J. R. Statist. Soc. B (1999)
61, *Part* 4, *pp.* 971–986

Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective

Marina Vannucci

Texas A&M University, College Station, USA

and Fabio Corradi

Università di Firenze, Italy

[Received January 1997. Final revision March 1999]

Summary. We present theoretical results on the random wavelet coefficients covariance structure. We use simple properties of the coefficients to derive a recursive way to compute the within- and across-scale covariances. We point out a useful link between the algorithm proposed and the two-dimensional discrete wavelet transform. We then focus on Bayesian wavelet shrinkage for estimating a function from noisy data. A prior distribution is imposed on the coefficients of the unknown function. We show how our findings on the covariance structure make it possible to specify priors that take into account the full correlation between coefficients through a parsimonious number of hyperparameters. We use Markov chain Monte Carlo methods to estimate the parameters and illustrate our method on bench-mark simulated signals.

Keywords: Bayesian inference; Nonparametric regression; Wavelet coefficients' correlation structure; Wavelet shrinkage

1. Introduction

Wavelets are now well established in the literature and have been successfully applied in many areas, such as mathematics, engineering and statistics. This paper develops models and theoretical results for the wavelet coefficients' covariance structure. Our main contribution is the development of a recursive algorithm to compute within- and across-scale covariances. The algorithm proposed has an interesting link to the two-dimensional discrete wavelet transform (DWT), making computations feasible. Our results are generally applicable in many problems that involve wavelet coefficient modelling. We use them in the context of wavelet shrinkage, a well-known application of wavelets to attempt the recovery of a signal from noisy data. Originally proposed by Donoho and Johnstone (1994, 1995, 1998) and Donoho *et al.* (1995), wavelet shrinkage has recently been considered within a Bayesian framework where a prior distribution is imposed on the wavelet coefficients of the unknown signal. Existing work (Chipman *et al.*, 1997; Clyde *et al.*, 1998; Abramovich *et al.*, 1998) assumes independent coefficients. We adopt an approach suggested by Vidakovic and Müller (1995) to incorporate correlation. Our recursive covariance structure gives rise to a model that allows for full correlation between coefficients. A different approach, allowing for partial correlation between coefficients, is proposed by Crouse *et al.* (1998).

Address for correspondence: Marina Vannucci, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.
E-mail: mvannucci@stat.tamu.edu

 1369–7412/99/62971

In addition to the theoretical appeal of totally relaxing the independence assumption, our proposal has the advantage of incorporating knowledge about stochastic relationships between wavelet coefficients. The practical implication of this is a model that depends on a parsimonious number of hyperparameters.

Inference is performed via Markov chain Monte Carlo methods, simulation techniques widely used in Bayesian statistics to produce samples from a posterior distribution. Here we use a hybrid algorithm, combining in a cycle Gibbs and Metropolis steps, as described by Chib and Greenberg (1994) and Müller (1992). Wavelet coefficients are then estimated by averaging over the simulated values.

The paper is organized as follows: Section 2 briefly reviews basic concepts about wavelets and wavelet shrinkage. Section 3 states the results about the wavelet coefficients' covariance structure. Section 4 presents the Bayesian shrinkage model. Applications to bench-mark signals are given in Section 5. In Section 6 a hierarchical structure is introduced and inference on the parameters of interest is made via a hybrid Markov chain Monte Carlo method. Section 7 provides further examples and Section 8 some concluding remarks.

2. Preliminaries and notation

2.1. Orthonormal wavelet bases and wavelet transforms

Wavelets are families of functions that can accurately describe other functions in a parsimonious way; see Daubechies (1992) and Meyer (1992), among others. In $L^2(\mathbb{R})$, for example, a wavelet basis is obtained by translations and dilations of a *scaling function* ϕ, constructed as a solution of a dilation equation

$$\phi(t) = \sqrt{2} \sum_{l \in \mathbb{Z}} h_l \, \phi(2t - l), \tag{1}$$

and a *mother wavelet* ψ, defined from ϕ as $\psi(t) = \sqrt{2}\, \Sigma_l \, g_l \, \phi(2t - l)$, with filter coefficients g_l often defined as $g_l = (-1)^l h_{1-l}$. The wavelet collection is obtained by translations and dilations as $\phi_{j,k}(t) = 2^{j/2} \, \phi(2^j t - k)$ and $\psi_{j,k}(t) = 2^{j/2} \, \psi(2^j t - k)$ and the family of wavelets $\{\psi_{j,k}(t), j, k \in \mathbb{Z}\}$ forms an orthonormal basis in $L^2(\mathbb{R})$. Any $L^2(\mathbb{R})$ function f can then be represented by a wavelet series as $f(t) = \Sigma_{j,k \in \mathbb{Z}} \, d_{j,k} \, \psi_{j,k}(t)$ with wavelet coefficients

$$d_{j,k} = \langle f, \psi_{j,k} \rangle = \int f(t) \, \psi_{j,k}(t) \, \mathrm{d}t$$

that describe features of the function f at the spatial location $2^{-j}k$ and frequency proportional to 2^j (or scale j).

Interesting recursive relationships hold between the coefficients $d_{j,k}$ and the scaling coefficients $c_{j,k} = \langle f, \phi_{j,k} \rangle = \int f(t) \, \phi_{j,k}(t) \, \mathrm{d}t$; see Mallat (1989). For example, using equation (1), coefficients at scale j can be obtained from scaling coefficients at the finer scale $j + 1$ as

$$\begin{aligned} c_{j,k} &= \sum_{m \in \mathbb{Z}} h_{m-2k} c_{j+1,m}, \\ d_{j,k} &= \sum_{m \in \mathbb{Z}} g_{m-2k} c_{j+1,m}. \end{aligned} \tag{2}$$

These equations can be written using signal processing terminology. Let F indicate a linear filter defined by an infinite sequence f_l of coefficients and acting as $(Fa)_k = \Sigma_{n \in \mathbb{Z}} \, f_{n-k} a_n$, with a_n an infinite sequence. In this paper we are concerned with Daubechies (1992) wavelets. These wavelets have compact support, implying filters with a finite number of non-zero

coefficients f_l, and issues of convergence do not arise. The filter H, defined by the sequence h_l, is a *low pass* filter capturing low frequency components, whereas G, defined by the sequence g_l, is a *high pass* filter capturing high frequency components. Now let D_0 indicate the down-sampling operator $(D_0 a)_j = a_{2j}$ that chooses every other element of a sequence. With $c^{(j)}$ and $d^{(j)}$ indicating the coefficient vectors at scale j, equations (2) can be expressed as

$$\begin{aligned} c^{(j)} &= H_{j+1} c^{(j+1)}, \\ d^{(j)} &= G_{j+1} c^{(j+1)}, \end{aligned} \tag{3}$$

where H_{j+1} and G_{j+1} are the linear functions corresponding to the application of the filters $D_0 H$ and $D_0 G$. Index $j+1$ indicates that the dimensions of the matrices change with the scale, owing to the down-sampling operation. Equations (2) are used in wavelet theory to derive a fast algorithm, known as the DWT, for decomposing a function into a set of wavelet coefficients. The algorithm for the inverse construction is called the *inverse wavelet transform* (IWT). See Strang (1989) for a detailed exposition of these algorithms.

2.2. Wavelet shrinkage

Let $y_1, \ldots, y_n$, $n = 2^J$, be a sequence of observations modelled as

$$y_i = f(t_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{4}$$

where f is a function to be estimated, $t_i = i/n$ are equally spaced points and ϵ_i are independent and identically distributed normal random variables with zero mean and variance σ^2. The wavelet shrinkage of Donoho and Johnstone (1994, 1995, 1998) and Donoho *et al.* (1995) is a technique consisting of three steps: firstly, the DWT is applied to the data y_i to obtain a vector $\tilde{d}$ of empirical wavelet coefficients. Since the DWT is linear and orthogonal, $\tilde{d}$ can still be modelled as

$$\tilde{d} = d + \tilde{\epsilon}, \tag{5}$$

where d is the vector of wavelet coefficients of the unknown function f and $\tilde{\epsilon}$ is Gaussian white noise with mean 0 and variance–covariance matrix $\sigma^2 I$. Secondly, the noise is suppressed from the empirical coefficients by using a thresholding (and/or shrinkage) method. See Donoho and Johnstone (1994, 1995) for their *hard* and *soft threshold* policies. Finally, the IWT is applied, leading to an estimate of the unknown function. Donoho and Johnstone (1998) showed that wavelet shrinkage estimators are nearly minimax for a wide set of functional classes and a large class of loss functions. Johnstone and Silverman (1997) extended these results to the case of correlated noise. Several different thresholding rules have been proposed. Among others, Nason (1996) and Wang (1996) adjusted the well-known cross-validation criterion to choose the threshold.

An important choice is the coarsest level of the DWT. Thresholding (and/or shrinkage) methods, in fact, are not applied to all the coefficients. Donoho and Johnstone called the coarsest level *low resolution cut-off* j_0, pointing out its interpretation as a bandwidth. Hall and Patil (1995) discussed the importance of this parameter on mean-squared error performance and proposed choosing j_0 to increase with n.

3. Random coefficients' covariances

Here we consider f as a realization of a stochastic process $\{X_t,\ t \in \mathbb{R}\}$ and investigate the

statistical properties of random coefficients $d_{j,k} = \int X(t)\,\psi_{j,k}(t)\,\mathrm{d}t$ and $c_{j,k} = \int X(t)\,\phi_{j,k}(t)\,\mathrm{d}t$. We assume $X(t)$ to be a stochastic process with existing first and second moments. The results of the next section do not require any additional assumptions.

3.1. Recursive structure

We prove a result about the wavelet coefficients' covariances and use it to develop a recursive algorithm to calculate the *within-* and *across-scales* coefficients' covariances. To clarify the terminology, we call within scale the covariances between coefficients that belong to the same scale and across scales those between coefficients that belong to different scales.

Proposition 1. Given a wavelet basis in $L^2(\mathbb{R})$, the following results hold:

(a) $\operatorname{cov}(d_{j,k}d_{j',k'}) = \sum_m \sum_n g_{m-2k}g_{n-2k'} \operatorname{cov}(c_{j+1,m}c_{j'+1,n})$,

(b) $\operatorname{cov}(c_{j,k}c_{j',k'}) = \sum_m \sum_n h_{m-2k}h_{n-2k'} \operatorname{cov}(c_{j+1,m}c_{j'+1,n})$,

(c) $\operatorname{cov}(c_{j,k}d_{j',k'}) = \sum_m \sum_n h_{m-2k}g_{n-2k'} \operatorname{cov}(c_{j+1,m}c_{j'+1,n})$,

with j, j', k and k' integers.

Proof. The proof is straightforward given relationships (2). □

Assume that the within-scale variance–covariance matrix of scaling coefficients $c^{(j+1)}$ at scale $j+1$ is known and denote that matrix by $CC^{\{j+1,j+1\}}$. We state the following results in filter notation (see equation (3)). The within-scale covariances at the coarser scale j can be easily computed as

$$DD^{\{j,j\}} = G_{j+1}[CC^{\{j+1,j+1\}}]G_{j+1}^{\mathrm{T}}, \tag{6}$$

$$CC^{\{j,j\}} = H_{j+1}[CC^{\{j+1,j+1\}}]H_{j+1}^{\mathrm{T}}, \tag{7}$$

$$CD^{\{j,j\}} = G_{j+1}[CC^{\{j+1,j+1\}}]H_{j+1}^{\mathrm{T}}, \tag{8}$$

where $DD^{\{j,j\}}$ indicates the within-scale variance–covariance matrix of wavelet coefficients at scale j and $CD^{\{j,j\}}$ the within-scale variance–covariance matrix of scaling and wavelet coefficients at scale j. Moreover, the across-scales covariances between scales $j-1$ and j are

$$CD^{\{j-1,j\}} = H_jH_{j+1}[CC^{\{j+1,j+1\}}]G_{j+1}^{\mathrm{T}}, \tag{9}$$

$$DD^{\{j-1,j\}} = G_jH_{j+1}[CC^{\{j+1,j+1\}}]G_{j+1}^{\mathrm{T}}, \tag{10}$$

with $CD^{\{j-1,j\}}$ the across-scales variance–covariance matrix of scaling coefficients at scale $j-1$ and wavelet coefficients at scale j, and $DD^{\{j-1,j\}}$ the across-scales variance–covariance matrix of wavelet coefficients at scales $j-1$ and j.

These formulae can be applied to the matrix $CC^{\{j,j\}}$ to obtain the within-scale covariances at the coarser scale $j-1$ and the across-scales covariances between scales $j-2$ and $j-1$, and so on until a desired scale is reached. Fig. 1 shows the algorithm for the first two scales j and $j-1$. The resultant matrix is symmetric.

$CC^{\{j+1,j+1\}} \longrightarrow$

$CC^{\{j-1,j-1\}} = H_j[CC^{\{j,j\}}]H_j^{\mathrm{T}}$	$DC^{\{j-1,j-1\}} = H_j[CC^{\{j,j\}}]G_j^{\mathrm{T}}$	$CD^{\{j-1,j\}} = H_jH_{j+1}[CC^{\{j+1,j+1\}}]G_{j+1}^{\mathrm{T}}$
$[DC^{\{j-1,j-1\}}]^{\mathrm{T}}$	$DD^{\{j-1,j-1\}} = G_j[CC^{\{j,j\}}]G_j^{\mathrm{T}}$	$DD^{\{j-1,j\}} = G_jH_{j+1}[CC^{\{j+1,j+1\}}]G_{j+1}^{\mathrm{T}}$
$[CD^{\{j-1,j\}}]^{\mathrm{T}}$	$[DD^{\{j-1,j\}}]^{\mathrm{T}}$	$DD^{\{j,j\}} = G_{j+1}[CC^{\{j+1,j+1\}}]G_{j+1}^{\mathrm{T}}$

(a) (b)

Fig. 1. (a) Variance–covariance matrix of the vector $c^{(j+1)}$ and (b) variance–covariance matrix of ($c^{(j-1)}$, $d^{(j-1)}$, $d^{(j)}$) obtained from $CC^{\{j+1,j+1\}}$

$C \longrightarrow$

$H_j[B]H_j^{\mathrm{T}}$	$H_j[B]G_j^{\mathrm{T}}$	$H_{j+1}[C]G_{j+1}^{\mathrm{T}}$
$G_j[B]H_j^{\mathrm{T}}$	$G_j[B]G_j^{\mathrm{T}}$	
$G_{j+1}[C]H_{j+1}^{\mathrm{T}}$		$G_{j+1}[C]G_{j+1}^{\mathrm{T}}$

Fig. 2. Two-dimensional DWT ($B = H_{j+1}[C]H_{j+1}^{\mathrm{T}}$)

3.2. Link to the two-dimensional discrete wavelet transform

To understand further the algorithm described, consider the *two-dimensional DWT*. Given C, a $2^{j+1} \times 2^{j+1}$ matrix of pixels, a wavelet decomposition of the matrix can be calculated; this results in first applying the linear filters to the rows of the matrix C, obtaining two matrices $H_{j+1}[C]$ and $G_{j+1}[C]$, and then to the columns of $H_{j+1}[C]$ and $G_{j+1}[C]$, obtaining four matrices $H_{j+1}[C]H_{j+1}^{\mathrm{T}}$, $H_{j+1}[C]G_{j+1}^{\mathrm{T}}$, $G_{j+1}[C]H_{j+1}^{\mathrm{T}}$ and $G_{j+1}[C]G_{j+1}^{\mathrm{T}}$ of dimensions $2^j \times 2^j$. This procedure can be repeated with the matrix $B = H_{j+1}[C]H_{j+1}^{\mathrm{T}}$ (Fig. 2), and so on until a desired scale is reached.

Our recursive algorithm has an interesting link to the two-dimensional DWT. A comparison of Figs 1 and 2 shows that, having applied the two-dimensional DWT to the matrix $CC^{\{j+1,j+1\}}$, the diagonal blocks will correspond to the within-scale variance–covariance matrices; moreover, the across-scales variance–covariance matrices will be obtained by suitably applying the one-dimensional DWT to the rows of the non-diagonal blocks. Since $CC^{\{j+1,j+1\}}$ is a symmetric matrix, the matrices $G_{j+1}[CC^{\{j+1,j+1\}}]H_{j+1}^{\mathrm{T}}$ and $H_{j+1}[CC^{\{j+1,j+1\}}]G_{j+1}^{\mathrm{T}}$ are transposes of each other. This link to the two-dimensional DWT makes the implementation of our algorithm extremely simple in any of the available wavelet packages.

3.3. Further results

Let us now assume f a realization of a stochastic process $X(t)$ having finite (constant) mean $E[X(t)] = c$ (i.e. $c = 0$ without loss of generality) and finite $E[X(t)\,X(s)]$. The following result

states that, using minimum phase Daubechies (1992) wavelets, within- and across-scale covariance matrices have zero entries outside diagonal bands that depend on the number of vanishing moments of the wavelet family. Similar arguments can be made for different Daubechies wavelet families.

Proposition 2. Consider the Daubechies minimum phase wavelets. Then, at fixed integer scales j and j' with $j' - j = l$, where l is a positive integer, scaling coefficients $c_{j,k}$ and $c_{j',k'}$ are uncorrelated for integer k and k' such that $k' - 2^l k > 2^l(2N-1)$ or $k' - 2^l k < 1 - 2N$, and wavelet coefficients $d_{j,k}$ and $d_{j',k'}$ are uncorrelated for k and k' such that $k' - 2^l k > (1+2^l)N - 1$ or $k' - 2^l k < 2^l - (1+2^l)N$, where N is the number of vanishing moments of the wavelet family.

Proof. Note that

$$E[c_{j,k}\, c_{j',k'}] = E\left[\int X(t)\, \phi_{j,k}(t)\, \mathrm{d}t \int X(s)\, \phi_{j',k'}(s)\, \mathrm{d}s \right]$$

$$= \int \int E[X(t)\, X(s)]\, \phi_{j,k}(t)\, \phi_{j',k'}(s)\, \mathrm{d}s\, \mathrm{d}t.$$

Daubechies minimum phase wavelets satisfy

$$\mathrm{supp}(\phi_{j,k}) = \left[\frac{k}{2^j},\ \frac{2N-1+k}{2^j} \right]$$

and $E[c_{j,k}\, c_{j',k'}]$ will be 0 when $\phi_{j,k}$ and $\phi_{j',k'}$ have disjoint support, i.e. for $k' - 2^l k > 2^l(2N-1)$ or $k' - 2^l k < 1 - 2N$ with $j' - j = l$, and similarly for the $d_{j,k}$, considering that

$$\mathrm{supp}(\psi_{j,k}) = \left[\frac{1-N+k}{2^j},\ \frac{N+k}{2^j} \right]. \qquad \square$$

We remark here that more specific results can be derived if $X(t)$ is more precisely specified. The decay properties of the coefficients' correlation are now well known for a large variety of stationary and non-stationary stochastic processes. In the non-stationary case, for example, the coefficients' correlation structure has been investigated for the class of fractional Brownian motions, non-stationary zero-mean Gaussian processes with stationary increments: among others, Tewfik and Kim (1992) and Dijkerman and Mazumdar (1994) proved that the correlation between coefficients decreases exponentially fast across scales and hyperbolically fast through time. Flandrin (1992) studied the variance structure and showed that it is scale dependent.

4. Bayesian wavelet shrinkage

Recently, Bayesian approaches to wavelet shrinkage have attracted much attention in the literature. A common feature of the existing proposals is that the coefficient vector d in equation (5) is assumed to be a random variable and a prior distribution is imposed on it. The shrinkage step then becomes the result of deriving a Bayes rule from the posterior distribution of d. Vidakovic (1998), Chipman *et al*. (1997), Clyde *et al*. (1998) and Abramovich *et al*. (1998) explored different ways to specify the priors. All these contributions share the assumption that the components of d are independent.

The results of Section 3 can be brought to bear in developing algorithms for Bayesian wavelet shrinkage that do not rely on the assumption of independence. The first approach that takes into account the coefficients' correlation structure was proposed by Vidakovic and Müller (1995) in the context of density estimation and applied by Vannucci (1996) to denoise data. We shall refer to it as the *VM* model.

4.1. The Vidakovic–Müller model

According to equation (5), uncertainty about the empirical coefficients is described by

$$\tilde{d}\,|d, \sigma^2 \sim \mathcal{N}(d, \sigma^2 I). \tag{11}$$

A closed form learning procedure about the ds can be achieved (see, for example, O'Hagan (1994)), assuming

$$d, \sigma^2 \sim \mathcal{NIG}_{n+1}(\alpha, \delta, m, \Sigma) \tag{12}$$

where $\mathcal{NIG}_{n+1}$ denotes the multivariate normal–inverse gamma distribution with dimension $n+1$. The model requires the specification of the hyperparameters α, δ, m and Σ. Since the marginal prior distribution of σ^2 is an inverse gamma $\mathcal{IG}(\alpha/2, \delta/2)$ distribution, the parameters α and δ can be used to specify beliefs about the noise scale, considering that $E[\sigma^2] = \alpha/(\delta - 2)$ and

$$\mathrm{var}(\sigma^2) = \frac{2\alpha^2}{(\delta - 1)(\delta - 2)^2}.$$

Moreover, for a given σ^2, the prior conditional distribution of d given σ^2 is a multivariate normal distribution with mean vector m and covariance matrix $\sigma^2\Sigma$. The posterior distribution for d and σ^2 is

$$d, \sigma^2 | \tilde{d} \sim \mathcal{NIG}_{n+1}(\alpha^*, \delta^*, m^*, \Sigma^*), \tag{13}$$

with updated parameters

$$\begin{aligned} \Sigma^* &= (I + \Sigma^{-1})^{-1}, \\ m^* &= \Sigma^*(\tilde{d} + \Sigma^{-1} m) \end{aligned} \tag{14}$$

and $\alpha^* = \alpha + m^{\mathrm{T}}\Sigma^{-1}m + \tilde{d}^{\mathrm{T}}\tilde{d} + m^{*\prime}\Sigma^{*-1}m^*$ and $\delta^* = \delta + n$.

Vidakovic and Müller (1995) chose $m = 0$ and structured Σ as a block diagonal matrix, assuming that the components of d at different scales are independent, but allowing for coefficients at the same scale to be correlated. They suggested $\sigma_{k,k'} = \rho^{|k-k'|}$, $|\rho| < 1$, as entries for the diagonal blocks. This is a parsimonious and attractive choice which leads to a within-scale correlation structure that is inversely proportional to the distance between coefficients. Moreover, each diagonal block was multiplied by a scale-dependent parameter λ_j. They suggested the choice of an exponentially decreasing sequence of λs. This model specification and the adoption of a squared error loss function lead to the posterior mean $m^* = \Sigma^2\tilde{d}$ as a Bayes estimator of d. Thus, the IWT can be applied to m^* to obtain the function estimate. Note that this estimation procedure does not require the specification of α and δ.

4.2. Recursive specification of Σ

Here we use the results of Section 3 to motivate a different way for specifying Σ. We first

specify the covariance matrix $\Sigma_{J,\phi}$ of scaling coefficients at scale J and then derive the matrix Σ from $\Sigma_{J,\phi}$ according to the recursive algorithm described in Sections 3.1 and 3.2.

Section 3.3 provides insights about correlation structures for large classes of common processes that may generate f. In the Daubechies wavelets case, the results given in proposition 2 impose zero covariances outside a certain diagonal band which depends on the number of vanishing moments. Within the diagonal band, we may specify a covariance structure that decreases in inverse proportion to the distance between coefficients, as intuitively suggested by Vidakovic and Müller (1995). Thus, we propose specifying $\Sigma_{J,\phi}$ as $\sigma_{k,k'} = \lambda\rho^{|k-k'|}$ for $|k-k'| < 2N-1$ and $\sigma_{k,k'} = 0$ otherwise. Because of the recursive structure, we have within- and across-scale band covariance matrices at each scale.

Our specification of Σ has several additional features compared with the original proposal of Vidakovic and Müller (1995): it incorporates knowledge about the wavelet coefficients' correlation structure; it allows within- and across-scale correlation modelling; it leads to a remarkable reduction in the number of parameters implied in the model. The matrix $\Sigma_{J,\phi}$ (and hence Σ) depends in fact only on the two parameters λ and ρ, and on the basis of the construction above we can write $\Sigma(\lambda, \rho) = \lambda\,\Sigma(\rho)$. The parameter λ is a smoothing parameter. Smaller values of λ imply a more precise prior, i.e. greater shrinkage of the wavelet coefficients towards the prior mean m. Suitable values for ρ are those that imply a (semi)positive definite matrix $\Sigma_{J,\phi}$, i.e. a positive minimum eigenvalue. In Appendix A we find $|\rho| < C(N, n)$ where $C(N, n)$ is a constant depending on N, the wavelet number, and n, the dimension of the matrix. Exact results are given for $N = 1$, whereas numerical techniques are employed for larger values of N, given the complex structure of $\Sigma_{J,\phi}$.

As an illustration, Fig. 3 shows the covariance matrix Σ obtained for $J = 9$, $\lambda = 1$ and $\rho = 0.5$ using Daubechies minimum phase wavelets with six vanishing moments. Wavelet transforms have been applied with coarsest scale equal to 4. The scales are graphed from

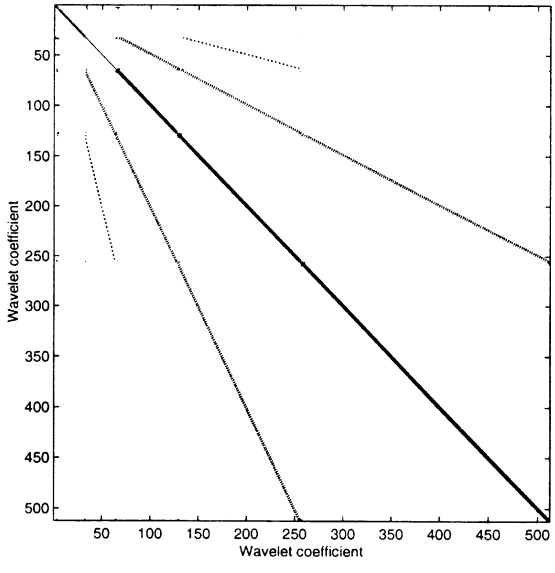

Fig. 3. Wavelet coefficients' covariance matrix $\Sigma(\lambda, \rho)$ with $\lambda = 1$ and $\rho = 0.5$ for Daubechies minimum phase wavelets with six vanishing moments: the highest grey scales values of the images correspond to the largest entries of the matrices; the coefficients are ordered from coarse to fine

coarse to fine. Plots were obtained by using the MATLAB (MathWorks, 1996) function `imagesc` that displays a matrix as an image. Each element of the matrix specifies the colour of a rectilinear patch in the image. The highest grey scale values of the image correspond to the largest entries of the matrix. Fig. 3 highlights the existence of non-zero covariances between coefficients at different scales.

5. Examples: simulated signals

To illustrate our Bayesian shrinkage strategy we use the functions *HeaviSine*, *Blocks*, *Bumps* and *Doppler* of Donoho and Johnstone (1994, 1995) as representative of signals with different characteristics which arise in several scientific fields. Fig. 4 shows the four signals (512 observations) and Fig. 5 the same signals corrupted by additive Gaussian white noise $N(0, \sigma^2)$ with signal-to-noise ratio $\mathrm{SNR} = \mathrm{sd}(f)/\sigma$ equal to 5.

We need to apply the DWT to the noisy data, to calculate the updated variance–covariance matrix Σ^* and posterior mean m^* as in equations (14) and to apply the inverse wavelet transform to m^* to obtain the smoothed data. We specify the vector m by centring wavelet coefficients on 0 and scaling coefficients at the coarsest scale on the empirical values. Our recursive specification of Σ requires choosing λ and ρ. Using an empirical Bayes approach, we search over a grid for the values that minimize a general measure of discrepancy. A suitable score, that measures the goodness of fit of the wavelet estimator, is the mean-squared error

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2.$$

A more realistic method, that learns about λ and ρ using a Bayesian hierarchical model, will be presented in the next section.

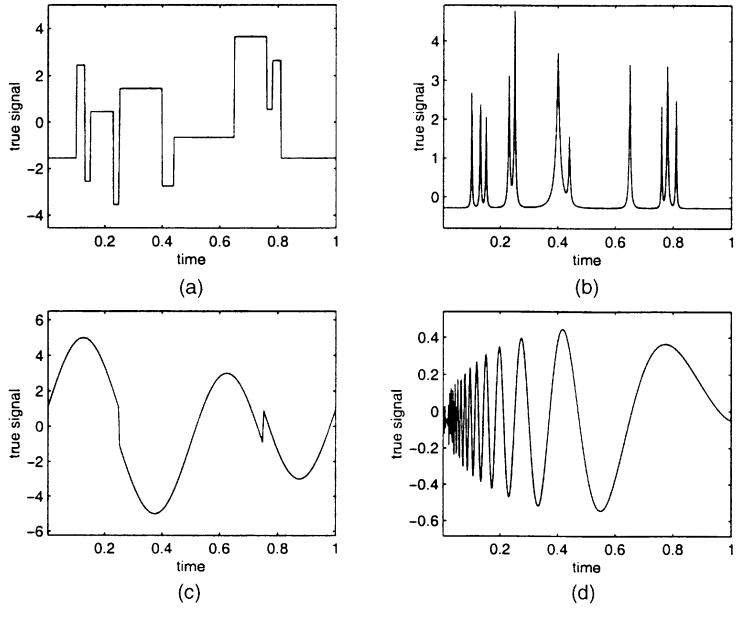

Fig. 4. Test signals (a) Blocks, (b) Bumps, (c) HeaviSine and (d) Doppler

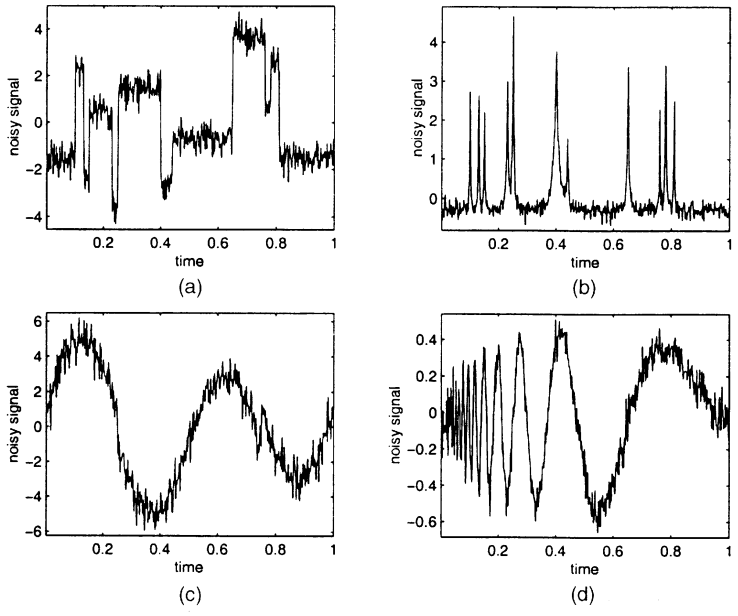

Fig. 5. Test signals corrupted by Gaussian white noise with signal-to-noise ratio SNR = sd(f)/σ equal to 5: (a) Blocks; (b) Bumps; (c) HeaviSine; (d) Doppler

Fig. 6 shows the smoothed signals. The Daubechies minimum phase wavelet family with seven vanishing moments was used for the HeaviSine, Doppler and Bumps signals. The Haar family was chosen for the Blocks signal. The selected values of λ and ρ were $(\lambda, \rho) = (1/3^2, 0.8)$ for the HeaviSine signal, $(\lambda, \rho) = (1/2, 0.5)$ for the Bumps signal, $(\lambda, \rho) = (1/3, 0.8)$ for the Doppler signal and $(\lambda, \rho) = (1/2^4, 0.5)$ for the Blocks signal. Among different values of the coarsest scale j_0 of the DWT, those that gave the best results were $j_0 = 6$ in the Bumps case and $j_0 = 5$ otherwise. In the assessment of j_0, cross-validation techniques may be used; alternatively, a fully Bayesian model could include uncertainty about j_0 over the small number of possible integer values.

To highlight the fact that allowing for across-scales correlation between coefficients gives a better reconstruction of the signals, we compare our results with those obtained by using the VM model in the original formulation of Vidakovic and Müller (1995). The sequence of λs was chosen as $[100, 2^{-j}\tau]$ with j ranging from the coarsest to the finest scale. We used $\tau = 2^4$ for the Bumps and Doppler signals and $\tau = 2^3$ otherwise. The value 100 is simply a large value that avoids the shrinkage of scaling coefficients at the coarsest scale. Numerical summaries can help in comparing the two methods. Ratios of the mean-squared errors

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2$$

of our estimates over those of Vidakovic and Müller were 0.8269 (HeaviSine), 0.8813 (Blocks), 0.4327 (Doppler) and 0.4116 (Bumps). Ratios of the mean absolute errors

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{f}_i - f_i|$$

were 0.9747 (HeaviSine), 0.9308 (Blocks), 0.7138 (Doppler) and 0.6569 (Bumps).

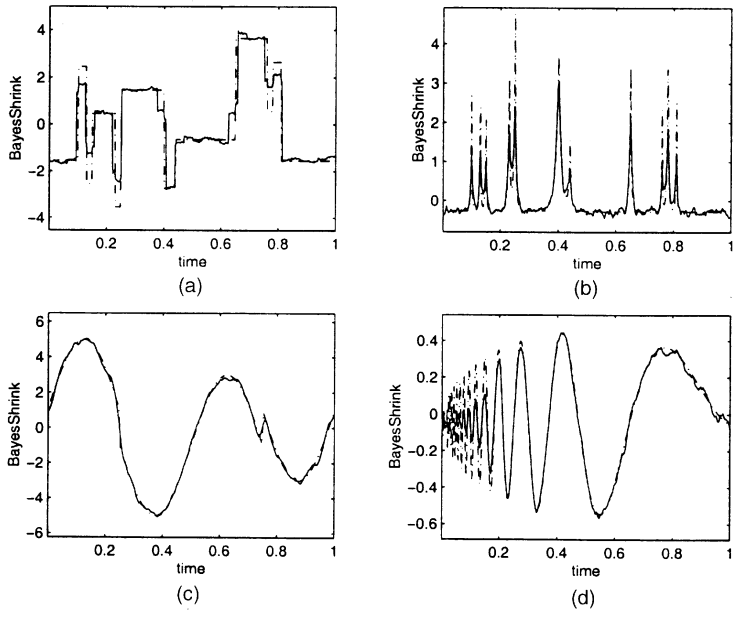

Fig. 6. Signals reconstructed by using the VM model with the proposed recursive specification of matrix Σ (the true signals are superimposed): (a) Blocks; (b) Bumps; (c) HeaviSine; (d) Doppler

6. Learning about λ and ρ

Because results can be sensitive to the choice of λ and ρ, it can be attractive to include them in the inferential process. A possible solution is to specify a third scale in the VM model. Empirical coefficients $\tilde{d}$ are still modelled through distribution (11) with prior distribution on d and σ^2 expressed in the conjugate form (12). Then, λ and ρ are also random and can be assumed independent and *a priori* distributed as

$$\lambda \sim \mathcal{IG}(p/2, q/2), \tag{15}$$

where $\mathcal{IG}$ denotes the inverse gamma distribution, and

$$p(\rho) \propto (C-\rho)^{r_1-1}(C+\rho)^{r_2-1}, \qquad |\rho| < C, \tag{16}$$

such that $(C-\rho)/2C$ is proportional to a beta distribution with parameters r_1 and r_2. The constant C takes into account the constraints on the support of ρ, as discussed in Section 4.2 and Appendix A.

6.1. The inference strategy

The complex structure of $\Sigma(\lambda, \rho)$ does not allow for inference in closed form. A sample from the posterior distribution of the parameters can be obtained by a Markov chain Monte Carlo method. The Metropolis–Hastings algorithm of Metropolis *et al.* (1953) and Hastings (1970) can be considered the archetype of this large variety of algorithms. Values are sampled from proposal distributions and accepted on the basis of suitable acceptance rules. The Gibbs sampler (see, for example, Tierney (1994) and Smith and Roberts (1993)) is a special case. Proposal distributions are represented by full conditional distributions of the parameters and sampled values are always accepted. Hybrid algorithms can be used when intractable full

conditional distributions arise for some of the parameters, as described by Chib and Greenberg (1994) and Müller (1992).

In our model we can easily calculate the full conditional distributions of d, σ^2 and λ, whereas it is more difficult to specify the full conditional distribution of ρ. Consequently, the chain is simulated by combining in a cycle Gibbs steps for the parameters d, σ^2 and λ with a Metropolis step for ρ. More precisely, given starting values for σ^2, λ and ρ, we sample the parameters in the following order. The vector d is simulated from

$$d|\tilde{d}, \sigma^2, \lambda, \rho \sim \mathcal{N}(m^*, \sigma^2\Sigma^*) \tag{17}$$

with $\Sigma^* = \{I + \lambda^{-1}\,\Sigma(\rho)^{-1}\}^{-1}$ and $m^* = \Sigma^*\{\tilde{d} + \lambda^{-1}\,\Sigma(\rho)^{-1}m\}$. This is done in a single step to exploit the correlation structure and to improve the speed of convergence. The variance noise σ^2 is simulated from

$$\sigma^2|\tilde{d}, d, \lambda, \rho \sim \mathcal{IG}(\alpha^*/2, \delta^*/2) \tag{18}$$

with $\alpha^* = \alpha + (d-m)^{\mathrm{T}}\lambda^{-1}\,\Sigma(\rho)^{-1}(d-m) + (\tilde{d}-d)^{\mathrm{T}}(\tilde{d}-d)$ and $\delta^* = \delta + 2n$. The parameter λ is simulated from

$$\lambda|d, \sigma^2, \rho \sim \mathcal{IG}(p^*/2, q^*/2) \tag{19}$$

with $p^* = p + \{(d-m)^{\mathrm{T}}\,\Sigma(\rho)^{-1}(d-m)\}/\sigma^2$ and $q^* = q + n$. Finally, choosing $\mathcal{N}(\hat{\rho}|\rho^{(j)}, \sigma_\rho^2)$, $\hat{\rho} < |C|$, as a proposal distribution, we simulate ρ by

(a) sampling $\hat{\rho}$ from $\mathcal{N}(\hat{\rho}|\rho^{(j)}, \sigma_\rho^2)$, $\hat{\rho} < |C|$, with $\rho^{(j)}$ the sample value generated from the previous cycle,

(b) computing

$$a = \min\left\{1,\ \frac{p(\hat{\rho}|d, \sigma^2, \lambda)}{p(\rho^{(j)}|d, \sigma^2, \lambda)}\right\},$$

(c) accepting $\hat{\rho}$ if $0 < U(0, 1) < a$.

After running the transient phase of the chain, the mean vector of d can be estimated by averaging over the simulated samples of the ds. Finally, the IWT can be applied to obtain the function estimate.

7. An example

We illustrate the performance of the hierarchical model of Section 6 on the Blocks signal.

To obtain *a posteriori* inferences we need to choose the hyperparameters m, α, δ, p, q, r_1 and r_2 and the starting values σ_0^2, λ_0 and ρ_0. We specify the mean vector m as described in Section 5. We assume ignorance about σ^2 by setting α and δ equal to 0 and we set σ_0^2 to the estimate suggested by Donoho and Johnstone (1994), i.e. the median absolute deviation of the wavelet coefficients at the finest scale, divided by a constant. We specify the prior distribution on λ so that the variability is large: we use $(p, q) = (1/32, 4)$ and starting value $\lambda_0 = 0.5$. Consider now ρ. Previous experience has shown that the positive part of the parameter space supports most of the probability mass. We use Haar wavelets, requiring $C = 0.5$; we set $(r_1, r_2) = (3, 12)$ and starting value $\rho_0 = 0.3$.

We simulated 256 observations and applied the DWT with coarsest scale equal to 5. A chain was run for 5000 iterations. Metropolis steps on ρ were performed using $\sigma_\rho = 0.005$, a value that turned out to be a good compromise between two needs: exploring the whole

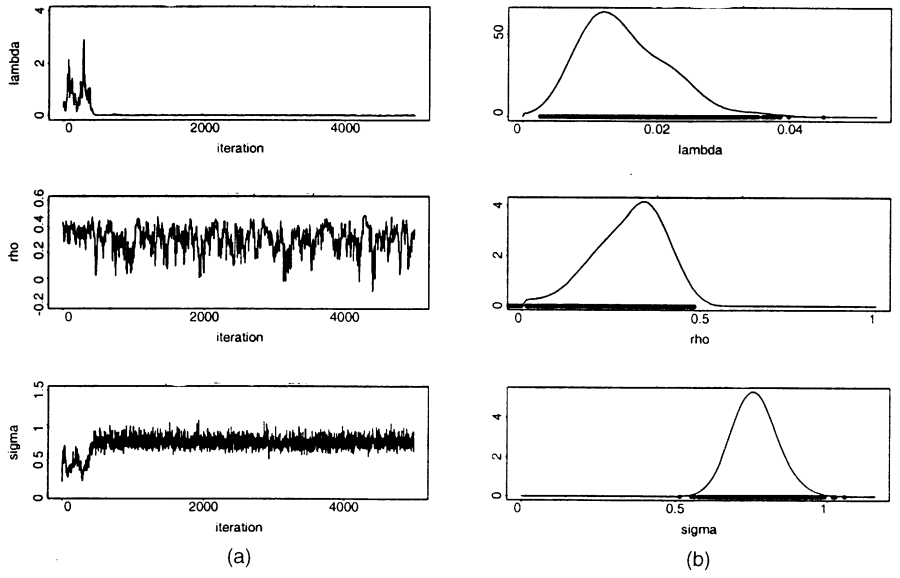

Fig. 7. Blocks signal–Markov chain Monte Carlo output analysis: (a) 5000 simulated values of the parameters λ, ρ and σ^2; (b) corresponding posterior density estimates (last 3000 values)

parameter space and obtaining an acceptance ratio of at least 40% (for the simulation presented here the exact acceptance ratio was 68%). Sampled values for the nuisance parameters σ^2, λ and ρ are shown in Fig. 7. The transient phase for these parameters seems to last for 400–500 iterations so the choice of a burn-in of 2000 iterations is conservative. Posterior density estimates are also given in Fig. 7. We produced the autocorrelation plots of values selected every 2, 5, 10 and 15 iterations of the last 3000 and decided to take one simulated value every 10 iterations. The autocorrelation becomes negligible after lag 4 for the λ- and ρ-parameters and at lag 2 for σ^2. To assess whether there was a lack of convergence we used some of the diagnostics implemented in the CODA software. The variables passed the Heidelberger and Welch (1983) stationary test. Furthermore, values of the Geweke (1992) Z-score diagnostic were -0.35 (λ), 1.07 (ρ) and 1.62 (σ), providing no evidence against convergence. In case problems in the convergence of the chain arise, modelling $\log(\rho)$, rather than ρ, may help.

Before evaluating the smoothing performance on the signal, it is interesting to consider the shrinkage on the wavelet coefficients. Fig. 8 compares the empirical coefficients with the average over the simulated values of the ds. As was expected, shrinkage has a very moderate effect on large coefficients whereas it reduces the smaller coefficients to values that are very close to 0. A histogram of the 217 coefficients that would have been set to 0 by using the SureShrink method of Donoho and Johnstone (1994, 1995) is also given in Fig. 8.

After averaging over the simulated values of the ds, we applied the IWT. Figs 9(a) and 9(b) respectively show the noisy signal and the smoothed signal superimposed on the original. Substantial smoothing has been performed.

8. Concluding remarks

We have investigated the correlation structure of wavelet coefficients and proposed a recursive algorithm to calculate within- and across-scale covariances. Focusing on Bayesian

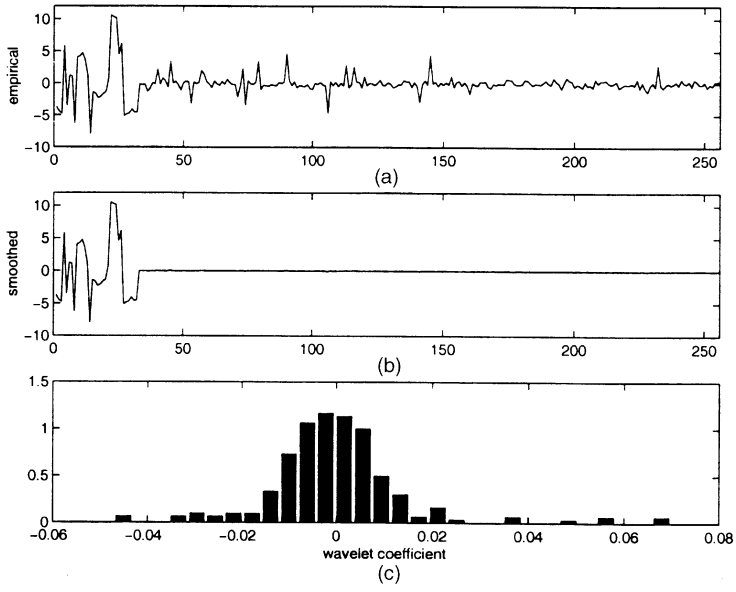

Fig. 8. Blocks signal: (a) empirical wavelet coefficients, ordered from coarse to fine; (b) averages of the simulated values of the *d*s; (c) histogram of the simulated values of the 217 coefficients that would have been set to 0 by SureShrink

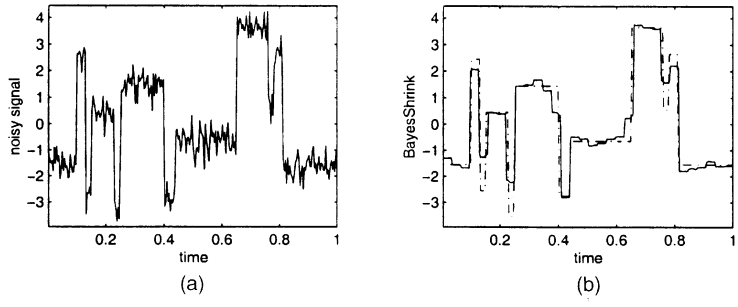

Fig. 9. Blocks signal: (a) noisy signal; (b) reconstructed signal superimposed on the original

wavelet shrinkage, we have used these findings to motivate a model specification leading to a parsimonious solution that depends only on two parameters. Inference on these parameters has been made by using Markov chain Monte Carlo methods.

The results of Section 3 can be viewed in the more general context of modelling wavelet coefficients which is implied in the nonparametric estimation of densities and regression functions involving wavelets. Kovac and Silverman (1998) have independently explored the recursive way of computing variances and covariances of wavelet coefficients. They concentrated only on variances and within-scale covariances and investigated the use of the algorithm in wavelet regression methods with irregularly spaced data, regularly spaced data sets of arbitrary size and correlated data. In the context of wavelet analysis of long memory processes, Vannucci *et al.* (1998) employed the variance recursive algorithm to derive Bayesian estimates of characteristic parameters of the process.

Different implementations of our model need to be explored. Notably, under study is the possibility of specifying two noise variance parameters for scaling and wavelet coefficients: this solution allows modelling of different levels of corruption of the two kinds of coefficients. Another possible improvement could derive from using a different way to model the dependence of the within-scale covariances on the distance between coefficients.

Acknowledgements

We thank Brani Vidakovic, Duke University, Antonio Moro, University of Florence, Alan B. Evans, University of Kent at Canterbury, and Bernard Silverman and Guy Nason, University of Bristol, for helpful discussions. Comments by the Associate Editor, one of the referees and by Giovanni Parmigiani, Duke University, greatly enhanced the paper. This work was partially supported by Consiglio Nazionale delle Ricerche and Fondi ex 40%. Part of this work was done while M. Vannucci was a research fellow at the University of Kent at Canterbury, supported under grant GK/K73343 from the Engineering and Physical Sciences Research Council. The MATLAB (MathWorks, 1996) toolbox *Wavbox 4.3* (Taswell, 1995) was used to calculate the wavelet transforms.

Appendix A

Here we discuss some properties of the eigenvalues of the matrix $\Sigma_{J_1+1,\phi}(\rho)$ defined as $\sigma_{k,k'} = \rho^{|k-k'|}$ for $|k-k'| < 2N-1$ and $\sigma_{k,k'} = 0$ otherwise. The parameter N represents the wavelet number, $N = 1, 2, \ldots$. Exact results can be proved for $N = 1$. Larger values of N lead to a very complex structure of the matrix and we therefore employ numerical techniques.

If $N = 1$ it is possible to prove (see, for example, Basilevsky (1983), p. 223) that the eigenvalues of the matrix $\Sigma_{J_1+1,\phi}(\rho)$ are

$$\lambda_k = 1 + 2\rho \cos\left(\frac{k\pi}{n+1}\right), \qquad k = 1, \ldots, n,$$

where n is the dimension of the matrix. The smallest eigenvalue is λ_n if $\rho > 0$ or λ_1 if $\rho < 0$. Thus $\Sigma_{J_1+1,\phi}(\rho)$ is positive definite for each n if $|\rho| < \frac{1}{2}$.

If $N > 1$, the eigenvalues have the analytical expression

$$\lambda_k = 1 + 2\left\{\rho \cos\left(\frac{k\pi}{n+1}\right) + \rho^2 \cos\left(\frac{2k\pi}{n+1}\right) + \ldots + \rho^N \cos\left(\frac{Nk\pi}{n+1}\right)\right\} \tag{20}$$

but to locate the smallest eigenvalue for this expression is far more difficult. Numerically, we found that, given N and n, the minimum eigenvalue is a monotonic decreasing function of ρ, $\rho \geqslant 0$. Thus, the highest value of ρ that gives a positive minimum eigenvalue can be found by using the bisection method. A similar procedure can be used when ρ is negative.

References

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

Basilevsky, A. (1983) *Applied Matrix Algebra in the Statistical Sciences*. Amsterdam: North-Holland.

Chib, S. and Greenberg, E. (1994) Understanding the Metropolis-Hastings algorithm. *Am. Statistn*, **49**, 327–335.

Chipman, H., Kolaczyk, E. and McCulloch, R. (1997) Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Ass.*, **92**, 1413–1421.

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

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

Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.
Dijkerman, R. and Mazumdar, R. (1994) On the correlation structure of the wavelet coefficients of fractional Brownian motion. *IEEE Trans. Inform. Theory*, **40**, 1609–1612.
Donoho, D. and Johnstone, I. (1994) Ideal spatial adaption via wavelet shrinkage. *Biometrika*, **81**, 425–455.
———(1995) Adapting to unknown smoothness by wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.
———(1998) Minimax estimation via wavelet shrinkage. *Ann. Statist.*, **26**, 879–921.
Donoho, D., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia (with discussion)? *J. R. Statist. Soc.* B, **57**, 301–369.
Flandrin, P. (1992) Wavelet analysis and synthesis of fractional Brownian motion. *IEEE Trans. Inform. Theory*, **38**, 910–917.
Geweke, J. (1992) Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 169–194. Oxford: Oxford University Press.
Hall, P. and Patil, P. (1995) Discussion on Wavelet shrinkage: asymptopia? (by D. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard)? *J. R. Statist. Soc.* B, **57**, 355.
Hastings, W. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
Heidelberger, P. and Welch, P. (1983) Simulation run length control in the presence of an initial transient. *Oper. Res.*, **31**, 1109–1144.
Johnstone, I. M. and Silverman, B. W. (1997) Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc.* B, **59**, 319–351.
Kovac, A. and Silverman, B. (1998) Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Technical Report 5/1998*. University of Dortmund, Dortmund.
Mallat, S. G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.*, **11**, 674–693.
MathWorks (1996) *MATLAB Version 5.0.0 .4064 on SOLZ*. Matick: MathWorks.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.
Meyer, Y. (1992) *Wavelets and Operators*. Cambridge: Cambridge University Press.
Müller, P. (1992) Alternatives to the Gibbs sampling scheme. *Technical Report 92-14*. Institute of Statistics and Decision Sciences, Duke University, Durham.
Nason, G. P. (1996) Wavelet shrinkage using cross-validation. *J. R. Statist. Soc.* B, **58**, 463–479.
O'Hagan, A. (1994) *Kendall's Advanced Theory of Statistics*, vol. 2B. Cambridge: Cambridge University Press.
Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.
Strang, G. (1989) Wavelets and dilation equations: a brief introduction. *SIAM Rev.*, **31**, 614–627.
Taswell, C. (1995) WavBox 4: a software toolbox for wavelet transforms and adaptive wavelet packet decompositions. In *Wavelets and Statistics* (eds A. Antoniadis and G. Oppenheim). New York: Springer.
Tewfik, A. and Kim, M. (1992) Correlation structure of the discrete wavelet coefficients of fractional Brownian motion. *IEEE Trans. Inform. Theory*, **38**, 904–909.
Tierney, L. (1994) Markov chains for exploring posterior distributions. *Ann. Statist.*, **22**, 1701–1728.
Vannucci, M. (1996) On the application of wavelets in statistics (in Italian). *PhD Thesis*. Dipartimento di Statistica G. Parenti, University of Florence, Florence.
Vannucci, M., Brown, P. and Fearn, T. (1998) Wavelet analysis of long-memory processes. *Technical Report 98-22*. University of Kent at Canterbury, Canterbury.
Vidakovic, B. (1998) Nonlinear wavelet shrinkage with Bayes rules and Bayes factor. *J. Am. Statist. Ass.*, **93**, 173–179.
Vidakovic, B. and Müller, P. (1995) Wavelet shrinkage with affine Bayes rules with applications. *Technical Report 95-34*. Institute of Statistics and Decision Sciences, Duke University, Durham.
Wang, Y. (1996) Function estimation via wavelet shrinkage for long memory data. *Ann. Statist.*, **24**, 466–484.