WAVELET ESTIMATION OF AN UNKNOWN FUNCTION OBSERVED WITH CORRELATED NOISE

Xue Wang

Institute of Mathematics, Statistics and Actuarial Science

University of Kent at Canterbury,

Kent, UK

X.Wang@kent.ac.uk

Andrew T. A. Wood

School of Mathematical Sciences

University of Nottingham

Nottingham, UK, NG7 2RD

atw@maths.nottingham.ac.uk

Key Words: Bayes block shrinkage; correlation structure; Durbin-Levinson algorithm; innovations algorithm; pseudo likelihood estimation.

ABSTRACT

In many practical applications of nonparametric regression, it is desirable to allow for the possibility that the noise is correlated. In this paper, we focus on wavelet-based nonparametric function estimation and propose two distinct methods for estimating the correlation structure of the noise, one based in the time domain and the other based in the wavelet domain. Once the correlation structure has been estimated, there are various methods that may be used for reconstructing the unknown signal; we focus here on the empirical Bayes block shrinkage method proposed by Wang and Wood (2006). A simulation study is described. Our numerical results indicate that the proposed methods do a good job of reconstructing the signal even when the noise is highly correlated.

Key words: Bayes block shrinkage; Correlation structure; Durbin-Levinson algorithm; innovations algorithm; pseudo likelihood estimation.

1 Introduction

Various approaches for thresholding and non-linear shrinkage of wavelet coefficients have been shown to perform well under the IID noise assumption; see for example Donoho and Johnstone (1994, 1995), Donoho *et al.* (1995). Bayesian wavelet shrinkage and thresholding approaches have become increasingly popular and have been shown to perform well in practice: for term-by-term Bayesian shrinkage approaches, see for example Chipman *et al.* (1997), Clyde *et al.* (1998), and Johnstone and Silverman (2005a, b); and for Bayesian block shrinkage approaches, see Abramovich *et al.* (2002), De Canditiis and Vidakovic (2004), and Wang and Wood (2006).

Some authors have also considered the situation in which the noise is correlated. Johnstone and Silverman (1997) pointed out that, if the noise in the data is sta-

tionary and correlated, then the variance of the wavelet coefficients will depend on the level in the wavelet decomposition but will be constant at each level. With this in mind, they proposed a level-dependent thresholding approach in which the noise variance at each level is estimated from the data. This is a quick and convenient way to cope with the problem of correlated noise which does not involve full estimation of the correlation structure. However, from the results in §4 we can see that the level-dependent methods do not always work so well with rougher signals (e.g. Bumps). We believe that it is of interest to develop methods for estimating the correlation structure.

From a theoretical perspective, quite a lot is known about how correlated noise affects theoretical performance in nonparametric regression; see Opsomer *et al.* (2001) for a review of how kernel, spline and wavelet approaches to nonparametric function estimation are affected by correlated noise. However, it is unclear to what extent the known theoretical results reflect and capture what happens in practical situations with correlated noise.

The main aim of this paper is to propose two procedures for estimating the correlation structure of the noise. Once the correlation structure is determined, any one of several wavelet block shrinkage methods may be used to reconstruct the unknown function, taking advantage of the information provided by the estimated covariance structure. Here we employ the empirical Bayes block (EBB) shrinkage method proposed by Wang and Wood (2006); other possibilities involving Bayes block shrinkage are to use the approach of either Abramovich *et al.* (2002) or De Canditiis and Vidakovic (2004).

Correlation structure of wavelet coefficients has been considered in some published papers (e.g. Abramovich *et al.*, 2002, Vannucci and Corradi, 1999), but with a somewhat different emphasis. To the best of our knowledge, this is the first paper which considers full estimation of the correlation structure in the

noise in a wavelet setting.

The outline of this paper is as follows. In §2, the basic model considered in this paper is specified and a preliminary study of the correlation structure of the wavelet coefficients demonstrates the potential importance of accounting for correlation of the noise. In §3 we propose two procedures for estimating correlation structure in the noise. The results of a simulation study are presented in §4. Relevant details of EBB method of Wang and Wood (2006) are sketched in the Appendix.

2 General Model Setup

2.1 A Model with Correlated Noise

The model to be considered in this paper is

$$y_i = f(x_i) + \epsilon_i \qquad i = 1, \cdots, n \tag{1}$$

where f is the unknown function to be estimated, $\{y_i\}$ is a set of observations, $x_i = i/n$, $\{\epsilon_i\}$ is a stationary Gaussian sequence with $E(\epsilon_i) = 0$ and stationary covariance function $\gamma(\cdot)$ given by

$$\gamma(i-j) = cov(\epsilon_i, \epsilon_j) = E(\epsilon_i \epsilon_j). \tag{2}$$

For simplicity, we assume that n is an integer power of 2. In our simulation study, described later, we mainly focus on the cases in which $\{\epsilon_i\}$ is an autoregressive process (AR) of order p or a moving average process (MA) of order q,

which are given, respectively, by

$$AR(p) \qquad \epsilon_t = \alpha_1 \epsilon_{t-1} + \cdots + \alpha_p \epsilon_{t-p} + \eta_t, \tag{3}$$

and

$$MA(q) \qquad \epsilon_t = \beta_1 \eta_{t-1} + \cdots + \beta_q \eta_{t-q} + \eta_t, \tag{4}$$

where the η_t are independent $N(0, \sigma_\eta^2)$.

2.2 Variance Analysis

Since the properties of the discrete wavelet transform show that wavelets are "almost eigenfunctions" of many operators (see Frazier *et al.*, 1991, Meyer, 1992), which means that the autocorrelation of the wavelet coefficients of a noisy signal within each level often dies away rapidly, and little or no correlation between the wavelet coefficients at different levels exists (see Johnstone and Silverman, 1997), it is of interest to know how well standard methods (i.e. methods designed for denoising the IID noise) perform in the correlated noise model (1). Here, as a preliminary, we look at the differences between the covariance structure of wavelet coefficients in the presence of IID noise and correlated noise.

If $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T \sim N_n(0, \sigma^2 I_n)$, where I_n is the $n \times n$ identity matrix, then $\mathcal{W}\boldsymbol{\epsilon}$, the discrete wavelet transform (DWT) of $\boldsymbol{\epsilon}$, is also distributed as $N_n(0, \sigma^2 I_n)$, because the DWT matrix $\mathcal{W}$ is orthogonal. In the case of a general correlation structure, $\boldsymbol{\epsilon} \sim N_n(0, V)$, where V is the covariance matrix of the noise, in which case $\mathcal{W}\boldsymbol{\epsilon} \sim N_n(0, \Sigma)$, where $\Sigma = \mathcal{W}V\mathcal{W}^T$.

Using the recursive algorithm proposed by Vannucci and Corradi (1999), which calculates the covariance of wavelet coefficients within and across levels, two plots are obtained to show the difference in the above situations. Fig. 1(a)

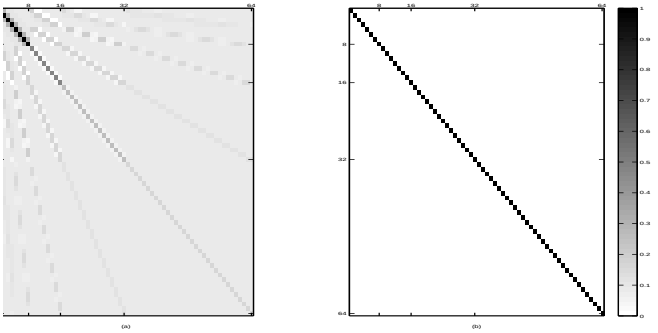

Figure 1: (a): covariance structure of DWT of AR(1) noise with $\alpha = 0.7$. The horizontal line from right to left and vertical line from bottom to top show from the finest level to the coarsest level. The darker the colour of the squares the higher the wavelet coefficients. (b): covariance structure of DWT of IID noise.

shows the covariance matrix of $\mathcal{W}\boldsymbol{\epsilon}$ for AR(1) noise with $\alpha_1 = 0.7$. The small squares along the diagonal of the matrix mark the existing correlation within each level after the DWT has been applied to the (correlated) data. From the finest level to coarsest level (along horizontal line from right to left), the colour of the squares are darker when the correlation of the wavelet coefficients is higher. As a comparison, Fig. 1(b) shows the covariance matrix of IID noise. In this case the covariance matrix is an identity matrix and therefore the variances at all levels are the same.

Further results have shown that, if standard methods are used on correlated data, it may seriously affect the quality of the reconstruction of f, particularly when the data are highly correlated.

3 Estimation of covariance structure

3.1 The procedure

We now propose a four-step procedure for estimating an unknown function in the presence of correlated noise. Each step is discussed in more detail below.

Step 1: identify a parametric model for the correlation structure.

Step 2: estimate the correlation parameters for the model obtained in Step 1.

Step 3: using the model identified in Step 1 with estimated parameters obtained in Step 2, calculate estimates for the variances and covariances of the wavelet coefficients in each block.

Step 4: estimate the signal f, making use of the estimated variances and covariances obtained in Step 3.

If the parametric structure of the covariance matrix $V(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter vector involved in determining the specific noise process, is assumed known, there is no need for Step 1. However, in many situations, $V(\boldsymbol{\theta})$ will be unknown, in which case we suggest implementing Step 1 as follows. Starting with the model $\mathbf{y} = \mathbf{f} + \boldsymbol{\epsilon}$, obtain a preliminary estimate $\hat{\mathbf{f}}$ of $\mathbf{f}$ using a suitable estimation procedure, such as the level-dependent universal threshold method due to Johnstone and Silverman (1997). Then estimate the (unobserved) noise vector by $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{f}}$, and use standard time series model identification techniques on $\hat{\boldsymbol{\epsilon}}$ to determine a suitable parametric covariance structure for $\hat{\boldsymbol{\epsilon}}$. We consider two illustrative examples below which use the Durbin-Levinson algorithm and innovations algorithm, respectively; see e.g. Brockwell and Davis (1991) for details of these algorithms.

Once a parametric model has been identified, we may estimate the unknown parameters using a standard procedure such as the Durbin-Levinson algorithm in the AR case or the innovations algorithm in the MA case. We refer to such procedures as time domain procedures. A second option for Step 2, referred to as a wavelet domain procedure, is discussed in subsection 3.2. We mention two further possibilities which we do not pursue here: (i) the use of the Whittle likelihood (see e.g. Hannan, 1994); and (ii) nonparametric estimation of the correlation function $\gamma(\cdot)$ in (2) (see e.g. Hall *et al.*, 1994), which in effect avoids the need for Step 1.

For Step 3, we simply pick out the required elements of the estimated covariance matrix of the wavelet coefficients, $\mathcal{W}\hat{V}\mathcal{W}^T$, where $\hat{V} = V(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix of $\boldsymbol{\epsilon}$, $\hat{\boldsymbol{\theta}}$ is the estimate of the unknown covariance parameter vector obtained in Step 2, and $\mathcal{W}$ is the discrete wavelet transform.

For Step 4, we use the approach proposed by Wang and Wood (2006); a brief outline of this approach is given in the Appendix.

We now present two examples which illustrate Step 1.

Example 1 *: 1024 data from a simulated AR(2) process with coefficients* $\alpha_1 = 0.7$ *and* $\alpha_2 = -0.2$ *are added to the HeaviSine signal* $\mathbf{f}$ *(see §4 for further details of this signal). Using the level-dependent universal threshold method, we obtain the smoothed signal* $\hat{\mathbf{f}}$*. Hence we estimate the noise vector* $\boldsymbol{\epsilon}$ *using* $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{f}}$*. By applying the Durbin-Levinson algorithm to fit successively higher order autoregressive processes to* $\hat{\boldsymbol{\epsilon}}$*, we obtain the sample partial autocorrelation function (the sample pac.f)* $\hat{\alpha}_{jj}$*. The first 40 numbers of the sample pac.f with the bounds* $\pm 1.96n^{-1/2}$ *are shown in Fig. 2. Inspection of the graph supports the view that the appropriate model for the noise is an AR(2) process because the sample pac.f is near zero after lag-2.*

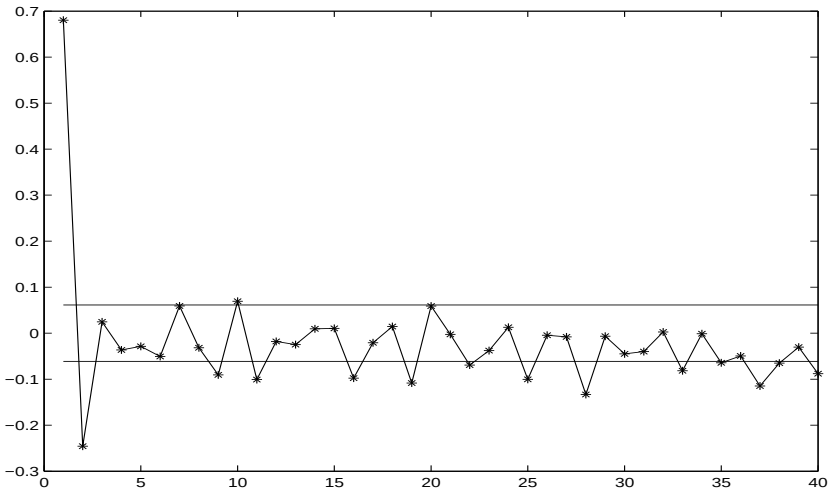

Figure 2: The first 40 numbers of the sample pac.f for the estimated data $\hat{\epsilon}_i$ with the bounds $\pm 1.96 n^{-1/2}$.

Example 2 *: 512 data from a simulated MA(1) process with coefficients* $\beta = 0.5$ *are added to the Doppler signal* $\mathbf{f}$ *(see §4 for further details of this signal). Using the same steps as in Example 1, we obtain* $\hat{\boldsymbol{\epsilon}}$*. By applying the innovations algorithm to fit successively higher order moving average processes to* $\hat{\boldsymbol{\epsilon}}$*, we obtain the estimated coefficient values* $\hat{\beta}_{mj}$ *and noise variances* $\hat{v}_m$*. Table 1 shows* $\hat{\beta}_{mj}$*,* $j = 1, \ldots, 8$ *and* $\hat{v}_m$*,* $m = 1, \ldots, 10, 50, 100$*. This table suggests that MA(1) is the appropriate model for the noise since the estimated coefficients for the orders higher than* 1 *are close to zero.*

3.2 A wavelet domain procedure for Step 2

We now consider a wavelet domain procedure for Step 2 which uses the finest-level wavelet coefficients only; the rationale is that the finest-level coefficients tend to be less affected than coefficients at other levels by the smooth part of the signal. Thus we only use a part of the wavelet transform, represented by

	$\hat{\beta}_{mj}$								$\hat{s}_m$
$m \setminus j$	1	2	3	4	5	6	7	8	
1	0.437								1.021
2	0.513	0.051							0.919
3	0.527	0.059	0.013						0.904
4	0.532	0.053	0.025	-0.017					0.902
5	0.532	0.055	0.021	-0.005	-0.029				0.900
6	0.532	0.054	0.023	-0.009	-0.020	-0.029			0.900
7	0.533	0.054	0.023	-0.008	-0.021	-0.028	-0.015		0.900
8	0.534	0.051	0.023	-0.015	-0.012	-0.048	0.032	-0.095	0.899
9	0.544	0.051	0.024	-0.015	-0.011	-0.050	0.037	-0.101	0.893
10	0.550	0.050	0.024	-0.016	-0.011	-0.051	0.037	-0.102	0.889
50	0.526	-0.002	-0.024	-0.053	-0.039	-0.067	0.008	-0.115	0.842
100	0.534	0.007	-0.034	-0.053	-0.031	-0.067	0.007	-0.129	0.804

Table 1: The estimated coefficients $\hat{\beta}_{mj}$, $j = 1, \cdots, 8$ and noise standard deviation $\hat{s}_m$, $m = 1, \cdots, 10, 50, 100$ for the estimated error vector $\hat{\boldsymbol{\epsilon}}$.

the $n/2 \times n$ matrix $\mathcal{W}_J$, a submatrix of the DWT $\mathcal{W}$, i.e. $\mathcal{W}_J$ maps $\mathbf{y}$ to the finest-level wavelet coefficients according to $\tilde{\mathbf{d}}_J = \mathcal{W}_J\mathbf{y}$. Under the model (1), $\tilde{\mathbf{d}}_J$ has distribution $N_{n/2}(\mathbf{d}_{signal}, \Sigma_J)$, where $\Sigma_J = \mathcal{W}_J V \mathcal{W}_J^T$. Consider the decomposition $\tilde{\mathbf{d}}_J = \mathbf{d}_{signal} + \mathbf{d}_{noise}$, where $\mathbf{d}_{signal} = \mathcal{W}_J\mathbf{f}$, and $\mathbf{d}_{noise} = \mathcal{W}_J\boldsymbol{\epsilon}$, where $\mathbf{d}_{noise} \sim N_{n/2}(\mathbf{0}, \Sigma_J)$. The wavelet domain procedure for Step 2, referred to as Step 2W with sub-steps (a)–(c), estimates $\boldsymbol{\theta}$ using an estimate $\hat{\mathbf{d}}_{noise}$ of $\mathbf{d}_{noise}$ obtained as follows.

Step 2W(a) Shrink (or threshold) the finest-level wavelet coefficients, $\tilde{\mathbf{d}}_J$ say, to obtain $\hat{\mathbf{d}}_{signal}$;

Step 2W(b) Estimate the portion of the finest-level wavelet coefficients attributable to the noise by $\hat{\mathbf{d}}_{noise} = \tilde{\mathbf{d}}_J - \hat{\mathbf{d}}_{signal}$;

Step 2W(c) Use maximum likelihood, or if more convenient, a pseudo-likelihood procedure, with estimated data $\hat{\mathbf{d}}_{noise}$, to estimate the unknown covariance parameters of $V(\boldsymbol{\theta})$.

The reason for including Step 2W parts (a) and (b) is as follows: if the signal has a few discontinuities, then the finest-level coefficients may have a few very large values due to discontinuities in the signal rather than due to the noise. The purpose of Step 2W(a) and (b) is to remove the large coefficients due to the signal and identify the wavelet coefficients which come from the noise.

Once $\hat{\mathbf{d}}_{noise}$ has been obtained, we may implement Step 2W(c) as follows. To simplify the presentation, we write $\mathbf{d}$ rather than $\hat{\mathbf{d}}_{noise}$. The approximate log-likelihood based on $\mathbf{d}$ is given by

$$l(\boldsymbol{\theta}) = \log\{f(\mathbf{d}|\boldsymbol{\theta})\} = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\{\det(\Sigma_J)\} - \frac{1}{2}\mathbf{d}^T\Sigma_J^{-1}\mathbf{d}, \qquad (5)$$

and approximate maximum likelihood estimation of $\boldsymbol{\theta}$ may be carried out by maximizing (5) over the valid parameter space. For example, in the case of AR(1) defined in (3), we have $\sigma^2 > 0$, $0 < \alpha < 1$. However, when the dimension of $\mathbf{d}$ increases, it becomes more difficult to calculate the inverse and determinant of Σ_J. For this reason, we have also investigated various pseudo-likelihood approaches (Besag, 1975, 1977). To implement this approach we split the data into a small number of large blocks, i.e. we split the vector $\mathbf{d}$ into k subvectors of equal dimension, where k is relatively small. For simplicity, we assume that $h = n/(2k)$ is an integer. Then each subvector has h elements, denoted as $\mathbf{d}_i = (d_{i1}, d_{i2}, ..., d_{ih})^T \sim N_h(0, \Sigma_{Ji})$, $i = 1, \ldots, k$. For each block i, we may write log-likelihood as

$$l_i(\boldsymbol{\theta}) = \log\{f(\mathbf{d}_i|\boldsymbol{\theta})\} = const - \frac{1}{2}\log\{\det(\Sigma_{Ji})\} - \frac{1}{2}\mathbf{d}_i^T\Sigma_{Ji}^{-1}\mathbf{d}_i \qquad (6)$$

and the pseudo log-likelihood, $\sum_{i=1}^{k} l_i(\boldsymbol{\theta})$, is the sum of these component log-likelihoods to be maximised. It is worth noting that this pseudo-likelihood approach ignores correlations between blocks. However, when the blocks are

large, inter-block correlations should have a negligible impact on the estimate of $\boldsymbol{\theta}$ provided long-range dependence is not present. In the simulation study presented in §4, the log-likelihood (6) is used in the corresponding simulations, where $k = 4$ was chosen in all cases.

4 Simulation Study

4.1 Specific Covariance Matrices

In this section, we present the results of some simulations to illustrate the procedures proposed above. Three types of correlated noise considered here are AR(1), AR(2) and MA(1). If the noise model is known, we can write down the covariance matrix, see e.g. Cox and Miller (1965). For an AR(1) process, $\epsilon_t = \alpha\epsilon_{t-1} + \eta_t$, with independent $\eta_t \sim N(0, \sigma_\eta^2)$, $t = 1, \ldots, n$, the $n \times n$ covariance matrix of $\boldsymbol{\epsilon}$, with parameters α and $\sigma^2 = \sigma_\eta^2/(1-\alpha^2)$, is given by $\sigma^2 V(\alpha)$, where $V(\alpha) = [V_{ij}(\alpha)]_{i,j=1}^n$ has entries $V_{ij}(\alpha) = \alpha^{|i-j|}$.

For an AR(2) process $\epsilon_t = \alpha_1\epsilon_{t-1} + \alpha_2\epsilon_{t-2} + \eta_t$, where $\eta_t \sim N(0, \sigma_\eta^2)$ independently, if the process is stationary (i.e. $\alpha_1 + \alpha_2 < 1$, $\alpha_2 - \alpha_1 < 1$ and $-1 < \alpha_2 < 1$), the covariance matrix of $\boldsymbol{\epsilon}$ is given by $\sigma^2 V(\alpha_1, \alpha_2)$, where

$$\sigma^2 = \frac{1-\alpha_2}{(1+\alpha_2)[(1-\alpha_2)^2 - \alpha_1^2]}\sigma_\eta^2$$

and the autocorrelation matrix $V(\alpha_1, \alpha_2) = [V_{i,j}(\alpha_1, \alpha_2)]_{i,j=1}^n$ with entries $V_{i,j}(\alpha_1, \alpha_2) = \gamma_{|i-j|}$ where $\gamma_0 = 1$, $\gamma_1 = \alpha_1(1-\alpha_2)^{-1}$, $\gamma_2 = \alpha_2 + \alpha_2^2(1-\alpha_1)^{-1}$ and $\gamma_i = \alpha_1\gamma_{i-1} + \alpha_2\gamma_{i-2}$ for $i > 2$.

For the MA(1) process, the covariance matrix of $\boldsymbol{\epsilon}$ is given by $\sigma^2 V(\beta)$ with parameters β and $\sigma^2 = \sigma_\eta^2$, and $V(\beta) = [V_{ij}(\beta)]_{i,j=1}^n$, where $V_{ii}(\beta) = 1 + \beta^2$, $V_{i-1,i}(\beta) = V_{i,i+1}(\beta) = \beta$, and $V_{ij}(\beta) = 0$ whenever $|i-j| > 1$.

For higher-order AR(p) processes ($p > 2$), it becomes more difficult to write down the explicit form of the covariance matrix $V(\boldsymbol{\theta})$. However, we can obtain the inverse of the covariance matrix, $V^{-1}(\boldsymbol{\theta})$, in the way proposed by Siddiqui (1958), and then we can calculate $\Sigma^{-1} = \mathcal{W}V(\boldsymbol{\theta})^{-1}\mathcal{W}^T$.

In order to gain insight into the behavior at low, medium and high signal-to-noise ratio (SNR), we normalise the noise by using

$$\boldsymbol{\epsilon}_{norm} = \frac{\boldsymbol{\epsilon}}{std(\boldsymbol{\epsilon})} \tag{7}$$

and rescale the signal by using

$$f_{scaled} = f \times (\frac{SNR}{std(f)}), \tag{8}$$

where *std* is the standard deviation of the test function or noise and the value of SNR indicates signal to noise ratio, which can be controlled in the simulation study.

4.2 Simulation Results

To compare the performance of existing methods with the proposed methods in this paper, we use four signals, "HeaviSine", "Blocks", "Bumps" and "Doppler", first proposed in Donoho and Johnstone (1994, 1995) as test functions for wavelet estimators.

In what follows we used the notation WDmean-m and WDmed-m to denote the wavelet domain (WD) procedure outlines in §3.2, with Step 4 implemented using the posterior mean and posterior median version of the EBB procedure of Wang and Wood (2006), respectively; see the Appendix. The quantity m denotes the block size used. Similarly, TDmean-m and TDmed-m denote the corresponding

time domain (TD) procedures, in which the parameter estimates are obtained as a by-product of Step 1 of the general procedure described in §3.1. Also, JS denotes the level-dependent universal thresholding method by Johnstone and Silverman (1997), ETLmean, ETLmed, ETCmean and ETCmed denote, respectively, the posterior mean and posterior median of the level-dependent "EbayesThresh"(ET) method with Laplace (L), where better results are obtained by only estimating the scale parameters empirically from the data, and Cauchy (C) prior proposed by Johnstone and Silverman (2005a, b).

In Table 2, the new methods are compared with a number of existing methods designed for the same situation. The MSE of 9 methods using 100 simulation runs with n=1024, SNR=7, and signals HeaviSine, Block, and Bumps and Doppler are listed. The figure in brackets indicates the relative MSE. For each signal, the relative MSE of the j-th estimator is defined as $\min_k(\text{MSE}_k)/\text{MSE}_j$. The table shows the new methods $1-4$ are quite competitive with the existing methods, especially for the rough signals (e.g Bumps). In two of the four cases with three noise situations (Bumps and Doppler), all the new procedures $1-4$ do better than $5-9$. For the Bumps signal the MSEs of the TD and WD methods are less than one tenth of that of the JS method for three signals. For the remaining signals (HeaviSine and blocks), the results of the new methods are very close (larger than 0.8 of the relative MSE) to the results of the best of the published methods, and WDmean-2 is the best for HeaviSine signal. The TD and WD methods are quite competitive although the pseudo-likelihood estimation based on large blocks of the WD method is computationally intensive, especially when the order of the noise is high. However, considering the improvement of the average MSE and the widespread availability of high-powered computers, this cost is worthwhile.

Fig. 3 shows the reconstructions of Bumps signal from AR(1) ($\alpha = 0.7$) noise

by 9 methods (see details above). Generally speaking, all the methods can give approximately noise-free reconstructions except for the JS method although it gives the right direction of denoising. Even though the TD and WD methods use the JS method as a preliminary, they still can give the excellent denoising results; this is confirmed by the simulations (see Table 2).

Some further comments now follow.

Remark 1: Boundary correction. One possibility we have not discussed so far is the use of boundary correction, based on symmetric reflection of the function f beyond the boundaries of the domain; see for example Abramovich and Benjamini (1996). A simulation study, not reported here, was undertaken to compare the performance of the proposed methods with and without boundary correction when applying the DWT to the signals. The overall performance of boundary correction did not show clear superiority over implementation without boundary correction. For this reason, we did not include boundary correction methods in the results reported here.

Remark 2: Choice of block size. The choice of block size is an important issue for block thresholding methods, and has been discussed by several authors in an asymptotic frequentist framework; see, for example, Hall *et al.* (1998), Cai (2002), and Cai and Zhou (2008). Numerical simulations for different block sizes (1,2,4,8,16) with different signal-to-noise ratios, SNR, (3,5,7) have been undertaken, and results suggest that the block size favored by each signal is quite stable across the different signal-to-noise ratios. The numerical results for different block sizes (1,2,4,8,16) with sample size 1024 and SNR 7 are presented in Table 3. Table 3 suggests that a small block size, (1,2,4), is appropriate for the given examples. It is interesting to note that the best choice of block size for each signal contaminated by the three types of noises agreed with the signal had IID Gaussion noise; see Wang and Wood (2006).

signal	methods	AR(1)	AR(2)	MA(1)
HeaviSine	1. WDmean-2	0.2555∗ (1.0)	0.1408∗ (1.0)	0.1158∗ (1.0)
	2. TDmean-2	0.2976 (.859)	0.1776 (.793)	0.1286 (.901)
	3. WDmed-2	0.2861 (.893)	0.1615 (.872)	0.1263 (.917)
	4. TDmed-2	0.2897 (.882)	0.1725 (.816)	0.1239 (.935)
	5. JS	0.3337 (.766)	0.2316 (.608)	0.1848 (.626)
	6. ETLmean	0.2636 (.969)	0.1543 (.913)	0.1236 (.937)
	7. ETLmed	0.2803 (.912)	0.1607 (.876)	0.1291 (.897)
	8. ETCmean	0.2628 (.972)	0.1537 (.916)	0.1229 (.942)
	9. ETCmed	0.2803 (.912)	0.1607 (.876)	0.1287 (.900)
Bumps	1. WDmean-2	0.5444∗ (1.0)	0.4261∗ (1.0)	0.3599 (.994)
	2. TDmean-2	0.5493 (.991)	0.4816 (.885)	0.3576∗ (1.0)
	3. WDmed-2	0.5703 (.955)	0.4423 (.963)	0.3796 (.942)
	4. TDmed-2	0.6055 (.899)	0.4696 (.907)	0.3629 (.985)
	5. JS	6.8436 (.079)	6.9480 (.061)	6.9175 (.052)
	6. ETLmean	0.8402 (.648)	0.6606 (.645)	0.5634 (.635)
	7. ETLmed	1.0525 (.517)	0.8144 (.523)	0.6859 (.521)
	8. ETCmean	0.8682 (.627)	3.3679 (.127)	0.5727 (.624)
	9. ETCmed	1.0973 (.496)	0.8451 (.504)	0.7093 (.504)
Blocks	1. WDmean-2	0.3612 (.920)	0.2667 (.880)	0.2221 (.839)
	2. TDmean-2	0.4102 (.810)	0.2939 (.798)	0.2329 (.800)
	3. WDmed-2	0.3569 (.931)	0.2701 (.869)	0.2231 (.836)
	4. TDmed-2	0.3679 (.903)	0.2917 (.804)	0.2057 (.906)
	5. JS	0.4650 (.715)	0.3560 (.659)	0.2827 (.659)
	6. ETLmean	0.3482 (.954)	0.2494 (.941)	0.2007 (.929)
	7. ETLmed	0.3480 (.955)	0.2505 (.937)	0.1981 (.941)
	8. ETCmean	0.3323∗ (1.0)	0.2346∗ (1.0)	0.1864∗ (1.0)
	9. ETCmed	0.3381 (.983)	0.2416 (.971)	0.1894 (.984)
Doppler	1. WDmean-2	0.3038∗ (1.0)	0.2473∗ (1.0)	0.1875∗ (1.0)
	2. TDmean-2	0.3732 (.814)	0.2932 (.844)	0.2035 (.844)
	3. WDmed-2	0.3211 (.946)	0.2927 (.845)	0.2221 (.845)
	4. TDmed-2	0.3542 (.856)	0.2962 (.835)	0.1971 (.835)
	5. JS	0.5463 (.556)	0.4342 (.569)	0.3747 (.570)
	6. ETLmean	0.3825 (.794)	0.2787 (.887)	0.2326 (.887)
	7. ETLmed	0.3984 (.763)	0.3009 (.822)	0.2538 (.822)
	8. ETCmean	0.3790 (.802)	0.2773 (.892)	0.2312 (.882)
	9. ETCmed	0.3985 (.762)	0.3059 (.808)	0.2569 (.808)

Table 2: The comparison of MSE of 9 methods using 100 simulation runs with n= 1024, SNR= 7, signals HeaviSine, Blocks, Bumps and Doppler contaminated with noises AR(1) with $\alpha = 0.7$, AR(2) with $\alpha(1) = 0.7$ and $\alpha(2) = -0.2$ and MA(1) with $\beta = 0.5$. The figures in brackets are the relative MSEs, which indicate the ratios of the MSE with the minimum MSE achieved for each example.

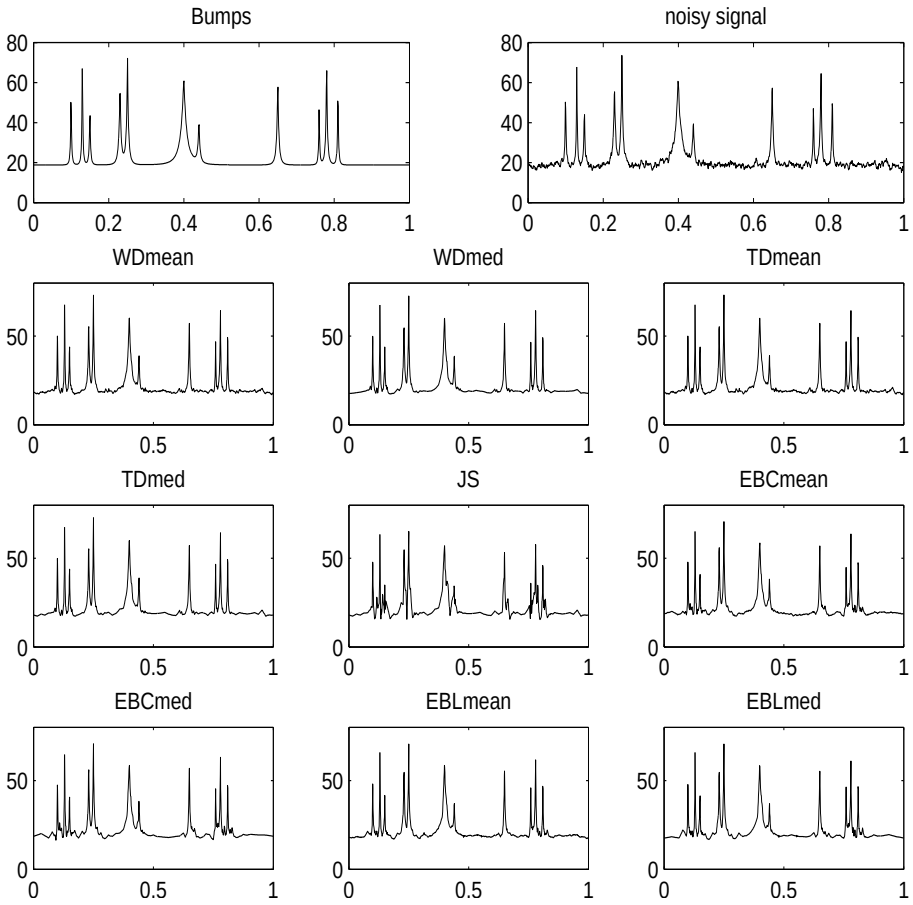

Figure 3: Bumps signal with AR(1), $\alpha = 0.7$, based on sample size $n = 1024$ and $SNR = 0.7$. The reconstructions are obtained using 9 methods (see details in text).

HeaviSine	methods	m=1	m=2	m=4	m=8	m=16
AR(1)	WDmean	0.2536∗	0.2555	0.2988	0.3112	0.3222
	WDmed	0.2837∗	0.2861	0.3078	0.3149	0.323
AR(2)	WDmean	0.1414	0.1408∗	0.1684	0.1850	0.1980
	WDmed	0.1678	0.1615∗	0.1694	0.1879	0.2010
MA(1)	WDmean	0.1171	0.1158∗	0.1351	0.1485	0.1655
	WDmed	0.1331	0.1263∗	0.1424	0.1569	0.1692
Blocks	methods	m=1	m=2	m=4	m=8	m=16
AR(1)	WDmean	0.3152∗	0.3612	0.4602	0.5700	0.6804
	WDmed	0.2967∗	0.3569	0.4335	0.5604	0.6738
AR(2)	WDmean	0.2231∗	0.2667	0.3469	0.4415	0.5558
	WDmed	0.2094∗	0.2701	0.3384	0.4445	0.5596
MA(1)	WDmean	0.1823∗	0.2221	0.2905	0.3759	0.4851
	WDmed	0.1796∗	0.2231	0.2827	0.3722	0.4869
Bumps	methods	m=1	m=2	m=4	m=8	m=16
AR(1)	WDmean	0.5946	0.5444∗	0.5879	0.6846	0.7883
	WDmed	0.6616	0.5703∗	0.5566	0.6681	0.7863
AR(2)	WDmean	0.4848	0.4261∗	0.4634	0.5633	0.6852
	WDmed	0.5409	0.4423∗	0.4463	0.5525	0.6874
MA(1)	WDmean	0.4126	0.3599∗	0.3876	0.4779	0.5973
	WDmed	0.4737	0.3796∗	0.3764	0.4706	0.5993
Doppler	methods	m=1	m=2	m=4	m=8	m=16
AR(1)	WDmean	0.3467	0.3038∗	0.3054	0.3649	0.3570
	WDmed	0.3762	0.3211∗	0.3265	0.3637	0.3482
AR(2)	WDmean	0.2893	0.2473	0.2134∗	0.2411	0.2360
	WDmed	0.3302	0.2927	0.2335∗	0.2334	0.2288
MA(1)	WDmean	0.2330	0.1875	0.1706∗	0.1883	0.1815
	WDmed	0.2823	0.2035	0.1891∗	0.1884	0.1797

Table 3: Simulation results for WDmean and WDmed comparing different block sizes m based on 100 simulation runs. The number in each cell is MSE and an asterisk is used to identify the optimum block size within a row.

	methods	HeaviSine	Blocks	Bumps	Doppler
AR(2) known	WDmean	0.1261	0.2466	0.3837	0.2212
	WDmed	0.1289	0.2510	0.4120	0.2628
AR(1) struc	WDmean	0.1961	0.2808	0.4468	0.2503
AR(2) noise	WDmed	0.2078	0.2644	0.4244	0.2461
MA(1) known	WDmean	0.1148	0.2238	0.3598	0.1860
	WDmed	0.1217	0.2132	0.3679	0.2037
AR(1) struc	WDmean	0.1418	0.2311	0.3682	0.1896
MA(1) noise	WDmed	0.1549	0.2283	0.3663	0.2170

Table 4: Simulation study for WDmean and WDmed comparing the noise process known and the noise process incorrectly specified. The numbers in the second and fourth rows are MSEs when the noise processes were known as AR(2) with $\alpha(1) = 0.7$ and $\alpha(2) = -0.2$ and MA(1) with $\beta = 0.5$. The numbers in the third and fifth rows are MSEs when the noise processes were AR(2) and MA(1) but incorrectly identified as AR(1) model.

Remark 3: Real noise structures known. It should be noted that, in the simulation study, the correct ARMA model was assumed when implementing the TD and WD approaches. Thus, the TD and WD methods had an advantage that they would not have in a real data example, where a parametric model for the correlation structure would be unknown. Before showing a real data example which suggests that satisfactory results may be expected even when the covariance structure is unknown, we consider the effect of incorrectly specifying the noise process. Table 4 shows the simulation results under the following four situations: the noise processes were known to be AR(2) with $\alpha(1) = 0.7$ and $\alpha(2) = -0.2$ and MA(1) with $\beta = 0.5$, the estimations were assuming an AR(1) model while the noises were AR(2) and MA(1). The first and third rows in the table showed the best denoising results we can hope to obtain since we know the correct noise processes. The second and fourth rows gave the results when we incorrectly specify the noise process. Compare these values with associated values in Table 2, we can see that it is not a disaster if we incorrectly specify the noise process although it does influence the denoising results.

4.3 The Real Data Case

Nason (1996) described a dataset obtained in an anesthesiological study using inductance plethsmography. The data was collected in an investigation of the recovery of patients after general anesthesia, which is available as part of the **wavethresh4** package (Nason 2006). The original data `ipd` are plotted in Fig. 4(a).

Johnstone and Silverman (2005b) presented two versions of this dataset, smoothed in two different ways. One way assumed that the original signal was observed with IID noise, where the ET method with the same noise estimate for all levels was used. The other way assumed stationary correlated noise and used the level-dependent ET method. Fig. 4(b) shows the smoothed version of the `ipd` data with IID noise assumption, and Fig. 4(c) with the stationary correlated noise assumption. From these two plots we can see that the reconstruction by assuming the stationary correlated noise, plot (c), removes some moderately high frequency effects which still exist in the reconstruction (b), for example, in the interval [1000, 2000].

To find the smoothed version of the data by the new methods proposed in this paper, we first try to find the noise model, and the Durbin-Levinson algorithm is used to fit successively higher order autoregressive processes to the estimation of $\boldsymbol{\epsilon}$. We then obtain the sample pac.f $\hat{\alpha}_{jj}$. The first 50 numbers of sample pac.f with the bounds $\pm 1.96 n^{-1/2}$ are shown in Fig. 5. The AR(2) model for the noise here with $\alpha_1 = 0.38$ and $\alpha_2 = 0.27$ will be used to balance the accuracy and simplicity. The smoothed versions of the `ipd` data by TDmean and TDmed are plotted in Fig. 4(d), (e). These plots show that the proposed methods are able to remove the noise effectively. Especially, in the interval [1000, 2000], the proposed methods remove the local variability better but still keep the peaks well.

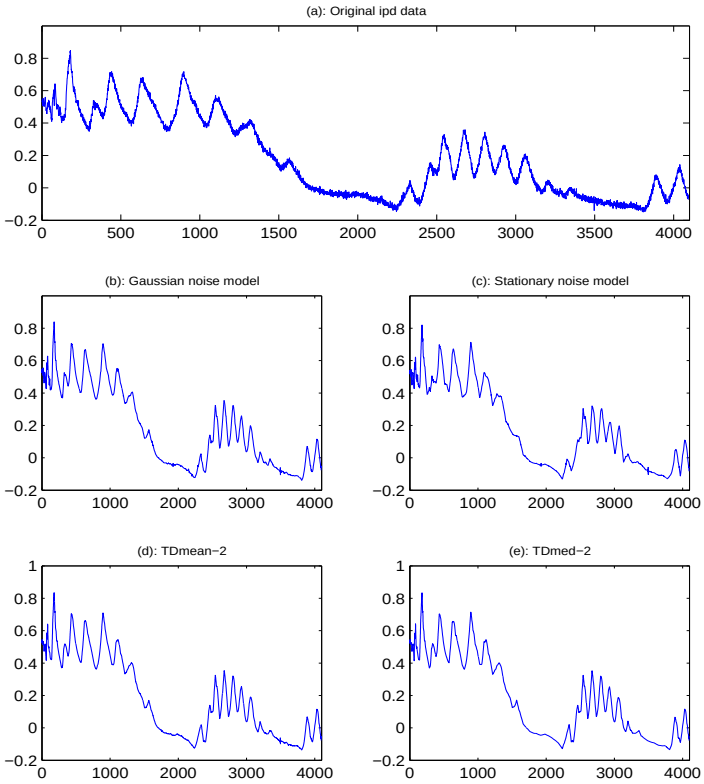

Figure 4: (a): the original ipd data; (b): the reconstruction obtained by ETCmean method with IID noise assumption; (c): the reconstruction obtained by ETCmean with stationary correlated noise assumption; (d): the reconstruction by TDmean-2; (e): the reconstruction by TDmed-2.

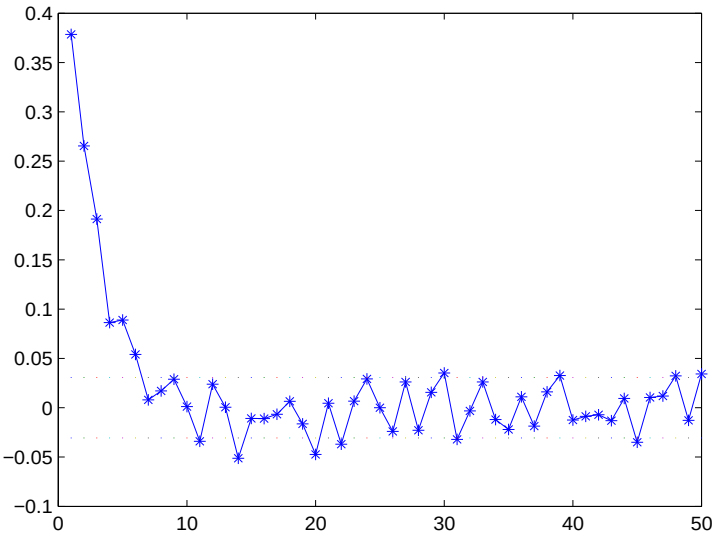

Figure 5: The first 50 numbers of the sample pac.f for the estimated noise $\boldsymbol{\epsilon}$ with the bounds $\pm 1.96 n^{-1/2}$.

5 Conclusions

In this paper we have considered two procedures for wavelet estimation of a signal in the presence of correlated noise, a time domain procedure and a wavelet domain procedure. Both procedures involve explicit estimation of the correlation structure of the noise. The results in §4 indicate that both of our proposals do better than the original level-dependent thresholding method of Johnstone and Silverman (1997), but when the level-dependent thresholding approach is used in combination with the empirical Bayes procedures of Johnstone and Silverman (2005a), the resulting performance is rather similar to that of our proposals. In Table 2, it is only in the case of the Bumps signal that there is a noticeable difference (with our proposals doing better). However, even though level-dependent procedures are simpler to implement than those procedures in

which the correlation structure is estimated, it is nevertheless worthwhile to develop estimation procedures of the latter type. Our numerical results indicate that the two procedures of the latter type proposed here may be expected to perform well in practice.

References

[1] Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. Computational Statistics and Data Analysis 22: 351–361.

[2] Abramovich, F., Besbeas, P. and Sapatinas, T. (2002). Empirical Bayes approach to block wavelet function estimation. Computational Statistics and Data Analysis 39: 435–451.

[3] Besag, J. E. (1975). Statistical analysis of non-lattice data. The Statistician 24: 179–195.

[4] Besag, J. E. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. Biometrika 64: 616–618.

[5] Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. New York: Spring-Verlag.

[6] Cai, T. T. (2002). On the block thresholding in wavelet regression: adaptivity, block size, and threshold level. Statistica Sinica 12: 1241–1237.

[7] Cai, T. T. and Zhou, H. H. (2008). A data-driven block thresholing approach to wavelet estimation. Annals of Statistics 36

[8] Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. Journal of American Statistical Association 92: 1413–1421.

[9] Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. Biometrika 85: 391–402.

[10] Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes.* London: Chapman and Hall.

[11] De Canditiis, D. and Vidakovic, B. (2004). Wavelet Bayesian block shrinkage via mixtures of normal-inverse gamma priors. Journal of Computational and Graphical Statistics 13: 383–398.

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

[13] Donoho, D. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. Journal of American Statistical Association 90: 1200–1224.

[14] Donoho, D., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). Journal of Royal Statistical Society (B) 57: 301–369.

[15] Frazier, M., Jawerth, B. and Weiss, G. (1991). Littlewood-Paley theory and the study of function spaces. CBMS Regional Conference Series 79.

[16] Hall, P., Fisher, N.I. and Hoffmann, B. (1994). On the nonparametric estimation of covariance functions. Annals of Statistics 22: 2115–2134.

[17] Hannan, E. J. (1994). The Whittle likelihood and frequency estimation. In:*Probability, Statistics and Optimisation: a tribute to Peter Whittle.* Wiley Series in Probability and Mathematical Statistics. Wiley, Chichester.

[18] Johnstone, I. M. and Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise. Journal of Royal Statistical Society (B) 59: 319–351.

[19] Johnstone, I. M. and Silverman, B. W. (2005a). Empirical Bayes selection of wavelet thresholds. Annals of Statistics 33: 1700–1752.

[20] Johnstone, I. M. and Silverman, B. W. (2005b). EbayesThresh: R programs for empirical Bayes thresholding. Journal of Statistical Software 12(8).

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

[22] Nason, G. P. (1996). Wavelet shrinkage using cross-validation. Journal of Royal Statistical Society (B) 58: 463–479.

[23] Nason, G. P. (2006). *WaveThresh4 Software*. Department of Mathematics, University of Bristol, UK. URL http://www.stats.bris.ac.uk/ wavethresh/.

[24] Opsomer, J., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. Statistical Science 16: 134–153.

[25] Hall, P., Kerkyacharian, G. and Picard, D. (2001). Block threshold rules for curve estimation using kernel and wavelet methods. Annals of Statistics 26: 922–942.

[26] Siddiqui, M. M. (1958). On the inversion of the sample covariance matrix in a stationary autoregressive process. Annals of Mathematical Statistics 29: 585–588.

[27] Vannucci, M. and Corradi, F. (1999). Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. Journal of Royal Statistical Society (B) 61: 971–986.

[28] Wang, X. and Wood, A. T. A (2006) Empirical Bayes block shrinkage of wavelet coefficients via the non-central χ^2 distribution. Biometrika 93: 705–722.

Appendix: empirical Bayes block shrinkage approach

In Bayesian wavelet shrinkage methods, a prior distribution is specified on the wavelet coefficients which is designed to capture the sparseness of the wavelet expansions that is common to most applications. The function can then be estimated by applying a suitable Bayes rule to the resulting posterior distribution of wavelet coefficients.

A popular prior model for each wavelet coefficient d_{jk}, where j is the resolution level and k is the location, is a mixture of a normal distribution and a point mass at zero. The normal distribution with positive variance, $N(0, \lambda_j^2)$, represents the possibility of a non-zero coefficient while a point mass at zero, $\delta(0)$, represents a negligible coefficient. A hierarchical model can be expressed as

$$d_{jk}|r_j \sim r_j N(0, \lambda_j^2) + (1 - r_j)\delta(0), \tag{9}$$

with $r_j \sim Bernoulli(p_j)$ for different resolution levels j. The binary random variable r_j determines whether the relevant wavelet coefficient is nonzero ($r_j = 1$) and comes from a normal distribution, or zero ($r_j = 0$), and arises from a point mass at zero. Suitable Bayesian wavelet shrinkage and thresholding estimators are the posterior mean and the posterior median.

The Bayesian wavelet shrinkage method used in Step 4 of the procedure in §3.1 is the empirical Bayes block (EBB) shrinkage method proposed by Wang and Wood (2006). The EBB method, which performs the shrinkage based on the sum of squares of wavelet coefficients in each single block, takes account of the information in neighbouring coefficients. In a correlated noise situation, the wavelet coefficients typically are highly correlated with their near neighbours. Thus a shrinkage procedure based on a quadratic form in the wavelet coefficients which takes account of these correlations makes good sense.

After performing the DWT on the noisy observations $\mathbf{y}$ in (1), we obtain the empirical wavelet coefficients $\tilde{d}_{jk}$, which are candidates for shrinkage. Let B represent a single block where, typically, a block would consist of neighboring coefficients at the same resolution level j. Define $\tilde{\mathbf{d}}_{jB} = \{\tilde{d}_{jk} : k \in B\}$ and let $n(B)$ denote the number of elements (i.e. labels) in B. Following the discussion in §2.2, we have

$$\tilde{\mathbf{d}}_{jB} \sim N_{n(B)}(\mathbf{d}_{jB}, \Sigma_{n(B)}), \tag{10}$$

where $\mathbf{d}_{jB}$ is the noiseless version of $\tilde{\mathbf{d}}_{jB}$ and $\Sigma_{n(B)}$ is the relevant $n(B) \times n(B)$ submatrix of $\Sigma = \mathcal{W}V(\boldsymbol{\theta})\mathcal{W}^T$, obtained in Step 3 of the procedure described in §3.1.

Define

$$z = \tilde{\mathbf{d}}_{jB}^T \Sigma_{n(B)}^{-1} \tilde{\mathbf{d}}_{jB} \quad \text{and} \quad \rho = \mathbf{d}_{jB}^T \Sigma_{n(B)}^{-1} \mathbf{d}_{jB}.$$

A shrinkage procedure can be derived by imposing a suitable prior on ρ. In this paper, we follow the EBB shrinkage method with the "power" prior (see section 2.2 in Wang and Wood, 2006).

The prior can be expressed as a mixture of a unit point mass $\delta(0)$ on $\rho = 0$ and a certain scaled central chi-squared distribution

$$\rho | r, \beta \sim r\delta(0) + (1 - r)\chi_m^2(\rho | 0, \beta^{-1}),$$

where r is the prior probability that $\rho = 0$, and β is the scale parameter has the distribution function

$$F(\beta | \sigma^2, \lambda) = \Big(\frac{\beta\sigma^2}{1 + \beta\sigma^2}\Big)^{\lambda+1}.$$

The likelihood of $z|\rho$ is given by a noncentral chi-squared distribution, $\chi^2_m(z|\rho, \sigma^2)$, and the mean and median of the resulting posterior distribution for ρ may be calculated numerically. The mean or median of this posterior distribution are chosen to be the estimators of the "true" wavelet coefficients. For full details of the EBB shrinkage method, see Wang and Wood (2006).