets, such as Daubechies (1992) or Chui (1992). The transform is defined by linear high- and low-pass filters $\mathcal{G}$ and $\mathcal{H}$, specified by coefficient sequences (g_k) and (h_k), respectively. For any sequence x, we have, for example:

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

The coefficient sequences satisfy $g_k = (-1)^k h_{1-k}$, and have finite (and usually short) support. Denote by $\mathcal{D}_0$ the 'binary decimation' operator that chooses every even member of a sequence, so that $(\mathcal{D}_0 x)_j = x_{2j}$.

The discrete wavelet transform (DWT) of a sequence y of length 2^J will then be carried out as in the schema shown in figure 7. The vectors c^j and d^j are the *smooth* and the *detail* at level j, and are of length n_j, where $n_j \approx 2^j$, depending on the treatment of boundary conditions. We denote by H_j and G_j the $n_{j-1} \times n_j$ matrices, such that $c^{j-1} = H_j c^j$ and $d^{j-1} = G_j c^j$. The data vector y is decomposed into vectors of wavelet coefficients $d^{J-1}, d^{J-2}, \ldots, d^0, c^0$, also written as an array d_{jk} or $\mathcal{W}y$.

References

Abramovich, F. & Silverman, B. W. 1998 Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**, 115–129.

Abramovich, F., Sapatinas, T. & Silverman, B. W. 1998 Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc.* B **60**, 725–749.

Abramovich, F., Sapatinas, T. & Silverman, B. W. 2000 Stochastic atomic decompositions in a wavelet dictionary. *Prob. Theory Related Fields.* (In the press.)

Amit, Y., Grenander, U. & Piccioni, M. 1991 Structural image restoration through deformable templates. *J. Am. Statist. Ass.* **86**, 376–387.

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

Cai, T. T. & Silverman, B. W. 1998 Incorporating information on neighboring coefficients into wavelet estimation. Technical Report, Department of Mathematics, University of Bristol.

Chipman, H. A., Kolaczyk, E. D. & McCulloch, R. E. 1997 Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Ass.* **92**, 1413–1421.

Chui, C. K. 1992 *An introduction to wavelets.* Academic.

Clyde, M., Parmigiani, G. & Vidakovic, B. 1998 Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391–402.

Coifman, R. R. & Donoho, D. L. 1995 Translation-invariant denoising. In *Wavelets and statistics* (ed. A. Antoniadis & G. Oppenheim). Springer Lecture Notes in Statistics, no. 103, pp. 125–150.

Daubechies, I. 1992 *Ten lectures on wavelets.* Philadelphia, PA: SIAM.

Donoho, D. L. & Johnstone, I. M. 1994 Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425–455.

Donoho, D. L. & Johnstone, I. M. 1995 Adapting to unknown smoothness by wavelet shrinkage. *J. Am. Statist. Ass.* **90**, 1200–1224.

Donoho, D. L. & Johnstone, I. M. 1998 Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879–921.

Donoho, D. L. & Yu, T. P. Y. 1998 Nonlinear 'wavelet transforms' based on median-thresholding. Technical Report, Department of Statistics, Stanford University.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. & Picard, D. 1995 Wavelet shrinkage: asymptopia? (with discussion). *J. R. Statist. Soc.* B **57**, 301–369.

Downie, T. R. & Silverman, B. W. 1998 The discrete multiple wavelet transform and thresholding methods. *IEEE Trans. Sig. Proc.* **46**, 2558–2561.

Downie, T. R., Shepstone, L. & Silverman, B. W. 1998 Economical representation of image deformation functions using a wavelet mixture model. Technical Report, Department of Mathematics, University of Bristol.

Eisenberg, R. 1994 Biological signals that need detection: currents through single membrane channels. In *Proc. 16th Ann. Int. Conf. IEEE Engineering in Medicine and Biology Society* (ed. J. Norman & F. Sheppard), pp. 32a–33a.

Green, P. J. 1984 Iterated reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *J. R. Statist. Soc.* B **46**, 149–192.

Green, P. J. & Silverman, B. W. 1994 *Nonparametric regression and generalized linear models: a roughness penalty approach*. London: Chapman & Hall.

Hall, P., Kerkyacharian, G. & Picard, D. 1998 Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26**, 922–942.

Johnstone, I. M. & Silverman, B. W. 1997 Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc.* B **59**, 319–351.

Johnstone, I. M. & Silverman, B. W. 1998 Empirical Bayes approaches to wavelet regression. Technical Report, Department of Mathematics, University of Bristol.

Kovac, A. & Silverman, B. W. 1998 Extending the scope of wavelet regression methods by coefficient-dependent thresholding. Technical Report, Department of Mathematics, University of Bristol.

Lang, M., Guo, H., Odegard, J. E., Burrus, C. S. & Wells, R. O. 1996 Noise reduction using an undecimated discrete wavelet transform. *IEEE Sig. Proc. Lett.* **3**, 10–12.

Mallat, S. G. 1989*a* Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathrm{R})$. *Trans. Am. Math. Soc.* **315**, 69–89.

Mallat, S. G. 1989*b* A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Patt. Analysis Mach. Intell.* **11**, 674–693.

McCullagh, P. & Nelder, J. A. 1989 *Generalized linear models*, 2nd edn. London: Chapman & Hall.

Nason, G. P. & Silverman, B. W. 1995 The stationary wavelet transform and some statistical applications. In *Wavelets and statistics* (ed. A. Antoniadis & G. Oppenheim). Springer Lecture Notes in Statistics, no. 103, pp. 281–300.

Shensa, M. J. 1992 The discrete wavelet transform: wedding the à trous and Mallat algorithms. *IEEE Trans. Sig. Proc.* **40**, 2462–2482.

Shepstone, L., Rogers, J., Kirwan, J. & Silverman, B. 1999 The shape of the distal femur: a palaeopathological comparison of eburnated and non-eburnated femora. *Ann. Rheum. Dis.* **58**, 72–78.

Silverman, B. W. 1998 Wavelets in statistics: some recent developments. In *CompStat: Proc. in Computational Statistics 1998* (ed. R. Payne & P. Green), pp. 15–26. Heidelberg: Physica.

Walden, A. T. & Contreras Cristan, A. 1998 Matching pursuit by undecimated discrete wavelet transform for non-stationary time series of arbitrary length. *Statistics Computing* **8**, 205–219.