Statistica Sinica **17**(2007), 1457-1481

BIVARIATE HARD THRESHOLDING IN WAVELET FUNCTION ESTIMATION

Piotr Fryzlewicz

University of Bristol

Abstract: We propose a generic bivariate hard thresholding estimator of the discrete wavelet coefficients of a function contaminated with i.i.d. Gaussian noise. We demonstrate its good risk properties in a motivating example, and derive upper bounds for its mean-square error. Motivated by the clustering of large wavelet coefficients in real-life signals, we propose two wavelet denoising algorithms, both of which use specific instances of our bivariate estimator. The BABTE algorithm uses basis averaging, and the BITUP algorithm uses the coupling of "parents" and "children" in the wavelet coefficient tree. We prove the L_2 near-optimality of both algorithms over the usual range of Besov spaces, and demonstrate their excellent finite-sample performance. Finally, we propose a robust and effective technique for choosing the parameters of BITUP in a data-driven way.

Key words and phrases: Chi-square, discrete wavelet transform, nonparametric regression, translation-invariance, universal threshold, wavelet shrinkage.

1. Introduction

A paradigmatic problem in non-parametric regression is the estimation of a one-dimensional function $f : [0, 1] \mapsto R$ from noisy observations X_i taken on an equispaced grid:

$$X_i = f(i/n) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where the ε_i's are random variables with $E(\varepsilon_i) = 0$. Various subclasses of the problem can be identified, depending on the joint distribution of $(\varepsilon_i)_{i=1}^n$ and on the smoothness of f. In particular, substantial research effort has been and is being expended on developing denoising techniques under the assumption that $(\varepsilon_i)_{i=1}^n \sim N(\underline{0}, \sigma^2\mathbf{I})$. In this paper, we also restrict ourselves to the i.i.d. Gaussian case.

Since the seminal work of Donoho and Johnstone (1994), techniques based on wavelets have become a commonly used tool in nonparametric regression, extensively studied in the statistical literature. Many of them combine excellent finite-sample performance, linear computational complexity, and optimal (or near-optimal) asymptotic Mean-Square Error behaviour over a variety of function smoothness classes. (A general overview of wavelet methods in statistics can

be found, for example, in Vidakovic (1999).) The main idea underlying most of these techniques is that upon transforming the original regression problem (1.1) via a "multiscale" orthonormal linear transform W called the *Discrete Wavelet Transform* (DWT), we obtain the regression formulation

$$Y_{j,k} = d_{j,k} + \varepsilon_{j,k}, \quad j = 0, \ldots, \log_2 n - 1, \quad k = 1, \ldots, 2^j, \tag{1.2}$$

and $k = 1$ for $j = -1$, where j and k are (respectively) scale and location parameters, $Y_{j,k}$ are the empirical wavelet coefficients of X_i, $d_{j,k}$ are the true wavelet coefficients of $f(i/n)$ which have to be estimated, and $\varepsilon_{j,k}$ are again i.i.d. $N(0, \sigma^2)$. The sequence $d_{j,k}$ is often *sparse*, with most $d_{j,k}$'s being equal, or close, to zero, which motivates the use of simple thresholding estimators $\hat{d}_{j,k}$ which do not estimate $d_{j,k}$ by zero if and only if the corresponding empirical wavelet coefficient $Y_{j,k}$ exceeds a certain threshold in absolute value. This ensures that a large proportion of the noise $\varepsilon_{j,k}$ gets removed. The inverse DWT of the thresholded coefficients then yields an estimate $\hat{f}$ of the original signal f. Note that, due to the orthonormality of W, the following Parseval-type relation holds:

$$\frac{1}{n}\sum_{i=1}^{n} E\left\{\hat{f}(\frac{i}{n}) - f(\frac{i}{n})\right\}^2 = \frac{1}{n}\sum_{j,k} E\left(\hat{d}_{j,k} - d_{j,k}\right)^2.$$

This motivates our interest in studying the Mean-Square Error properties of the thresholding estimators $\hat{d}_{j,k}$.

In many thresholding schemes, wavelet coefficients are considered individually, i.e., given the value of the threshold, each d_{j_0,k_0} is estimated using knowledge of the corresponding Y_{j_0,k_0} only. In this paper, we refer to such threshold estimators as *univariate*. A variety of methods for selecting threshold values in univariate thresholding estimators have been shown to attain (near-)optimal Mean-Square Error convergence rates over a range of smoothness classes of f. Examples include the *minimax* and *universal* thresholds (Donoho and Johnstone (1994), thresholds based on the *False Discovery Rate* (Abramovich and Benjamini (1996), and Abramovich, Benjamini, Donoho and Johnstone (2000)), thresholds based on *Stein's unbiased risk criterion* Donoho and Johnstone (1995)) or the *empirical Bayes* procedure of Johnstone and Silverman (2005).

Motivated by the observation that in a variety of real-life signals, significant wavelet coefficients often occur at adjacent scales and locations, several authors have studied risk properties of various *multivariate* thresholding rules (often referred to as "block thresholding rules"), whereby each d_{j_0,k_0} is estimated using knowledge of not only Y_{j_0,k_0}, but also other neighbouring coefficients $Y_{j,k}$, often within the same scale $j = j_0$. Examples of such techniques include the nonoverlapping block thresholding methods of Cai (1999) and Hall, Kerkyacharian and Picard (1999), as well as the *NeighBlock* and *NeighCoeff* methods of

Cai and Silverman (2001), which use overlapping block thresholding. Other interesting wavelet thresholding algorithms in which the coefficients are considered jointly include Crouse, Nowak and Baraniuk (1998), Sendur and Selesnick (2002) and Dragotti and Vetterli (2003); however, none of them are accompanied by a theoretical risk analysis. Barber and Nason (2004) develop various thresholding rules for complex-valued wavelet coefficients, where the real and imaginary parts are considered jointly.

In this paper, we are also motivated by the clustering of significant wavelet coefficients. However, we adopt a different approach, and estimate each d_{j_0,k_0} using knowledge of not only Y_{j_0,k_0}, but also another "generic" normally distributed quantity, which we provisionally denote here by Z_{j_0,k_0}. Typically, Z_{j_0,k_0} will contain some local information, taken from the neighbourhood of Y_{j_0,k_0}. Both Y_{j_0,k_0} and Z_{j_0,k_0} are used to form a chi-squared statistic, which is then used in estimating d_{j_0,k_0}. In Section 2, we motivate this approach with a toy example which demonstrates a substantial risk reduction of the proposed "bivariate" thresholding rule, compared to the analogous univariate rule. Without specifying the exact choice of Z_{j_0,k_0}, we derive generic risk properties of the proposed estimator of d_{j_0,k_0} in Section 3. Sections 4 and 5 are devoted to specific choices of Z_{j_0,k_0}. In Section 4, we choose Z_{j_0,k_0} to be an empirical wavelet coefficient of X_i at scale j_0 and location k_0, but computed using a different wavelet family than Y_{j_0,k_0}. In Section 5, we take Z_{j_0,k_0} to be the "parent" coefficient of Y_{j_0,k_0} in the binary tree of wavelet coefficients. For both choices, we propose universal-type thresholding rules, prove their near-optimal Mean-Square Error behaviour over a range of smoothness classes (using the generic risk properties shown in Section 3), and demonstrate their very good finite-sample performance. Proofs are deferred to the Appendix.

2. Motivation and preliminaries

As in Section 1, consider the regression model (1.2), where the $Y_{j,k}$ arise as empirical wavelet coefficients of $(X_i)_{i=1}^n$, computed using a real-valued DWT. The reader is invited to think here, for example, of a DWT with periodic boundary conditions that uses Daubechies' (1992) compactly supported wavelets. Typically, such a wavelet decomposition is performed using the $O(n)$ "pyramid" algorithm of Mallat (1989). Throughout the paper, we assume that ε_i are i.i.d. $N(0,\sigma^2)$, which implies that $\varepsilon_{j,k}$ are also i.i.d. $N(0,\sigma^2)$.

For each j and k, our aim is to estimate $d_{j,k}$. As we consider each coefficient in turn, we drop the subscripts j and k to shorten notation. The univariate hard thresholding rule (Donoho and Johnstone (1994)) estimates d by $\hat{d}^U(t) = Y\,I(Y^2/\sigma^2 > t^2)$, where t is the threshold and $I(\cdot)$ is the indicator function. The choice $t = (2\log n)^{1/2}$ yields the univariate *universal* hard thresholding rule

(again, see Donoho and Johnstone (1994)). Note that if $d = 0$, then the LHS of the argument of the indicator function in $\hat{d}^U(t)$ is distributed as χ^2_1.

In *soft thresholding*, "surviving" coefficients Y are not left intact as in $\hat{d}^U(t)$, but get shrunk towards zero. Since in simulated examples, hard thresholding typically achieves much lower Mean-Square Error (MSE) than soft thresholding (especially in the translation-invariant setting; see e.g., Antoniadis, Bigot and Sapatinas (2001)), we do not consider soft thresholding in this paper. However, corresponding theoretical results could also be developed for the soft thresholding case.

Assume now that Z is a normally distributed quantity such that (Y, Z) is bivariate normal with mean $\underline{d}$ and variance-covariance matrix Σ, where $\underline{d} = (d, d')^T$,

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

We propose to estimate d by a "bivariate" hard thresholding estimator

$$\hat{d}^B(t) = Y\, I\{(Y, Z)\Sigma^{-1}(Y, Z)^T > t^2\}, \tag{2.1}$$

where t is the threshold and ^{T} denotes the transpose. Note that if $\underline{d} = \underline{0}$, then the LHS of the argument of the indicator function in (2.1) is distributed as χ^2_2.

The idea of using a χ^2_m variable (with $m \geq 2$) as a "thresholding statistic" is not new: it was used, for example, by Downie and Silverman (1998) in the context of multiwavelet nonparametric regression, by Barber and Nason (2004) in the context of real-valued nonparametric regression using complex wavelets (with $m = 2$), and by Olhede and Walden (2004) in wavelet thresholding which incorporates information from the discrete Hilbert transform of the signal (also with $m = 2$). However, our approach is different from the above in that we work in the classical real-valued wavelet context, and we derive the risk properties of $\hat{d}^B(t)$ *without* specifying the exact meaning of Z. This permits us to obtain a general mean-square risk bound for bivariate thresholding in Propositon 3.2. The latter result offers a flexible, easy-to-use device for assessing the risk of a bivariate estimator by "reducing" the problem to the univariate case. Benefits of this modular approach are demonstrated in Sections 4 and 5, where two estimators involving specific instances of the variable Z are introduced. By applying Proposition 3.2 in those two cases, we are easily able to prove the MSE near-optimality of our two estimators over a range of Besov smoothness classes. By contrast, we note that no such results were obtained for the related techniques cited above.

We now introduce notation for the mean-square risk of $\hat{d}^U(t)$ and $\hat{d}^B(t)$:

$$R^U_{d,\sigma}(t) = E\left\{\hat{d}^U(t) - d\right\}^2, \tag{2.2}$$

$$R^B_{\underline{d},\Sigma}(t) = E\left\{\hat{d}^B(t) - d\right\}^2. \tag{2.3}$$

The following toy example demonstrates the potential usefulness of our approach.

Example. We assume that $d' = d$ (the means of Z and Y are equal), $\sigma = 1$, $\rho = 0$ (Y and Z are uncorrelated). We take $t = 1, 2, 3, 4$ and plot $R^U_{d,\sigma}(t)$ and $R^B_{\underline{d},\Sigma}(t)$ (computed using numerical integration of (2.2) and (2.3)) against d. The plots are shown in Figure 2.1: irrespective of t, the risk of $\hat{d}^B(t)$ is almost always substantially lower than the risk of $\hat{d}^U(t)$, except for "small" values of d where it is slightly higher. The message here is that however one chooses the value of t for univariate thresholding, the bivariate estimator with this specific choice of Z, and the same value of t, can achieve better performance.

Obviously, in practical situations, the availability of Z such that $d' = d$ and $\rho = 0$ is not guaranteed. However, the hope is that the following observation might lead to a successful choice of Z: in a variety of real-life signals, large (small) wavelet coefficients often come in clusters. Thus, for each Y_{j_0,k_0}, the corresponding Z_{j_0,k_0} might represent, for example, a neighbouring coefficient of Y_{j_0,k_0} of a similar magnitude (so that hopefully $d' \approx d$, which would then lead to a risk reduction by the above example). This is the main idea behind the two specific instances of our estimator, described in Sections 4 and 5: by constructing Z so that it contains "local" information, drawn from roughly the same location as Y, we are able to achieve very good finite-sample performance.

Bivariate versus m-variate thresholding. We now describe our motivation for focusing on the bivariate case instead of the more general m-variate thresholding where m empirical coefficients are grouped together and a χ^2_m variable is used as a thresholding statistic. Our motivation can be summarised as follows.

1. There has been much interest in the bivariate case in the recent literature. Apart from Olhede and Walden (2004) and Barber and Nason (2004), Sendur and Selesnick (2002) study a specific case of the bivariate thresholding estimator. While all of these articles make interesting algorithmic contributions, none of them provides a mean-square risk theory for bivariate thresholding which would lead to (near-)optimal rates of convergence over a range of Besov classes. Our work fills this gap, and provides two new, well-performing algorithms; see Sections 4 and 5.
2. We have found no empirical evidence of m-variate estimators with $m > 2$ performing better than bivariate ones and, in a number of cases, their performance was found to be significantly inferior. In Section 5.2, we report the outcome of a simulation study designed to illustrate this.
3. In Section 5.3, we introduce a computational procedure for choosing the auxiliary variable Z from the data for a specific instance of our bivariate thresholding estimator. An analogous procedure for an m-variate estimator with $m > 2$ would be significantly more computationally demanding, which would

have a detrimental effect on the speed of such an m-variate wavelet estimation algorithm.

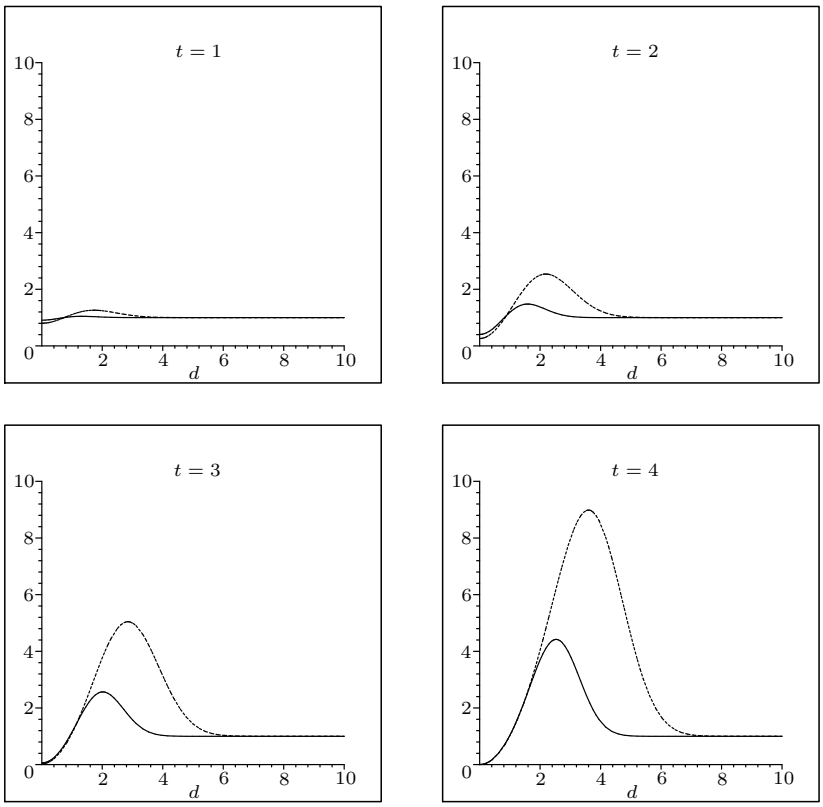

Figure 2.1. $R^B_{\underline{d},\Sigma}(t)$ (solid lines) and $R^U_{d,\sigma}(t)$ (dashed lines) for values of t as in the titles, and values of d', σ, ρ as in the Example in Section 2; plotted against d.

3. Generic risk properties of $\hat{d}^B(t)$

3.1. Upper bound for $R^B_{\underline{d},\Sigma}(t)$

As before, we drop the subscripts j, k to shorten notation. Consider a change of variable $U = (Z - \rho Y)/(1-\rho^2)^{1/2}$. Denote $d'' := E(U)$ and note that $d'' = (d' - \rho d)/(1-\rho^2)^{1/2}$, $\mathrm{cov}(Y, U) = 0$ and $\mathrm{Var}(U) = \sigma^2$. In the new coordinates, $\hat{d}^B(t) = Y\, I(Y^2 + U^2 > t^2\sigma^2)$, which leads to the representation of $R^B_{\underline{d},\Sigma}(t)$ as

$$R^B_{\underline{d},\Sigma}(t) = \sigma^2 + \iint_{y^2+u^2 \le t^2\sigma^2} \left\{ d^2 - (y-d)^2 \right\} f_{Y,U}(y,u)\,dy\,du, \tag{3.1}$$

where $f_{Y,U}$ is the joint density of (Y,U):

$$f_{Y,U}(y,u) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{(y-d)^2+(u-d'')^2}{2\sigma^2}\right\}.$$

Note that due to the particular shape and location of the integration region in (3.1), it is not possible to compute $R^B_{\underline{d},\Sigma}(t)$ exactly, or indeed to express it as a simple formula involving the pdf or cdf of the normal distribution. However, we establish simple upper bounds for $R^B_{\underline{d},\Sigma}(t)$ below. In order to do so, we introduce the following notation:

$$\delta = \left\{d^2+(d'')^2\right\}^{\frac{1}{2}},$$

$$\tilde{R}^{B,1}_{\underline{d},\Sigma}(t) = d^2\left[1-\exp\left\{-\frac{(t\sigma+\delta)^2}{2\sigma^2}\right\}\right]+\sigma^2\exp\left\{-\frac{(t\sigma-\delta)^2}{2\sigma^2}\right\}\left\{\frac{(t\sigma-\delta)^2}{2\sigma^2}+1\right\}, \quad (3.2)$$

$$\tilde{R}^{B,2}_{\underline{d},\Sigma}(t) = \frac{d^2 t\sigma}{2\delta}\left[\exp\left\{-\frac{(t\sigma-\delta)^2}{2\sigma^2}\right\}-\exp\left\{-\frac{(t\sigma+\delta)^2}{2\sigma^2}\right\}\right]+\sigma^2. \quad (3.3)$$

Denote also

$$\tilde{R}^B_{\underline{d},\Sigma}(t) = \begin{cases} \tilde{R}^{B,1}_{\underline{d},\Sigma}(t) \text{ if } \delta < t\sigma \\ \tilde{R}^{B,2}_{\underline{d},\Sigma}(t) \text{ if } \delta \geq t\sigma. \end{cases}$$

Proposition 3.1. *We have* $R^B_{\underline{d},\Sigma}(t) \leq \tilde{R}^B_{\underline{d},\Sigma}(t)$.

Using (3.1), (3.2) and (3.3), it is easy to see that both $R^B_{\underline{d},\Sigma}(t)$ and $\tilde{R}^B_{\underline{d},\Sigma}(t)$ satisfy the following conditions, which are indeed natural for the L_2 risk of a wavelet thresholding rule: $R^B_{\underline{d},\Sigma}(\infty) = d^2$, $R^B_{\underline{d},\Sigma}(0^+) = R^B_{\underline{d},\Sigma}(0) = \sigma^2$; $\tilde{R}^B_{\underline{d},\Sigma}(\infty) = d^2$, $\tilde{R}^B_{\underline{d},\Sigma}(0^+) = \tilde{R}^B_{\underline{d},\Sigma}(0) = \sigma^2$.

3.2. Comparison with the univariate risk

In order to gain a better understanding of the behaviour of the bound $\tilde{R}^B_{\underline{d},\Sigma}(t)$, we now compare it to the (better known) quantity $R^U_{\delta,\sigma}(t)$: the MSE in estimating δ from a $N(\delta,\sigma^2)$-distributed observation by means of a univariate hard thresholding estimator with threshold t. Note that it seems natural to compare $\tilde{R}^B_{\underline{d},\Sigma}(t)$ and $R^U_{\delta,\sigma}(t)$, as both functions contain exponentials of the terms $(t\sigma\pm\delta)^2/\sigma^2$. The quantity $R^U_{\delta,\sigma}(t)$ was studied in detail, for example, by Donoho and Johnstone (1994). In the following proposition, we obtain an upper bound for $\tilde{R}^B_{\underline{d},\Sigma}(t)$ in terms of $R^U_{\delta,\sigma}(t)$.

Proposition 3.2. *We have* $\tilde{R}^B_{\underline{d},\Sigma}(t) \leq \max\left(\sqrt{\frac{\pi}{2}}t+\sqrt{2\pi}, \sqrt{2\pi}t\right) R^U_{\delta,\sigma}(t)$.

Even though the Example of Section 2 suggests that $\hat{d}^B(t)$ is potentially a "better" estimator of d than $\hat{d}^U(t)$, note that Proposition 3.2 does *not* establish anything resembling "$\tilde{R}^B_{\underline{d},\Sigma}(t) \leq R^U_{d,\sigma}(t)$". Indeed, continuing the example of Section 2, it is easy to see that this inequality is not true: naturally enough, for small values of d we have $\tilde{R}^B_{\underline{d},\Sigma}(t) > R^U_{d,\sigma}(t)$.

On the other hand, it is obviously possible to establish a bound of the type $\tilde{R}^B_{\underline{d},\Sigma}(t) \leq g_{\underline{d},\Sigma}(t) R^U_{d,\sigma}(t)$ for an appropriate choice of g. However, we find that the bound in Proposition 3.2 is sufficient for our purposes, so we stick to it for simplicity.

The potential applicability of Proposition 3.2 is wide. It can be used to assess the risk of our bivariate estimator for any procedure of choosing t for which the mean-square properties of the univariate estimator are known.

The following corollary is a consequence of Proposition 3.2. It establishes a simple upper bound for the risk of our bivariate thresholding estimator with the universal threshold $t = (2 \log n)^{1/2}$. Indeed, Downie and Silverman (1998) recommend $(2 \log n)^{1/2}$ as a "universal" threshold suitable for bivariate thresholding estimators. Note that by Proposition 3.2, $\tilde{R}^B_{\underline{d},\Sigma}\{(2 \log n)^{1/2}\}$ is bounded by a quantity involving $R^U_{\delta,\sigma}\{(2 \log n)^{1/2}\}$, the latter being well-known and easy to work with.

Corollary 3.1. *For any $n \geq 4$, we have*

$$R^B_{\underline{d},\Sigma}\{(2 \log n)^{\frac{1}{2}}\} \leq \frac{q\{(\log n)^{\frac{1}{2}}\}}{1 - |\rho|} \left\{ \frac{\sigma^2}{n} + \min(d^2, \sigma^2) + \min(d'^2, \sigma^2) \right\}, \qquad (3.4)$$

where q is a cubic polynomial.

The form of inequality (3.4) is reminiscent of the "oracle inequality" of Donoho and Johnstone (1994).

4. Algorithm I: Basis Averaging using Bivariate Thresholding

In this section, we introduce one of our two wavelet denoising algorithms that involve specific instances of our bivariate hard thresholding estimator.

One way of improving the quality of wavelet-based function estimators is to compute several estimates using different wavelet families, and then to average them to obtain the final estimate. This is often referred to as basis averaging and was studied, for example, by Kohn, Marron and Yau (2000) (who, however, did not consider any bi- or multivariate estimation ideas). In this section, we propose a basis averaging algorithm which uses our bivariate hard thresholding estimator. The algorithm is called BABTE (Basis Averaging using the Bivariate Thresholding Estimator). We note that BABTE is different than the biwavelet thresholding

algorithm of Downie and Silverman (1998) in that it averages over two distinct real-valued wavelet bases and thus requires no data preprocessing. (It is known that preprocessing can hamper the practical performance of multiwavelet thresholding estimators, see e.g. the simulation study in Barber and Nason (2003).) Furthermore, unlike the above approaches, we prove the MSE near-optimality of BABTE over a range of Besov smoothness classes.

The BABTE algorithm proceeds as follows.

1. Given the regression problem (1.1), compute the DWT of X_i using two distinct Daubechies' (1992) compactly supported orthonormal wavelet bases $\psi^{(1)}$ and $\psi^{(2)}$ to obtain, respectively, the following regression problems in the wavelet domain:
$$Y_{j,k}^{(1)} = d_{j,k}^{(1)} + \varepsilon_{j,k}^{(1)}, \qquad Y_{j,k}^{(2)} = d_{j,k}^{(2)} + \varepsilon_{j,k}^{(2)},$$
for $j = 0, \ldots, \log_2 n - 1$ and $k = 1, \ldots, 2^j$ (the meaning of the symbols is as in (1.2)). Let $J = \log_2 n$. Recall that $j = 0$ ($j = J - 1$) is the coarsest (finest) detail resolution scale: the only "smooth" coefficients are indexed by $(j, k) = (-1, 1)$, so that we have $Y_{-1,1}^{(l)} = d_{-1,1}^{(l)} + \varepsilon_{-1,1}^{(l)}$ for $l = 1, 2$. The variance of $\varepsilon_{j,k}^{(l)}$ ($l = 1, 2$), denoted by σ^2, is assumed known for the theory; in practice, it is estimated via the MAD estimator on the finest resolution level $J - 1$ (see e.g., Donoho and Johnstone (1994)).
2. For each $l = 1, 2$, and each $j = J-2, J-3, \ldots, 0$, compute the *discrete wavelet vectors* $\psi_j^{(l)}$ using the formula $\{\psi_j^{(l)}\}_m = \sum_k \{\psi_{j+1}^{(l)}\}_k h_{m-2k}^{(l)}$, where $\{\psi_{J-1}^{(l)}\}_m = g_m^{(l)}$, and $h^{(l)}$, $g^{(l)}$ are low- and high-pass, respectively, quadrature mirror filters associated with $\psi^{(l)}$ (for details of this computation, see Daubechies (1992, p.204)). For each scale $j = 0, \ldots, J - 1$, compute the *inter-basis correlation* ρ_j as $\rho_j = \sum_m \{\psi_j^{(1)}\}_m \{\psi_j^{(2)}\}_m$.
3. Let $\overline{\rho} \in (0, 1)$ denote the maximum accepted correlation level, specified by the user. For each $j = 0, \ldots, J - 1$,
   - if $|\rho_j| \le \overline{\rho}$, then estimate $d_{j,k}^{(l)}$, $l = 1, 2$, by the bivariate universal thresholding estimator
   $$\hat{d}_{j,k}^{(l)}\{(2 \log\, n)^{\frac{1}{2}}\} = Y_{j,k}^{(l)} I\left\{(Y_{j,k}^{(1)}, Y_{j,k}^{(2)}) \Sigma_j^{-1} (Y_{j,k}^{(1)}, Y_{j,k}^{(2)})^T > 2 \log\, n\right\}, \quad (4.1)$$
   where
   $$\Sigma_j = \sigma^2 \begin{bmatrix} 1 & \rho_j \\ \rho_j & 1 \end{bmatrix};$$
   - otherwise, estimate $d_{j,k}^{(l)}$, $l = 1, 2$, by the classical univariate universal thresholding estimator
   $$\hat{d}_{j,k}^{(l)}\{(2 \log\, n)^{\frac{1}{2}}\} = Y_{j,k}^{(l)} I\left\{\frac{(Y_{j,k}^{(l)})^2}{\sigma^2} > 2 \log\, n\right\}.$$

In either case, leave the smooth coefficient intact: $\hat{d}^{(l)}_{-1,1} = Y^{(l)}_{-1,1}$.

4. For $l = 1, 2$, perform the inverse DWT of $\hat{d}^{(l)}_{j,k}$ to produce $\hat{f}^{(l)}(i/n)$, $i = 1, \ldots, n$. The BABTE estimate is obtained as $\hat{f} = (\hat{f}^{(1)} + \hat{f}^{(2)})/2$.

The upper bound $\overline{\rho} < 1$ for the inter-basis correlations ρ_j is required to keep the risk bound in (3.4) finite: note the $1 - |\rho|$ in the denominator. Note also that the argument of the indicator function in (4.1) does not depend on l.

The rationale behind estimating $d^{(l)}_{j,k}$ using our bivariate estimator involving $Y^{(1)}_{j,k}$ and $Y^{(2)}_{j,k}$ is as in the Example of Section 2: the hope is that the true coefficients $d^{(1)}_{j,k}$ and $d^{(2)}_{j,k}$ are both simultaneously either "large" or "small" in magnitude, since they both carry information extracted from the signal f at scale j and location k. If that is indeed the case, then we would expect a significant risk reduction of our bivariate estimator, compared to the univariate estimator with the same threshold. This heuristic observation will be confirmed by the simulation results reported in Section 4.2.

4.1. Near-optimality of the BABTE algorithm

In this section, we show that BABTE attains near-optimal MSE behaviour for a variety of signals. Note that from Figure 1, it is apparent that the bivariate thresholding estimator, which forms a basis of BABTE, does not always improve on the univariate thresholding estimator with the same threshold. Hence the near-optimality of BABTE is not a simple consequence of the near-optimality of the univariate thresholding estimator with the universal threshold. However, it can be established using the generic results obtained in Section 3.

We consider a wide range of function spaces for f, corresponding to sequence space models for its wavelet coefficients $d^{(l)}_{j,k}$, $l = 1, 2$. A flexible scale of function spaces is given by the Besov family, which is specified in the sequence space form via the wavelet coefficients of f in the following way. Let $\|d^{(l)}_j\|_p = n^{-1/2}(\sum_{k=1}^{2^j} |d^{(l)}_{j,k}|^p)^{1/p}$ and

$$b^{\nu}_{p,q}(C) = \left\{ d^{(l)}_{j,k} : \sum_{j=0}^{\infty} 2^{jsq} \|d^{(l)}_j\|^q_p \le C^q \right\},$$

where $s = \nu + 1/2 - 1/p$.

Heuristically speaking, the (not necessarily integer) parameter ν is a smoothness parameter which indicates the number of derivatives which the function f possesses in L_p, while the additional parameter q provides a further, finer gradation. The family of Besov spaces includes the Hölder and Sobolev spaces (for $p = q = \infty$ and $p = q = 2$, respectively), as well as the class of functions of

bounded variation, "sandwiched" between $b^1_{1,\infty}$ and $b^1_{1,1}$. The reader is referred to Meyer (1992) for rigorous definitions and a detailed discussion of Besov spaces.

Our quantity of interest is $\mathrm{MSE}(f,\hat{f}) = (1/n)\sum_{i=1}^n E\{f(i/n) - \hat{f}(i/n)\}^2$, where $\hat{f}$ is the BABTE estimator. It is easy to see that $\mathrm{MSE}(f,\hat{f}) \le \{\mathrm{MSE}(f,\hat{f}^{(1)}) + \mathrm{MSE}(f,\hat{f}^{(2)})\}/2$. Thus, we now focus on each $\mathrm{MSE}(f,\hat{f}^{(l)})$ individually.

Theorem 4.1. *If $0 < p,q \le \infty$ and $\nu > 1/p$, then for each $l = 1,2$,*

$$\sup_{d^{(1)}_{j,k}, d^{(2)}_{j,k} \in b^\nu_{p,q}(C)} \mathrm{MSE}(f,\hat{f}^{(l)}) = \frac{\sigma^2}{n} + \sup_{d^{(1)}_{j,k}, d^{(2)}_{j,k} \in b^\nu_{p,q}(C)} \frac{1}{n}\sum_{j=0}^{J-1}\sum_{k=1}^{2^j} E\left(\hat{d}^{(l)}_{j,k} - d^{(l)}_{j,k}\right)^2$$
$$\le C_0\, C^{\frac{2}{2\nu+1}}\, (1-\overline{\rho})^{-1}\, q\{(\log\, n)^{\frac{1}{2}}\}\, n^{-\frac{2\nu}{2\nu+1}}.$$

The rate $O(n^{-(2\nu)/(2\nu+1)})$ is the best possible MSE rate for Besov spaces, and BABTE achieves it up to a logarithmic factor, hence the term "near-optimality". BABTE is a non-linear estimator: linear estimators, such as kernel estimators, cannot attain the optimal rate of MSE convergence for $p < 2$.

4.2. Empirical performance of the BABTE algorithm

In this section, we compare the empirical performance of the BABTE algorithm to that of a simple algorithm which averages over two univariate estimators. Our test functions are the Donoho and Johnstone (1994) blocks, bumps, doppler and heavisine signals, sampled at $n = 1,024$ equispaced points and rescaled to have a unit variance. The standard deviation of the noise is $\sigma = 1/3$, so that the root signal-to-noise ratio is 3. The variance σ is not known to the algorithms and is estimated using the MAD estimator on the finest resolution level, $J - 1 = 9$.

Due to their superior performance, we only compare the *translation-invariant* (TI) versions of both algorithms, whereby the final estimator is the average of estimators obtained for all circular shifts of the data. This is common practice in wavelet regression. The fast $O(n \log n)$ implementation of the TI wavelet thresholding algorithms uses the Non-Decimated Wavelet Transform (NDWT), see e.g., Nason and Silverman (1995). For simplicity, we use periodic boundary conditions and set the primary resolution to zero.

For each signal, we chose the wavelet families $\psi^{(1)}$ and $\psi^{(2)}$ as follows: we performed an additional simulation study in which we estimated the signal in question using the TI wavelet estimator based on (univariate) universal hard thresholding, with the primary resolution set to 0. We compared the Integrated Square Error (ISE), averaged over 100 realisations, for each of the following wavelet families: Daubechies' Least Asymmetric (DLA) with $4,\ldots,10$ vanishing moments; Daubechies' Extremal Phase (DEP) with $1,\ldots,10$ vanishing moments.

We chose $\psi^{(1)}$ to be the wavelet which performed the best among the DLA family, and $\psi^{(2)}$ to be the wavelet which performed the best among the DEP family.

Having chosen the families $\psi^{(1)}$ and $\psi^{(2)}$ for each signal separately, we now compare the performance of our BABTE algorithm with $\overline{\rho} = 0.99$ (labelled BABTE-TI to emphasise the translation-invariance), and a simple algorithm which averages over two TI estimates based on univariate universal hard thresholding, obtained using $\psi^{(1)}$ and $\psi^{(2)}$ (we label the latter algorithm AVG-TI). Table 4.1 shows the results, and also indicates the families $\psi^{(1)}$ and $\psi^{(2)}$ used for each signal: notation DEP/DLA N means "the DEP/DLA wavelet with N vanishing moments".

Table 4.1. ISE averaged over 100 sample paths ($\times 10^5$ and rounded) for the competing methods based on the NDWT. The better results are boxed.

	$\psi^{(1)}$	$\psi^{(2)}$	AVG-TI	BABTE-TI
blocks	DLA 4	DEP 1	1,213	**1,121**
bumps	DLA 4	DEP 2	1,727	**1,560**
doppler	DLA 9	DEP 8	965	**814**
heavisine	DLA 8	DEP 3	505	**476**

The improvement in ISE, achieved by BABTE, ranges from 6 to 16%. This is indeed a very good result: the simple TI univariate universal hard thresholding estimator is one of the best performing competitors in the comprehensive simulation study reported in Antoniadis, Bigot and Sapatinas (2001). In the above simulated examples, it was observed that the ISE decreased as $\overline{\rho}$ increased (this is not surprising as BABTE-TI reduces to AVG-TI for $\overline{\rho} = 0$). Our recommendation is to set $\overline{\rho}$ "as close as possible" to one, but less than one, to ensure the theoretical mean-square consistency of the procedure, see Theorem 4.1.

The BABTE algorithm is fast, with the computational complexity of order n for the non-TI version, and of order $n \log n$ in the TI case. Note that the computation of the estimates in (4.1) requires little computational effort, which is yet another advantage of the proposed procedure.

5. Algorithm II: Bivariate Thresholding using Parent Coefficients

In this section, we propose a denoising algorithm based on a version of our bivariate thresholding estimator that uses "parent" wavelet coefficients: coefficients are computed at a coarser scale than their "children", but at roughly the same location. For completeness, we note that Sendur and Selesnick (2002)

proposed an algorithm which exploited the parent-child dependency and used a bivariate soft thresholding rule; however, it was not accompanied by a theoretical risk analysis. Indeed, due to the complicated form of the thresholds used by those authors, any theoretical analysis of their procedure appears challenging, if at all possible. In contrast, we propose an algorithm which is both tractable theoretically, and performs well in practice. The algorithm is called BITUP (BIvariate Thresholding Using Parents).

We note that BITUP is different from other block thresholding techniques for which a mean-square risk theory for Besov spaces exists (see the references in Section 1), in that it groups the empirical wavelet coefficients across scales, and not within the same scale. Johnstone and Silverman (2004) remark that a possible reason why within-scale block thresholding techniques often perform poorly is that for many signals, their neighbouring wavelet coefficients are only weakly related to one another. By coupling "parents and children", rather than "neighbours", BITUP is able to overcome this problem, and this is reflected in its good finite-sample performance reported in Section 5.2.

The BITUP algorithm proceeds as follows.

1. Given the regression problem (1.1), compute the DWT of X_i using a Daubechies' (1992) compactly supported orthonormal wavelet basis ψ to obtain the following regression problem in the wavelet domain: $Y_{j,k} = d_{j,k} + \varepsilon_{j,k}$, where the meaning of the symbols and the ranges of the parameters are as in the BABTE algorithm of Section 4. The variance of $\varepsilon_{j,k}$ is denoted by σ^2.
2. For each $j = 1, \ldots, J-1$, given a pre-set vector of integer shift parameters $(\Delta_j)_{j=1}^{J-1}$, estimate $d_{j,k}$ by the bivariate universal thresholding estimator

$$\hat{d}_{j,k}\{(2\log\, n)^{\frac{1}{2}}\} = Y_{j,k} I\left\{(Y_{j,k}, Y_{j-1,\lceil \frac{k}{2} \rceil + \Delta_j}) \Sigma_j^{-1} (Y_{j,k}, Y_{j-1,\lceil k/2 \rceil + \Delta_j})^T > 2\log\, n\right\}, \quad (5.1)$$

   where

$$\Sigma_j = \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

   and $Y_{j-1,l}$ is the shorthand notation for $Y_{j-1,((l-1) \bmod 2^{j-1})+1}$. Note that, given this particular form of Σ_j, formula (5.1) simplifies to

$$\hat{d}_{j,k}\{(2\log\, n)^{\frac{1}{2}}\} = Y_{j,k} I\left(Y_{j,k}^2 + Y_{j-1,\lceil k/2 \rceil + \Delta_j}^2 > 2\sigma^2 \log\, n\right).$$

3. For $j = 0$, estimate $d_{0,1}$ by the classical univariate universal hard thresholding estimator $\hat{d}_{0,1} = Y_{0,1} I(Y_{0,1}^2/\sigma^2 > 2\log\, n)$. Leave the smooth coefficient intact: $\hat{d}_{-1,1} = Y_{-1,1}$.

4. Perform the inverse DWT of $\hat{d}_{j,k}$ to produce the BITUP estimate $\hat{f}(i/n)$, $i = 1, \ldots, n$.

Note that $Y_{j-1,\lceil k/2 \rceil}$ is the "parent" of $Y_{j,k}$, that is, it is located directly above $Y_{j,k}$ in the binary tree of wavelet coefficients. The shift parameters Δ_j provide an extra degree of flexibility, making it possible to use "foster parent", instead of parent, coefficients. Also observe that in contrast to the BABTE algorithm, the correlations ρ_j do not come into play here, as the DWT is orthonormal and thus (foster) parent and child coefficients are mutually uncorrelated.

The motivation for BITUP is again as in the Example of Section 2: since both $d_{j,k}$ and $d_{j-1,\lceil k/2 \rceil + \Delta_j}$ carry information extracted from the signal at roughly the same location (provided that Δ_j is "small"), although at different scales, the hope is that they are both simultaneously either "large" or "small" in magnitude.

5.1. Near-optimality of the BITUP algorithm

We now show that the BITUP algorithm achieves near-optimal MSE behaviour for a variety of signals from Besov spaces. For the purpose of this section, we take $\mathrm{MSE}(f, \hat{f}) = (1/n) \sum_{i=1}^{n} E\{f(i/n) - \hat{f}(i/n)\}^2$, where $\hat{f}$ is the BITUP estimator.

Theorem 5.2. *Let $\hat{d}_{j,k}$ be the BITUP estimator of $d_{j,k}$. If $0 < p,\ q \leq \infty$ and $\nu > 1/p$, then*

$$\sup_{d_{j,k} \in b^{\nu}_{p,q}(C)} \mathrm{MSE}(f, \hat{f}) = \frac{\sigma^2}{n} + \sup_{d_{j,k} \in b^{\nu}_{p,q}(C)} \frac{1}{n} \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} E\left(\hat{d}_{j,k} - d_{j,k}\right)^2$$
$$\leq C_0 \, C^{\frac{2}{2\nu+1}} \, q\{(\log\, n)^{\frac{1}{2}}\} \, n^{-\frac{2\nu}{2\nu+1}}.$$

Note that we do not aim to show theoretically that the constants in the MSE rates attained by BABTE and BITUP are lower in value than those attained by any other thresholding estimators. Any comparisons of this kind would be difficult to accomplish, as the risk analysis of bivariate thresholding estimators involves integrals with respect to the bivariate Gaussian density over regions whose shapes and locations prevent explicit computation (see the discussion in Section 3.1). Thus, only "suboptimal" risk bounds are possible. Those bounds are sufficient to show the near-optimality of BABTE and BITUP; however, their suboptimality means that they cannot be used to investigate the exact magnitude of risk reduction compared to other algorithms, or whether the reduction is in any sense optimal. Instead, we take the option of demonstrating the excellent finite-sample performance of BABTE and BITUP in simulation studies, see Sections 4.2 and 5.2.

5.2. Empirical performance of the BITUP algorithm

In this section, we compare the empirical performance of BITUP to two state-of-the-art wavelet denoising technologies: the eBayes method of Johnstone and Silverman (2005), and a variety of methods based on complex-valued wavelets (Barber and Nason (2004)). EBayes was shown to outperform some earlier methods such as the classical univeral thresholding (Donoho and Johnstone (1994)), the SureShrink technique (Donoho and Johnstone (1995)), techniques based on the False Discovery Rate (FDR; Abramovich and Benjamini (1996)), the block thresholding techniques NeighBlock and NeighCoeff of Cai and Silverman (2001), as well as the QL method of Efromovich (1999). The techniques based on complex wavelets were shown to outperform the PostBlockMean procedure of Abramovich, Besbeas and Sapatinas (2002), the multiwavelet technique due to Downie and Silverman (1998), the FDR method, the cross-validation technique of Nason (1996), and, in most cases, eBayes.

Our simulation set-up and error measure are as in Section 4.2. As before, we only compare the TI versions of the algorithms, labelling them CPLX-TI (complex wavelets), EBAYES-TI (eBayes) and BITUP-TI (BITUP). We assume periodic boundary conditions and, in the BITUP-TI and EBAYES-TI procedures, set the primary resolution to zero for simplicity. The EBAYES-TI method uses the Laplace prior and the posterior median threshold. In our BITUP-TI method, we set $\Delta_j = 0$ for all j.

Table 5.2 summarises our findings. The results for the CPLX-TI are the best results (over the thresholding technique and wavelet used) quoted in Barber and Nason (2003), who use the same simulation set-up and error measure. The results for BITUP-TI and EBAYES-TI are based on 100 simulated sample paths and are optimised over the wavelet used (DEP $1, \ldots, 10$ and DLA $4, \ldots, 10$). The wavelets ψ which achieved the lowest ISE were, incidentally, the same for BITUP-TI and EBAYES-TI, and are also indicated in Table 5.2.

Table 5.2. ISE averaged over 100 sample paths ($\times 10^5$ and rounded) for the competing methods based on the NDWT. The best results are boxed.

	CPLX-TI	EBAYES-TI	BITUP-TI	ψ
blocks	1,727	932	**918**	DEP 1
bumps	1,603	1,956	**1,596**	DEP 2
doppler	**710**	1,021	1,033	DLA 9
heavisine	470	458	**409**	DLA 8

Our BITUP-TI technique outperformed the competitors, with the exception of the doppler signal. In the next section, we identify possible reasons for the weaker performance for the doppler signal and propose a remedy.

The poor performance of CPLX-TI on blocks was due to the fact that blocks is a piecewise constant signal, and we used the piecewise constant DEP 1 (Haar) wavelet to estimate it with EBAYES-TI and BITUP-TI, whereas the degree of smoothness of the "least smooth" complex-valued Daubechies' wavelet is the same as that of DEP/DLA 3.

While in practice the "optimal" analysing wavelet for the signal at hand is obviously unknown, a suitable wavelet can be chosen, for example, via the fast cross-validation algorithm of Nason (2002).

BITUP versus trivariate thresholding. As mentioned in Section 2, we found no empirical evidence of m-variate thresholding estimators, $m > 2$, performing better than bivariate ones and, in a number of cases, their performance was found to be significantly worse. To illustrate this, we assess the performance of a trivariate estimator constructed in an analogous way to BITUP, except involving both "parents" and "grandparents". We tested two TI versions of this trivariate estimator: one with the classical universal threshold $t = (2\log n)^{1/2}$ also used in BITUP-TI (TRITUP-TI-UNIV), and the other with the threshold of the form $t = (2\log n + \log\log n)^{1/2}$, recommended by Downie and Silverman (1998) as a "universal" threshold suitable for trivariate thresholding (TRITUP-TI-DOWSIL). The simulation set-up was identical as above. TRITUP-TI-DOWSIL was found to perform better than TRITUP-TI-UNIV, but BITUP-TI outperformed TRITUP-TI-DOWSIL by 21% to 36% percent, depending on the signal. The fact that "grandparents" are not helpful here might indicate that they do not provide much extra information apart from that already provided by the "parents".

5.3. Choosing Δ_j from the data

Previously, we used $\Delta_j = 0$, which did not lead to satisfactory results for the doppler signal. We now identify possible reasons for this, and suggest a remedy.

The doppler signal ($n = 1,024$) and its wavelet coefficients computed using the DLA 9 wavelet, are plotted in Figure 5.2. The example of Section 2 suggests that the successful performance of the BITUP estimator with $\Delta_j = 0$ would rely on the "child" and "parent" coefficients in the signal under consideration being simultaneously "large", at least in some of the cases. However, a careful look at the wavelet coefficients of the doppler signal reveals that it is not the case here: the groups of large coefficients at any two adjacent scales do not lie along a vertical line but the group at the coarser scale is translated to the right. Therefore, setting $\Delta_j > 0$ might be more appropriate in this case.

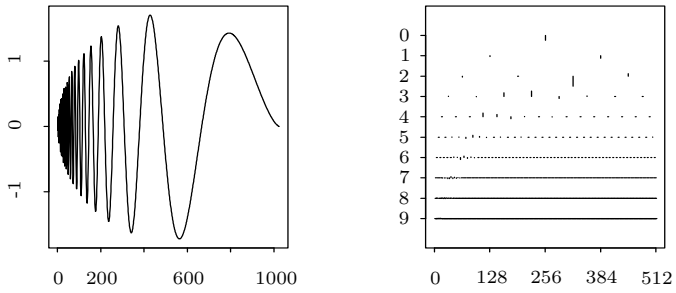

Figure 5.2. Left plot: the doppler signal. Right plot: its DLA 9 wavelet coefficients at scales $j = 0, \ldots, 9$ (the y-axis) plotted for each $k = 1, \ldots, 2^j$ (the x-axis).

In practice, a suitable value of Δ_j for $j = 1, \ldots, J-1$ might be found by aligning the sequences $(Y_{j,k})_k$ and $(Y_{j-1,k})_k$ using the following algorithm:

1. Upsample $(Y_{j-1,k})_{k=1}^{2^{j-1}}$ to create $Y^*_{j-1,k}$ such that $Y^*_{j-1,k} = Y_{j-1,\lceil k/2 \rceil}$ for $k = 1, \ldots, 2^j$.
2. For $c = -M, \ldots, M$, take a circular shift of $Y^*_{j-1,k}$ by $2c$ (denoted here by $Y^*_{j-1,k+2c}$ with a slight abuse of notation), and measure the distance between $|Y_{j,k}|$ and $|Y^*_{j-1,k+2c}|$.
3. Choose Δ_j to be the shift c which minimises the distance between $|Y_{j,k}|$ and $|Y^*_{j-1,k+2c}|$.

The rationale behind this algorithm is that by minimising the distance between $|Y_{j,k}|$ and $|Y^*_{j-1,k+2c}|$, we are forcing the large coefficients in $Y_{j,k}$ to lie directly underneath the large coefficients in $Y^*_{j-1,k+2\Delta_j}$. We then proceed with our bivariate thresholding estimator by coupling $Y_{j,k}$ and $Y_{j-1,\lceil k/2 \rceil + \Delta_j}$, rather than $Y_{j,k}$ and $Y_{j-1,\lceil k/2 \rceil}$, for the estimation of $d_{j,k}$.

Table 5.3. ISE averaged over 100 sample paths ($\times 10^5$ and rounded) for BITUP-TI and BITUP-TI with $M = 3$.

	BITUP-TI	BITUP-TI $M = 3$
blocks	918	909
bumps	1,596	1,602
doppler	1,033	766
heavisine	409	438

Table 5.3 shows the simulation results for our BITUP-TI algorithm combined with the above method for choosing Δ_j in a data-driven way, with the maximum number of shifts M set to 3 and the l_2 distance used in step 2. As expected, the ISE result for doppler is now significantly improved, while the results for the other functions are almost unaffected. This provides evidence for the effectiveness and

robustness of the algorithm. The overall algorithm is fast, and its computational complexity is $O(Mn \log n)$ (or $O(Mn)$ for the non-TI version). As in the case of BABTE, the computation of thresholds and thresholding statistics in BITUP is extremely rapid. Both BITUP and BABTE are easy to code in any package which implements the DWT, e.g., the *WaveThresh* package for the *R* environment.

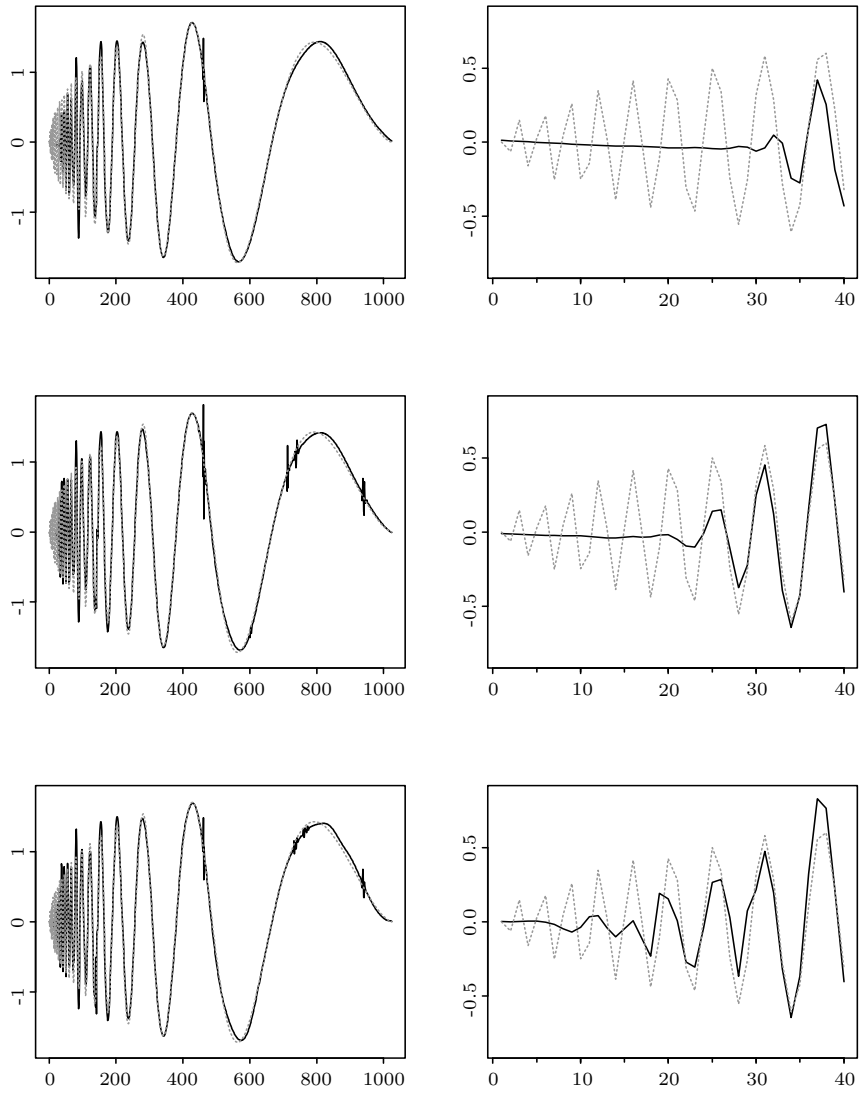

Figure 5.3. Results for the doppler signal. In all plots, solid lines are the estimates, and dotted lines are the doppler signal. Left column: the complete plots; right column: the first 40 observations. From top to bottom: results for UNIV-TI, NC-TI and BITUP-TI with $M = 3$.

We conclude with a noise removal example involving BITUP-TI with $M = 3$, TI univariate thresholding with the universal threshold (UNIV-TI), and the TI version of NeighCoeff by Cai and Silverman (2001) labelled NC-TI. We apply all three methods to the doppler signal, with the standard deviation of the noise $\sigma = 1/3$. The aim of this study is to illustrate how three different methods, each of which uses the universal threshold, is based on a single wavelet basis, and combines neighbouring coefficients (except UNIV-TI), compare on a "difficult" signal. All of the above use the DLA 9 wavelet and hard thresholding; results for soft thresholding were significantly worse and we do not report them.

The outcome is illustrated in Figure 5.3. Note that BITUP-TI does an excellent job in estimating the initial part of the signal. NC-TI does not perform so well, and UNIV-TI performs poorly. This is not surprising as the initial part of the doppler signal is highly structured across scales, see Figure 2, but only BITUP-TI takes advantage of this feature. On the other hand, both BITUP-TI and NC-TI display spurious blips in the final parts of the respective estimates. Again, this is not surprising as both use thresholds which are, effectively, lower than that used by UNIV-TI. Hoever, the spurious spikes in BITUP-TI are less pronounced than those in NC-TI.

Software. S-Plus code implementing BABTE and BITUP can be obtained from http://www.maths.bris.ac.uk/ mapzf/biv/biv.html.

Acknowledgement

I am grateful to the co-editors, an associate editor and a referee for their helpful comments which have improved the presentation of the results. I also thank the Department of Mathematics, Imperial College London, UK, where most of the work for the current version of this paper was done, for the excellent research environment.

Appendix 1. Proofs

Proof of Proposition 3.1. First consider the case $\delta < t\sigma$. Note the inequalities

$$\begin{aligned}
&\iint_{y^2+u^2\le t^2\sigma^2} f_{Y,U}(y,u)\,dy\,du \\
&\quad\le \frac{1}{2\pi\sigma^2}\iint_{(y-d)^2+(u-d'')^2\le(t\sigma+\delta)^2} \exp\left\{-\frac{(y-d)^2+(u-d'')^2}{2\sigma^2}\right\} dy\,du \\
&\quad= \frac{1}{\sigma^2}\int_0^{t\sigma+\delta} r\exp\left(-\frac{r^2}{2\sigma^2}\right) dr = 1-\exp\left\{-\frac{(t\sigma+\delta)^2}{2\sigma^2}\right\},
\end{aligned}$$

$$\begin{aligned}
&\iint_{y^2+u^2\le t^2\sigma^2} (y-d)^2 f_{Y,U}(y,u)dydu \\
&\quad\ge \frac{1}{2\pi\sigma^2}\iint_{(y-d)^2+(u-d'')^2\le(t\sigma-\delta)^2} (y-d)^2 \exp\left\{-\frac{(y-d)^2+(u-d'')^2}{2\sigma^2}\right\} dydu \\
&\quad= \frac{1}{2\pi\sigma^2}\int_0^{2\pi}\cos^2(\theta)d\theta\int_0^{t\sigma-\delta} r^3\exp\left(-\frac{r^2}{2\sigma^2}\right)dr \\
&\quad= \sigma^2\Big[1-\exp\Big\{-\frac{(t\sigma-\delta)^2}{2\sigma^2}\Big\}\Big\{\frac{(t\sigma-\delta)^2}{2\sigma^2}+1\Big\}\Big].
\end{aligned}$$

Using the above bounds in (3.1), we obtain $R^B_{\underline{d},\Sigma}(t)\le\tilde{R}^{B,1}_{\underline{d},\Sigma}(t)$ as required. Now consider the case $\delta\ge t\sigma$. Let D denote the smallest angular section of the smallest ring centred at (d,d'') containing $\{(y,u): y^2+u^2\le t^2\sigma^2\}$. We have

$$\begin{aligned}
\iint_{y^2+u^2\le t^2\sigma^2} f_{Y,U}(y,u)dydu &\le \frac{1}{2\pi\sigma^2}\iint_D \exp\left\{-\frac{(y-d)^2+(u-d'')^2}{2\sigma^2}\right\}dydu \\
&= \frac{1}{\pi\sigma^2}\arcsin\left(\frac{t\sigma}{\delta}\right)\int_{\delta-t\sigma}^{\delta+t\sigma} r\exp\left(-\frac{r^2}{2\sigma^2}\right)dr \\
&\le \frac{t\sigma}{2\delta}\left[\exp\left\{-\frac{(t\sigma-\delta)^2}{2\sigma^2}\right\}-\exp\left\{-\frac{(t\sigma+\delta)^2}{2\sigma^2}\right\}\right].
\end{aligned}$$

On the other hand, obviously $\iint_{y^2+u^2\le t^2\sigma^2}(y-d)^2 f_{Y,U}(y,u)dydu\ge 0$. Again using these two bounds in (3.1), we get $R^B_{\underline{d},\Sigma}(t)\le\tilde{R}^{B,2}_{\underline{d},\Sigma}(t)$ as required.

Proof of Proposition 3.2. Let $\phi_{\mu,\sigma}(x)=\phi\{(x-\mu)/\sigma\}$ and $\Phi_{\mu,\sigma}(x)=\Phi\{(x-\mu)/\sigma\}$, where ϕ and Φ are the pdf and the cdf of the standard normal, respectively. By straight intergration, we have

$$\begin{aligned}
R^U_{\delta,\sigma}(t) &= \delta^2\{\Phi_{\delta,\sigma}(t\sigma)-\Phi_{\delta,\sigma}(-t\sigma)\} \\
&\quad+\sigma^2\Big\{(t\sigma-\delta)\,\phi_{\delta,\sigma}(t\sigma)+(t\sigma+\delta)\,\phi_{\delta,\sigma}(-t\sigma) \\
&\quad+1-\Phi_{\delta,\sigma}(t\sigma)+\Phi_{\delta,\sigma}(-t\sigma)\Big\}. \qquad \text{(A.1)}
\end{aligned}$$

Consider the case $\delta<t\sigma$ and compare $R^U_{\delta,\sigma}(t)$ and $\tilde{R}^{B,1}_{\underline{d},\Sigma}(t)$. First examine

$$\frac{\sigma^2\{(t\sigma-\delta)\,\phi_{\delta,\sigma}(t\sigma)+1-\Phi_{\delta,\sigma}(t\sigma)\}}{\sigma^2\exp\left\{-\frac{(t\sigma-\delta)^2}{2\sigma^2}\right\}\left\{\frac{(t\sigma-\delta)^2}{2\sigma^2}+1\right\}} \qquad \text{(A.2)}$$

(note that the numerator is part of the second summand in (A.1), whereas the denominator is the second summand in (3.2)). With $x=(t\sigma-\delta)/\sigma$, (A.2)

becomes

$$\frac{x}{\sqrt{2\pi}\left(\frac{x^2}{2}+1\right)}+\frac{1-\Phi(x)}{\exp\left(-\frac{x^2}{2}\right)\left(\frac{x^2}{2}+1\right)}. \tag{A.3}$$

We now use the following result from Ito and McKean (1974, p.17) for all x,

$$1-\Phi(x)\geq\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{x^2+4}+x}\exp\left(-\frac{x^2}{2}\right). \tag{A.4}$$

Using (A.4), we bound (A.3) from below by

$$\frac{x\left(\sqrt{x^2+4}+x\right)+2}{\sqrt{2\pi}\left(\frac{x^2}{2}+1\right)\left(\sqrt{x^2+4}+x\right)}\geq\frac{1}{\sqrt{\frac{\pi}{2}}x+\sqrt{2\pi}}\geq\frac{1}{\sqrt{\frac{\pi}{2}}t+\sqrt{2\pi}},$$

which leads to

$$\begin{aligned}&\sigma^2\exp\left\{-\frac{(t\sigma-\delta)^2}{2\sigma^2}\right\}\left\{\frac{(t\sigma-\delta)^2}{2\sigma^2}+1\right\}\\&\quad\leq\left(\sqrt{\frac{\pi}{2}}t+\sqrt{2\pi}\right)\sigma^2\left\{(t\sigma-\delta)\,\phi_{\delta,\sigma}\,(t\sigma)+1-\Phi_{\delta,\sigma}\,(t\sigma)\right\}.\end{aligned} \tag{A.5}$$

We now consider the ratio

$$\frac{\delta^2\left\{\Phi_{\delta,\sigma}\,(t\sigma)-\Phi_{\delta,\sigma}\,(-t\sigma)\right\}}{d^2\left[1-\exp\left\{-\frac{(t\sigma+\delta)^2}{2\sigma^2}\right\}\right]}. \tag{A.6}$$

(note that the numerator is the first summand in (A.1), whereas the denominator is the first summand in (3.2)). With $y=-(t\sigma+\delta)/\sigma$, and with x as above, we bound (A.6) from below by

$$\frac{\delta^2\left\{\Phi(x)-\Phi(y)\right\}}{d^2\left\{1-\exp\left(-\frac{y^2}{2}\right)\right\}}\geq\frac{\delta^2\left\{1/2-\Phi(y)\right\}}{d^2\left\{1-\exp\left(-\frac{y^2}{2}\right)\right\}}\geq\frac{\delta^2}{2d^2}, \tag{A.7}$$

where the last inequality comes from the fact that $h(y):=\exp(-y^2/2)-2\Phi(y)\geq 0$ for $y\leq 0$ (note that $h(-\infty)=h(0)=0$ and differentiate twice to see that $h(y)$ has one maximum but no minima on $(-\infty,0)$). (A.7) implies that

$$d^2\left[1-\exp\left\{-\frac{(t\sigma+\delta)^2}{2\sigma^2}\right\}\right]\leq\frac{2d^2}{\delta^2}\delta^2\left\{\Phi_{\delta,\sigma}\,(t\sigma)-\Phi_{\delta,\sigma}\,(-t\sigma)\right\}. \tag{A.8}$$

By adding up (A.5) and (A.8), we obtain

$$\tilde{R}^{B,1}_{\underline{d},\Sigma}(t)\leq\max\left(\sqrt{\frac{\pi}{2}}t+\sqrt{2\pi},\frac{2d^2}{\delta^2}\right)R^U_{\delta,\sigma}(t)=\left(\sqrt{\frac{\pi}{2}}t+\sqrt{2\pi}\right)R^U_{\delta,\sigma}(t). \tag{A.9}$$

Now consider the case $\delta \geq t\sigma$ and compare $R^U_{\delta,\sigma}(t)$ and $\tilde{R}^{B,2}_{\underline{d},\Sigma}(t)$. We first examine the ratio

$$\frac{\frac{d^2 t\sigma}{2\delta}\left[\exp\left\{-\frac{(t\sigma-\delta)^2}{2\sigma^2}\right\}-\exp\left\{-\frac{(t\sigma+\delta)^2}{2\sigma^2}\right\}\right]}{\delta^2\left\{\Phi_{\delta,\sigma}\left(t\sigma\right)-\Phi_{\delta,\sigma}\left(-t\sigma\right)\right\}}$$

(note that the numerator is the first summand in (3.3), whereas the denominator is the first summand in (A.1)). With x and y as before, the extended Mean Value Theorem implies that there exists $\omega \in (y, x)$ such that

$$\frac{d^2 t\sigma\left\{\exp(-\frac{x^2}{2})-\exp(-\frac{y^2}{2})\right\}}{2\delta^3\left\{\Phi(x)-\Phi(y)\right\}}=\frac{-\sqrt{2\pi}d^2 t\sigma\omega\exp(-\frac{\omega^2}{2})}{2\delta^3\exp(-\frac{\omega^2}{2})}\leq\frac{d^2}{\delta^2}\sqrt{2\pi}t\leq\sqrt{2\pi}t, \quad \text{(A.10)}$$

where the last but one inequality follows from $-\omega \leq -y$ and $\delta \geq t\sigma$. (A.10) implies that

$$\frac{d^2 t\sigma}{2\delta}\left[\exp\left\{-\frac{(t\sigma-\delta)^2}{2\sigma^2}\right\}-\exp\left\{-\frac{(t\sigma+\delta)^2}{2\sigma^2}\right\}\right]\leq\sqrt{2\pi}t\delta^2\left\{\Phi_{\delta,\sigma}\left(t\sigma\right)-\Phi_{\delta,\sigma}\left(-t\sigma\right)\right\}. \quad \text{(A.11)}$$

Finally, consider the ratio

$$\frac{\sigma^2}{\sigma^2\left\{\left(t\sigma-\delta\right)\phi_{\delta,\sigma}\left(t\sigma\right)+1-\Phi_{\delta,\sigma}\left(t\sigma\right)\right\}}$$

(note that the numerator is the second summand in (3.3), and the denominator is part of the second summand in (A.1)). Since $u(x) := (2\pi)^{-1/2}x\exp(-x^2/2) + 1 - \Phi(x) \geq 1/2$ for $x \leq 0$ (as $u(0) = 1/2$ and $u'(x) < 0$), it follows that

$$\sigma^2 \leq 2\sigma^2\left\{\left(t\sigma-\delta\right)\phi_{\delta,\sigma}\left(t\sigma\right)+1-\Phi_{\delta,\sigma}\left(t\sigma\right)\right\}. \quad \text{(A.12)}$$

By adding up (A.11) and (A.12), we obtain

$$\tilde{R}^{B,2}_{\underline{d},\Sigma}(t) \leq \max\left(\sqrt{2\pi}t, 2\right) R^U_{\delta,\sigma}(t). \quad \text{(A.13)}$$

Combining (A.9) and (A.13) completes the proof.

Proof of Corollary 3.1. Using Proposition 3.1 and then Proposition 3.2,

$$\begin{aligned} R^B_{\underline{d},\Sigma}\{(2\log n)^{\frac{1}{2}}\} &\leq \tilde{R}^B_{\underline{d},\Sigma}\{(2\log n)^{\frac{1}{2}}\} \\ &\leq \max\left\{(\pi\log n)^{\frac{1}{2}}+(2\pi)^{\frac{1}{2}}, 2(\pi\log n)^{\frac{1}{2}}\right\} R^U_{\delta,\sigma}\{(2\log n)^{\frac{1}{2}}\}. \end{aligned}$$

By Theorem 7 from Donoho and Johnstone (1994), we have

$$R^U_{\delta,\sigma}\{(2\log n)^{\frac{1}{2}}\} \leq (2\log n + 2.4)\left\{\frac{\sigma^2}{n}+\min\left(\sigma^2,\delta^2\right)\right\}$$

for $n \geq 4$. From the definition of δ at (3.2), we get successively that

$$\delta^2 = \frac{d^2 + d'^2 - 2\rho d d'}{1-\rho^2} \leq \frac{d^2 + d'^2}{1-|\rho|},$$
$$\frac{\sigma^2}{n} + \min\left(\sigma^2, \delta^2\right) \leq \frac{1}{1-|\rho|}\left\{\frac{\sigma^2}{n} + \min(\delta^2, d^2 + d'^2)\right\}.$$

Finally, it is easy to see that $\min(\delta^2, d^2 + d'^2) \leq \min(\delta^2, d^2) + \min(\delta^2, d'^2)$. Combining the above in an obvious way leads to the result.

Proof of Theorem 4.1. The first equality comes from the orthonormality of the DWT. Assume $n \geq 4$. Using the bound (3.4), for those levels where the bivariate estimator is used ($|\rho_j| \leq \overline{\rho}$), we have that

$$E\left(\hat{d}_{j,k}^{(l)} - d_{j,k}^{(l)}\right)^2 \leq \frac{q\{(\log\, n)^{\frac{1}{2}}\}}{1-\overline{\rho}}\left[\frac{\sigma^2}{n} + \min\left\{\left(d_{j,k}^{(1)}\right)^2, \sigma^2\right\} + \min\left\{\left(d_{j,k}^{(2)}\right)^2, \sigma^2\right\}\right]. \tag{A.14}$$

For the remaining levels, by the oracle inequality of Donoho and Johnstone (1994),

$$E\left(\hat{d}_{j,k}^{(l)} - d_{j,k}^{(l)}\right)^2 \leq (2\log\, n + 2.4)\left[\frac{\sigma^2}{n} + \min\left\{\left(d_{j,k}^{(l)}\right)^2, \sigma^2\right\}\right]. \tag{A.15}$$

The bound on the RHS of (A.15) is majorised by the bound on the RHS of (A.14), so that (A.14) can be used for all levels j, irrespective of ρ_j. The rest of the proof proceeds *exactly* like the proof of Theorem 3 in Johnstone and Silverman (1997), but uses the bound (A.14) instead of the bound from Theorem 1 of Johnstone and Silverman (1997). The adaptation is straightforward and we omit the details.

Proof of Theorem 5.2. The first equality comes from the orthonormality of the DWT. Assume $n \geq 4$. Using the bound (3.4) (for $j = 1, \ldots, J-1$) and the oracle inequality of Donoho and Johnstone (1994) (for $j = 0$), we have

$$\begin{aligned}
&\sum_{j=0}^{J-1}\sum_{k=1}^{2^j} E\left(\hat{d}_{j,k} - d_{j,k}\right)^2 \\
&\quad \leq q\{(\log\, n)^{\frac{1}{2}}\}\Bigg[\sigma^2 + \sum_{j=1}^{J-1}\sum_{k=1}^{2^j}\left\{\min\left(d_{j,k}^2, \sigma^2\right) + \min\left(d_{j-1,\lceil\frac{k}{2}\rceil+\Delta_j}^2, \sigma^2\right)\right\} \\
&\qquad\qquad + \min\left(d_{0,1}^2, \sigma^2\right)\Bigg] \\
&\quad \leq q\{(\log\, n)^{\frac{1}{2}}\}\Big\{\sigma^2 + 3\sum_{j=0}^{J-1}\sum_{k=1}^{2^j}\min\left(d_{j,k}^2, \sigma^2\right)\Big\}.
\end{aligned} \tag{A.16}$$

The rest of the proof proceeds *exactly* like the proof of Theorem 3 in Johnstone and Silverman (1997), but uses the bound (A.16) instead of the bound from Theorem 1 of Johnstone and Silverman (1997). We omit the details.

References

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

Abramovich, F. and Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2000). Adapting to unknown sparsity by controlling the false discovery rate. Technical Report, Statistics Department, Stanford University.

Abramovich, F., Besbeas, P. and Sapatinas, T. (2002). Empirical Bayes approach to block wavelet function estimation. *Comput. Statist. Data Anal.* **39**, 435-451.

Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: a comparative simulation study. *J. Statist. Soft.* **6**, 1-83.

Barber, S. and Nason, G. P. (2003). Simulations comparing thresholding methods using real and complex wavelets. Technical Report 03:07, Department of Mathematics, University of Bristol.

Barber, S. and Nason, G. P. (2004). Real nonparametric regression using complex wavelets. *J. Roy. Stat. Soc. B* **66**, 927-939.

Cai, T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **27**, 898-924.

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

Crouse, M., Nowak, R. and Baraniuk, R. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. Signal Process.* **46**, 886-902.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadelphia, Pennsylvania.

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

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* **90**, 1200-1224.

Downie, T. and Silverman, B. W. (1998). The discrete multiple wavelet transform and thresholding methods. *IEEE Trans. Signal Process.* **46**, 2558-2561.

Dragotti, P. and Vetterli, M. (2003). Wavelet footprints: theory, algorithms, and applications. *IEEE Trans. Signal Process.* **51**, 1306-1323.

Efromovich, S. (1999). Quasi-linear wavelet estimation. *J. Amer. Statist. Assoc.* **94**, 189-204.

Hall, P., Kerkyacharian, G. and Picard, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* **9**, 33-50.

Ito, K. and McKean, H. P. (1974). *Diffusion Processes and Their Sample Paths.* Springer-Verlag, Berlin.

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.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and hay in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32**, 1594-1649.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33**, 1700-1752.

Kohn, R., Marron, J. S. and Yau, P. (2000). Wavelet estimation using Bayesian basis selection and basis averaging. *Statist. Sinica* **10**, 109-128.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal.* **11**, 674-693.

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

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

Nason, G. P. (2002). Choice of wavelet smoothness, primary resolution and threshold in wavelet shrinkage. *Statist. Comput.* **12**, 219-227.

Nason, G. P. and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. In *Lecture Notes in Statistics*, **103** (Edited by A. Antoniadis and G. Oppenheim), 281-300. Springer-Verlag, New York.

Olhede, S. and Walden, A. (2004). 'Analytic' wavelet thresholding. *Biometrika* **91**, 955-973.

Sendur, L. and Selesnick, I. (2002). Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Trans. Signal Process.* **50**, 2744-2756.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets.* Wiley, New York.

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

E-mail: p.z.fryzlewicz@bristol.ac.uk

(Received August 2005; accepted March 2006)