Some Findings on the Covariance Structure of Wavelet Coefficients: Theory and Models in a Bayesian Perspective

By

Marina Vannucci[1] and Fabio Corradi[2]

Report UKC/IMS/97/05, Institute of Maths & Stats, University of Kent at Canterbury, UK.

January 1997, revised November 1997

In this paper some theoretical results regarding the wavelet coefficients covariance structure between and within different scales are presented. These findings, generally applicable to the problem of modeling wavelet coefficients, are used here in the context of wavelet shrinkage where the purpose is to estimate a function from noisy data. The approach taken is Bayesian and uses an hierarchical structure. Applications to benchmark simulated signals are presented.

KEY WORDS: Nonparametric Regression, Wavelet Shrinkage, Wavelet Coefficients' Correlation Structure, Bayesian Inference.

1 Introduction

Wavelets are now well established in the literature as well as their usefulness in statistics as a powerful tool to the analysis and synthesis of the data. This paper explores the wavelet coefficients correlation structure. Our main contribution is the development of a recursive algorithm that leads to a well-specified variance-covariance structure between and within different scales. The given results seem to be generally applicable in every problem that involves wavelet coefficients' modeling. We use them in the context of wavelet shrinkage, a well known application of wavelets that attempts the recovery of a signal from noisy data. Originally proposed by Donoho and Johnstone (1994, 1995a, 1995b, 1997), wavelet shrinkage has been recently considered within a Bayesian framework where a prior distribution is imposed on the wavelet coefficients of the signal. While some of the existing contributions (see Chipman *et*

[1]Institute of Mathematics and Statistics, University of Kent at Canterbury, CT2 7NF, Canterbury, Kent, UK, *M.Vannucci@ukc.ac.uk*

[2]Dipartimento Statistico "G.Parenti", Università di Firenze, viale Morgagni 59, Firenze, 50134, Italy, *corradi@stat.ds.unifi.it*

al. (1997), Clyde *et al.* (1995) and Abramovich *et al.* (1998)) assume independent coefficients, we adopt a model firstly suggested by Vidakovic and Müller (1995) as an attempt to allow correlation. By employing the proposed recursive covariance structure we propose a specification of the model that depends on a parsimonious number of parameters. Inference is obtained via Markov Chain Monte Carlo methods. For a different approach that allows a partial correlation among coefficients see Crouse *et al.* (1998).

The paper is organized as follows: Basic concepts about orthonormal wavelet bases and wavelet transforms are briefly reviewed in Section 2 while wavelet shrinkage is described in Section 3. Section 4 states the results about the wavelet coefficients covariance structure. These results are merged in Section 5 with a Bayesian shrinking model. Applications to benchmark signals are given in Section 6. In Section 7 a hierarchical structure is introduced in order to make inference on the parameters of interest. Section 8 provides further examples and Section 9 some concluding remarks.

2 Basic Concepts

2.1 Orthonormal Wavelet Bases

Wavelets are families of basis functions able to accurately describe other functions in a parsimonious way. For a general exposition of the wavelet theory see Chui (1992), Daubechies (1992) and Meyer (1992). In $L^2(\mathbb{R})$, for example, a Daubechies wavelet basis is obtained as translations and dilations of two basis functions, a *scaling function* ϕ, constructed as a solution of a dilation equation

$$\phi(t) = \sqrt{2}\sum_l h_l \phi(2t-l), \tag{1}$$

and a *mother wavelet* ψ, defined from ϕ as $\psi(t) = \sqrt{2}\sum_l g_l \phi(2t-l)$. Filter coefficients h_l satisfy suitable conditions and $g_l = (-1)^l h_{1-l}$. 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 $L^2(\mathbb{R})$. Daubechies wavelets have compact support, maximum number of vanishing moments[3] and possess different degrees of smoothness. In contrast with the Fourier basis, wavelets are localized both in time and frequency/scale. Thus any $L^2(\mathbb{R})$ function f can be approximated by a wavelet series as $f(t) = \sum_{j,k\, \in \mathbb{Z}} d_{j,k}\psi_{j,k}(t)$ with wavelet coefficients defined as $d_{j,k} = \langle f, \psi_{j,k}\rangle = \int f(t)\psi_{j,k}(t)dt$ that describe features of the function f at the spatial location $2^{-j}k$ and frequency 2^j(or scale j).

[3]A function f has N vanishing moments if $\int t^q f(t)dt = 0, \quad q = 0, 1, \ldots, N-1$

2.2 The Discrete Wavelet Transform

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

$$c_{j,k} = \sum_m h_{m-2k} c_{j+1,m}, \quad d_{j,k} = \sum_m g_{m-2k} c_{j+1,m}. \tag{2}$$

The above equations are convolutions followed by a downsampling operation. Adopting signal processing terminology, let H indicate a *low pass* filter capturing low frequency components, G an *high pass* filter capturing high frequency components and D_0 the downsampling operator that chooses every other element of a sequence. With $c^{(j)}$ and $d^{(j)}$ indicating the coefficient vectors at level j equations (2) can be expressed as

$$c^{(j)} = H_{j+1} c^{(j+1)}, \quad d^{(j)} = G_{j+1} c^{(j+1)} \tag{3}$$

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 $j+1$ points out that the dimensions of the matrices change with the level, due to the downsampling operation.

Equations (2) are used in wavelet theory to derive a fast algorithm for decomposing a function into a set of wavelet coefficients. A vector of $n = 2^J$ (J positive integer) observations at equally spaced points is viewed as an approximation of the function at the scale J and relations (2) are applied to the vector getting scaling and wavelet coefficients at the coarser level $J-1$. Wavelet coefficients are kept and equations (2) applied again to the scaling coefficients, and so on until a certain level. This algorithm is known as *Discrete Wavelet Transform* (DWT).

The algorithm for the inverse construction is called the *Inverse Wavelet Transform* (IWT) and reconstructs the function from the coefficients.

DWT and IWT have operation count proportional to n. See Strang (1989) for an exhaustive exposition of these algorithms.

3 Wavelet Shrinkage

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

$$y_i = f(t_i) + \epsilon_i, \tag{4}$$

with f a function to be estimated, t_i, $i = 1, \ldots, n$, equally spaced points and ϵ_i, $i = 1, \ldots, n$, i.i.d. normal random variables with zero mean and variance σ^2. Wavelet shrinkage (see Donoho and Johnstone (1994, 1995a, 1995b, 1997)) is a technique consisting of three main steps: Firstly the discrete wavelet transform (DWT) is applied to the data obtaining a vector $\hat{d}$ of empirical wavelet coefficients. Since the DWT is linear and orthogonal, $\hat{d}$ can still be modeled as

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

where d is the vector of coefficients of the unknown function f and $\hat{\epsilon}$ is a Gaussian white noise with mean zero and variance-covariance matrix $\sigma^2 I$. Secondly the noise is suppressed from the coefficients by using a thresholding (and/or shrinking) method. Donoho and Johnstone suggested an *hard threshold* policy, $\hat{d}^S = \hat{d} I(|\hat{d}| \geq \lambda)$, with I the indicator function, and a *soft threshold* policy, $\hat{d}^S = \operatorname{sign}(\hat{d})(\hat{d} - \lambda)_+$, with $x_+ = \max(x, 0)$. Finally the inverse wavelet transform is applied to $\hat{d}^S$ leading to an estimate of the unknown function.

The key of successful shrinkage is the choice of the parameter λ. Donoho and Johnstone (1994, 1995a) proposed the *universal* threshold, known also as *VisuShrink*, $\lambda^U = \hat{\sigma}\sqrt{2 \log n}$, and a level-dependent threshold obtained by minimizing the Stein's unbiased estimator of the risk (*SureShrink* threshold). Moreover, they suggested an hybrid method which selects the SureShrink threshold or the universal threshold according to the sparseness of the wavelet coefficients. Donoho and Johnstone (1995b, 1997) 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 (1994) and Wang (1996) adjusted the well known cross-validation criterion to choose the threshold.

A good overview of statistical applications of the wavelet shrinkage can be found in Donoho (1993).

4 Wavelet Coefficients' Covariances

4.1 A Recursive Structure

Here we prove a general result about the wavelet coefficients covariances using the recursive relations (2).

Proposition 4.1 *Given a wavelet basis in $L^2(\mathbb{R})$, the following results hold*

$$i) \quad Cov[d_{j,k}d_{j',k'}] = \sum_m \sum_n g_{m-2k}g_{n-2k'}Cov[c_{j+1,m}c_{j'+1,n}]$$

$$ii) \quad Cov[c_{j,k}c_{j',k'}] = \sum_m \sum_n h_{m-2k}h_{n-2k'}Cov[c_{j+1,m}c_{j'+1,n}]$$

$$iii) \quad Cov[c_{j,k}d_{j',k'}] = \sum_m \sum_n h_{m-2k}g_{n-2k'}Cov[c_{j+1,m}c_{j'+1,n}]$$

Proof. Straightforward given the relations (2).

The results $i), ii)$ and $iii)$ given in the above proposition provide a useful recursive algorithm that can be used to calculate the *within- and across-level* coefficients covariances. To clarify the terminology, we call within-level the covariances between coefficients that belong to a same level and across-levels those between coefficients that belong to different levels. Let us assume that the within-level variance-covariance matrix of scaling coefficients $c^{(j+1)}$ at the level $(j+1)$ is known and indicate that matrix as $CC^{\{j+1,j+1\}}$. The within-level covariances at the coarser level j can be easily computed as

$$\mathrm{Cov}[d_{j,k}d_{j,k'}] = \sum_m \sum_n g_{m-2k}g_{n-2k'}\mathrm{Cov}[c_{j+1,m}c_{j+1,n}] \tag{6}$$

$$\mathrm{Cov}[c_{j,k}c_{j,k'}] = \sum_m \sum_n h_{m-2k}h_{n-2k'}\mathrm{Cov}[c_{j+1,m}c_{j+1,n}] \tag{7}$$

$$\mathrm{Cov}[c_{j,k}d_{j,k'}] = \sum_m \sum_n h_{m-2k}g_{n-2k'}\mathrm{Cov}[c_{j+1,m}c_{j+1,n}]. \tag{8}$$

The above equations can be readily expressed in filter notation (see equation (3)) as $DD^{\{j,j\}} = G_{j+1}[CC^{\{j+1,j+1\}}]G^T_{j+1}$, $CC^{\{j,j\}} = H_{j+1}[CC^{\{j+1,j+1\}}]H^T_{j+1}$ and $CD^{\{j,j\}} = G_{j+1}[CC^{\{j+1,j+1\}}]H^T_{j+1}$ where $DD^{\{j,j\}}$ indicates the within-level variance-covariance matrix of wavelet coefficients at the level j and $CD^{\{j,j\}}$ the within-level variance-covariance matrix of scaling and wavelet coefficients at the level j.

Moreover, the across-levels covariances between the levels $j-1$ and j are

$$\mathrm{Cov}[c_{j-1,k}d_{j,k'}] = \sum_m \sum_n \sum_l h_{m-2k}g_{n-2k'}h_{l-2m}\mathrm{Cov}[c_{j+1,l}c_{j+1,n}] \tag{9}$$

and

$$\mathrm{Cov}[d_{j-1,k}d_{j,k'}] = \sum_m \sum_n \sum_l g_{m-2k}g_{n-2k'}h_{l-2m}\mathrm{Cov}[c_{j+1,l}c_{j+1,n}] \tag{10}$$

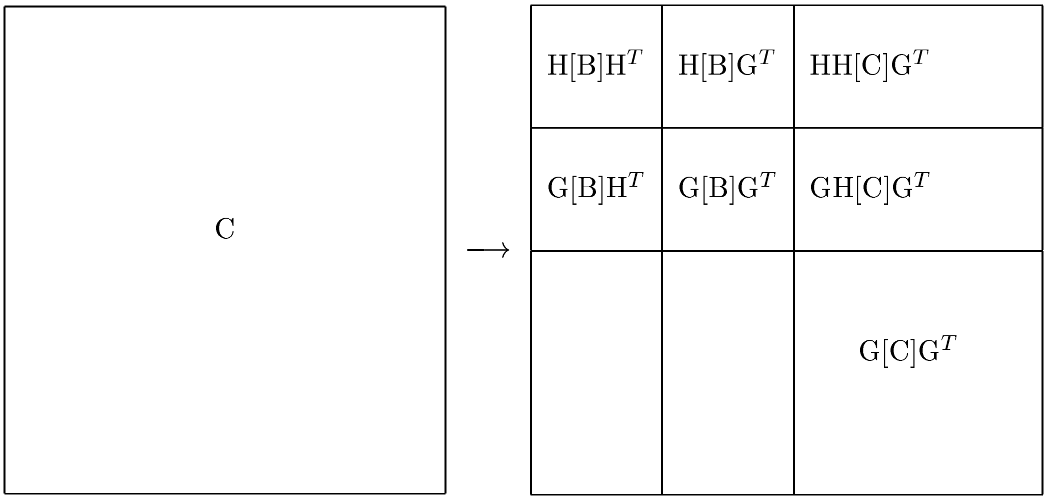

Figure 1: *Left side*: Variance-covariance matrix of the vector $c^{(j+1)}$. *Right side*: Variance-covariance matrix of $(c^{(j-1)}, d^{(j-1)}, d^{(j)})$ obtained by C (B stands for $H[C]H^T$).

that is, in filter notation, $CD^{\{j-1,j\}} = H_j H_{j+1}[CC^{\{j+1,j+1\}}]G_{j+1}^T$ and $DD^{\{j-1,j\}} = G_j H_{j+1}[CC^{\{j+1,j+1\}}]G_{j+1}^T$ with $CD^{\{j-1,j\}}$ the across-levels variance-covariance matrix of scaling coefficients at the level $j-1$ and wavelet coefficients at the level j and $DD^{\{j-1,j\}}$ the across-levels variance-covariance matrix of wavelet coefficients at the levels $j-1$ and j.

Clearly the above formulas can be applied to the matrix $CC^{\{j,j\}}$ to get the within covariances at the coarser level $j-1$ and the across covariances between levels $j-2$ and $j-1$, and so on until a certain level. Figure 1 shows the algorithm for the first two levels j and $j-1$ (the matrix $CC^{\{j+1,j+1\}}$ is indicated with C and the filter index dropped to simplify the notation).

A nice way to understand the described algorithm is to consider the *two dimensional discrete wavelet transform* (DWT2). 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 applying the linear filters firstly to the rows of the matrix C, obtaining two matrices $H_{j+1}[C]$ and $G_{j+1}[C]$ and secondly to the columns of $H_{j+1}[C]$ and $G_{j+1}[C]$ obtaining four matrices $H_{j+1}[C]H_{j+1}^T$, $H_{j+1}[C]G_{j+1}^T$, $G_{j+1}[C]H_{j+1}^T$ and $G_{j+1}[C]G_{j+1}^T$ of dimensions $2^j \times 2^j$. The procedure can be repeated with the matrix $B = H_{j+1}[C]H_{j+1}^T$ (see Figure 2), and so on until a certain level. A comparison of Figures 1 and 2 shows that having applied the two-dimensional discrete wavelet transform to the matrix $CC^{\{j+1,j+1\}}$,

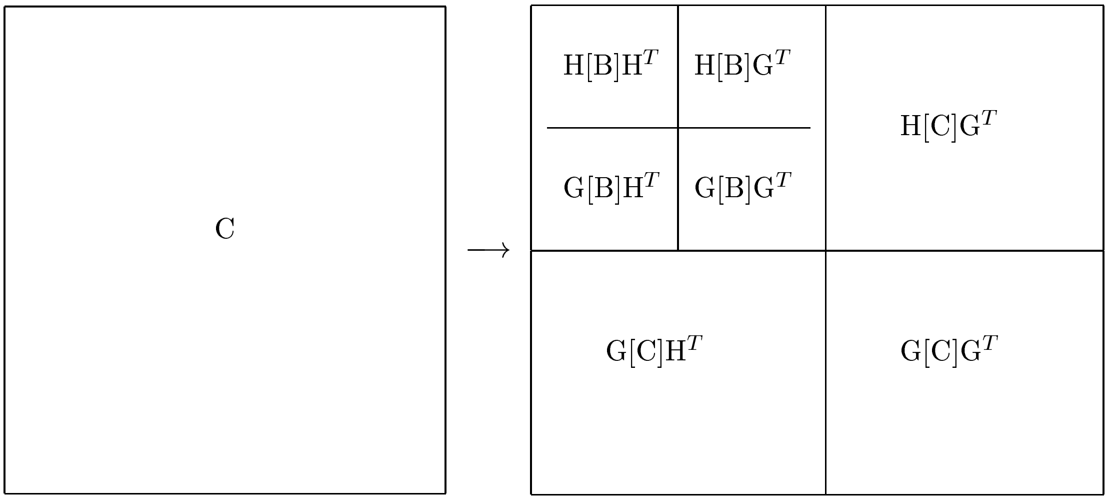

Figure 2: Two-dimensional discrete wavelet transform ($B = H[C]H^T$).

the diagonal blocks will correspond to the within-level variance-covariance matrices; moreover, the across-level variance-covariance matrices will be obtained by suitably applying the one-dimensional DWT to the rows of the non diagonal blocks. Notice that, since $CC^{\{j+1,j+1\}}$ is a symmetric matrix, the matrices $G_{j+1}[CC^{\{j+1,j+1\}}]H_{j+1}^T$ and $H_{j+1}[CC^{\{j+1,j+1\}}]G_{j+1}^T$ are transpose's of each other.

4.2 Further Results

Under mild assumptions more indications about the scaling and wavelet coefficients covariance structure can be given. Let's firstly assume that f in (4) is a realization of a random process $X(t)$ having finite $E[X(t)]$ and $E[X(t)X(s)]$. The following result uses minimum phase Daubechies wavelets (see Daubechies (1992)) but similar arguments can be made for different Daubechies wavelet families.

Proposition 4.2 *Let's consider the Daubechies minimum phase wavelets. Then, at fixed scales j and j' with $j' - j = l$, scaling coefficients $c_{j,k}$ and $c_{j',k'}$ are uncorrelated for k and k' such that*

$$(k' - 2^l k) > 2^l(2N - 1) \quad or \quad (k' - 2^l k) < 1 - 2N$$

and wavelet coefficients $d_{j,k}$ and $d_{j',k'}$ are uncorrelated for k and k' such that

$$(k' - 2^l k) > (1 + 2^l)N - 1 \quad or \quad (k' - 2^l k) < 2^l - (1 + 2^l)N$$

with N the number of vanishing moments of the wavelet family.

Proof. Notice that

$$\begin{aligned} E[c_{j,k}c_{j\prime,k\prime}] &= E\Big[\int X(t)\phi_{j,k}(t)dt \int X(s)\phi_{j',k'}(s)ds\Big] \\ &= \int\int E[X(t)X(s)]\phi_{j,k}(t)\phi_{j',k'}(s)dsdt. \end{aligned}$$

Daubechies minimum phase wavelets satisfy $Supp(\phi_{j,k}) = \left[\frac{k}{2^j}, \frac{2N-1+k}{2^j}\right]$ and $E[c_{j,k}c_{j',k'}]$ will be zero when $\phi_{j,k}$ and $\phi_{j',k'}$ have disjoint support, that is for $(k'-2^lk) > 2^l(2N-1)$ or $(k'-2^lk) < 1-2N$ with $j'-j=l$.

Similarly for the $d_{j,k}$. $\triangle$

More specific results can be provided if $X(t)$ is more precisely specified. For example if we assume wide-sense stationarity, that is $E[X(t)X(s)] = R(t-s)$, we have the following:

Proposition 4.3 *If $\hat{R}(\omega)$, $\hat{\psi}(\omega)$ and $\hat{\phi}(\omega)$ are C^p, $p>1$, then at fixed scale j*

$$E[c_{j,k}c_{j,k'}] = O(|k'-k|^{-p})$$

$$E[d_{j,k}d_{j,k'}] = O(|k'-k|^{-p})$$

that is the within-level correlation structure decreases with the distance between the time-localizations of the coefficients.

Proof. By considering the spectral representation of the covariance function we have

$$E[c_{j,k}c_{j',k'}] = \frac{2^{-j-l/2}}{2\pi}\int \hat{R}(\omega)\hat{\phi}(2^{-j}\omega)\hat{\phi}(2^{-j+l}\omega)e^{i\omega(2^{-l}k'-k)2^{-j}}d\omega$$

with $j'-j=l$. If $j'=j$

$$E[c_{j,k}c_{j,k'}] = \frac{2^{-j}}{2\pi}\int \hat{R}(\omega)|\hat{\phi}(2^{-j}\omega)|^2e^{i\omega(k'-k)}d\omega$$

and if $\hat{R}(\omega)$ and $\hat{\phi}(\omega)$ are C^p, $p>1$, we can conclude that for fixed j

$$E[c_{j,k}c_{j,k'}] = O(|k'-k|^{-p}).$$

Similarly for the $d_{j,k}$. $\triangle$

Remark 4.1: In the nonstationary case the coefficients correlation structure has been investigated for the class of fractional Brownian motions, nonstationary 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 along time. Flandrin (1992) studied the variance structure showing that it is scale dependent.

5 Bayesian wavelet shrinkage

Recently, Bayesian approaches to wavelet shrinkage have attracted much attention in the literature. A common feature of those proposals is that the coefficients vector d in (5) is assumed to be random and a prior distribution is imposed on it. The shrinking step becomes then the result of applying some Bayes rules to the posterior distribution of d. Vidakovic (1994), Chipman *et al.* (1997), Clyde *at al.* (1995) and Abramovich *et al.* (1998) explore different ways to specify the priors. At the meantime all the contributions share the assumption that the components of d are independent.

A first attempt to specify a model 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 a time series of IBM stock prices. We will refer to it as the $V\&M$ model.

5.1 The $V\&M$ Model

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

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

A closed form learning procedure about the d's can be achieved (see for example O'Hagan) assuming

$$[d, \sigma^2] \sim \mathcal{NIG}(\alpha, \delta, m, \Sigma) \tag{12}$$

where $\mathcal{NIG}$ stands for the multivariate normal-inverse gamma distribution.

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 $\mathcal{IG}(\alpha/2, \delta/2)$, the parameters α and δ allow for the specification of the belief about the noise considering that $E[\sigma^2] = \frac{\alpha}{\delta-2}$ and $Var(\sigma^2) = \frac{2\alpha^2}{(\delta-4)(\delta-2)^2}$. Moreover, for a given σ^2, the prior conditional distribution of d given σ^2 is a multivariate normal with mean vector m and covariance matrix $\sigma^2\Sigma$.

The posterior distribution for d and σ^2 is

$$[d, \sigma^2|\hat{d}] \sim \mathcal{NIG}(\alpha^\star, \delta^\star, m^\star, \Sigma^\star) \tag{13}$$

with updated parameters

$$\Sigma^\star = (I + \Sigma^{-1})^{-1}, \quad m^\star = \Sigma^\star(\hat{d} + \Sigma^{-1}m) \tag{14}$$

and $\alpha^\star = \alpha + m^T \Sigma^{-1} m + \hat{d}^T \hat{d} + (m^\star)^T (\Sigma^\star)^{-1} m^\star, \quad \delta^\star = \delta + n.$

Vidakovic and Müller (1995) chose $m = 0$ and structured Σ as a diagonal block matrix assuming that the components of d at different levels were independent but allowing coefficients at the same levels to be correlated. They suggested a parsimonious but sensible $\sigma_{k,k'} = \rho^{|k-k'|}$, $|\rho| < 1$ as entries for the diagonal blocks leading to a within-level correlation structure in inverse proportion to the distance $d(k,k') = |k - k'|$ between coefficients. Moreover each diagonal block was multiplied by a level-dependent parameter λ_j. The authors suggested to choose a sequence of $\lambda' s$ exponentially decreasing. In a slightly different context Müller and Vidakovic (1995) assume $\lambda_j = \tau 2^{-j}$.

This particular model specification and the adoption of an L^2-loss allow the use of the posterior mean $m^\star = \Sigma^\star \hat{d}$ as a Bayes estimator of d. Thus, the IWT can be applied to $m^\star$ to get the function estimate. Notice that the estimate procedure does not require the α and δ parameters specification.

5.2 A Recursive Specification of Σ

Here we use the results of Section 4 in order to propose a different way to specify Σ.

- First we specify a covariance matrix $\Sigma_{J,\phi}$ of scaling coefficients at the level J. In the Daubechies wavelets case, given N the number of vanishing moments, the results given in proposition 4.2 impose zero covariances outside a certain diagonal band which depends on N. Within the diagonal band we specify a covariance structure that decreases in inverse proportion to the distance between coefficients. According to the results of proposition 4.3 and remark 4.1, this structure, intuitively suggested by Müller and Vidakovic (1995), seems to be suitable for large classes of common processes that may originate f. Thus, we specify $\Sigma_{J,\phi}$ as
$$\begin{aligned} \sigma_{k,k'} &= \lambda \rho^{|k-k'|} \quad \text{for} \quad |k-k'| < 2N-1 \\ \sigma_{k,k'} &= 0 \qquad\qquad \text{otherwise.} \end{aligned}$$

- Then we derive the matrix Σ from $\Sigma_{J,\phi}$ by using the recursive relations given in proposition 4.1 according to the algorithm described in Section 4.1. Notice that, due to the recursive structure, we have within- and across-level band covariance matrices at each scale.

The proposed specification of Σ has a number of features not shared by the original proposal of Vidakovic and Müller (1995): it uses fundamental properties of wavelet coefficients; it allows within- and across-level correlation modeling; it leads to a remarkable reduction of the number of the parameters implied in the model. The matrix Σ depends in fact only on the two parameters λ and ρ, and based on the construction above it can be written as $\Sigma(\lambda, \rho) = \lambda\Sigma(\rho)$. The parameter λ is a smoothing parameter. Smaller values of λ imply a more precise prior, that is wavelet coefficients more shrunk towards the prior mean m. Suitable values for ρ are those that imply a (semi) positive definite matrix $\Sigma_{J,\phi}$, that is 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.

As an illustration, Figure 3 shows the covariance matrix Σ obtained for $J = 9$, $\lambda = 1$ and $\rho = 0.5$ using Haar wavelets and Daubechies minimum phase wavelets with 6 vanishing moments. Wavelet transforms have been applied with coarsest level equal to 4. That figure enlightens the existence of nonzero covariances between coefficients at different levels.

6 Examples: Simulated Signals

To illustrate the Bayesian shrinkage we use the functions *HeaviSine, Blocks, Bumps* and *Doppler* (Donoho and Johnstone (1994, 1995a)) representing signals with different characteristics that arise in several scientific fields. Figure 4 shows the four signals (512 observations) and Figure 5 the same signals corrupted by an additive Gaussian white noise $N(0, \sigma^2)$ with signal-to-noise ratio $SNR = sd(f)/\sigma$ equal to 5. The mean was subtracted from the data.

What needs to be done is to apply the discrete wavelet transform to the noisy data, calculate the updated variance-covariance matrix Σ^* and posterior mean m^* as in (14) and apply the inverse wavelet transform to m^* to get the smoothed data. We specify the vector m by centering wavelet coefficients on zero and scaling coefficients at the coarsest level on the empirical values. The recursive specification of Σ we adopt requires the choice of λ and ρ. We search over a grid the values that minimize a general measure of discrepancy. A suitable score, that measures the goodness-of-fit of the wavelet estimator, is $\mathrm{MSE}(\hat{f}, f) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_i - f_i)^2$. A somehow more realistic method, that "learns" about λ and ρ using a Bayesian hierarchical model, will be presented in the following section.

Figure 6 shows the smoothed signals. The Daubechies minimum phase wavelet family with 7 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 resulted to be $(\lambda, \rho) = (1/3^2, 0.8)$ for the *HeaviSine* signal, $(\lambda, \rho) = (1/2, 0.5)$

for the *Bumps*, $(\lambda, \rho) = (1/3, 0.8)$ for the *Doppler* and $(\lambda, \rho) = (1/2^4, 0.5)$ for the *Blocks*. The discrete wavelet transform was applied with coarsest level equal to 6 in the *Bumps* case and 5 otherwise.

Figure 7 shows the shrinkage obtained using the $V\&M$ model in the original formulation of Vidakovic and Müller (1995) (see Section 5.1). The sequence of λ's was chosen as $[100, \tau 2^{-j}]$ with j ranging from the coarsest to the finest level. 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 shrinking of scaling coefficients at the coarsest scale.

7 Learning about λ and ρ

Given the dependence of the results on the λ and ρ parameters, it seems essential to include them in the inferential process. A possible solution is to specify a third level in the $V\&M$ model. Empirical coefficients $\hat{d}$ are still modeled through (11) with prior distribution on d and σ^2 expressed in the conjugate form (12). Then, λ and ρ can be assumed independent and a priori distributed as follows:

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

and

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

such that $(C - \rho)/2C$ is proportional to a $\mathcal{B}eta(r_1, r_2)$. The constant C takes into account the constraints on the support of ρ, as explained in Section 5.2 and Appendix A.

7.1 The inference strategy

The complex structure of $\Sigma(\lambda, \rho)$ does not allow inference in closed form but a sample of the parameters posterior distribution can be obtained running a Markov Chain (see, for example, Tierney (1994) and Smith and Roberts (1993)). In the proposed model we can easily calculate the full conditional distributions of d, σ^2 and λ, while it seems not possible to specify the full conditional distribution of ρ. Consequently, the chain is simulated combining in a cycle Gibbs steps for the parameters d, σ^2 and λ with a Metropolis step for ρ. The possibility of using such an hybrid algorithms is detailed in Chib and Greenberg (1994) and Chan and Geyer (1994). 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

$$[d|\hat{d}, \sigma^2, \lambda, \rho] \sim \mathcal{N}(m^\star, \sigma^2\Sigma^\star) \tag{17}$$

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

$$[\sigma^2|\hat{d}, d, \lambda, \rho] \sim \mathcal{IG}(\alpha^\star/2, \delta^\star/2) \tag{18}$$

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

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

with $p^\star = p + [(d - m)^T(\Sigma(\rho))^{-1}(d - m)]/\sigma^2$ and $q^\star = q + n$. Finally, choosing $N(\hat{\rho}|\rho^{(j)}, \sigma^2_\rho)$, $\hat{\rho} < |C|$ as a jump distribution, we simulate ρ by
1) sampling $\hat{\rho}$ from $N(\hat{\rho}|\rho^{(j)}, \sigma^2_\rho)$, $\hat{\rho} < |C|$, with $\rho^{(j)}$ the sample value generated from the previous cycle;
2) computing

$$a = min\Big[1, \frac{p(\hat{\rho}|d, \sigma^2, \lambda)}{p(\rho^{(j)}|d, \sigma^2, \lambda)}\Big];$$

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

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

8 An Example

We illustrate the performance of the hierarchical model (12), (15) and (16) on the Blocks signal.

In order to run the chain we need to choose the hyperparameters m, α, δ, p, q, r_1 and r_2 and the starting values σ^2_0, λ_0 and ρ_0. We specify the mean vector m as described in Section 6. We assume *ignorance* on σ^2 setting α and δ equal to zero and choose as initial value σ^2_0 the estimate suggested by Donoho and Johnstone (1994) as the median absolute deviation of the wavelet coefficients at the finest level, divided by a constant. We specify the prior distribution on λ following the rationale of giving a mode equal to 1 (the indifference point) with a large amount of variance: among other possible choices 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 the main part of the probability, so we specify the prior distribution giving a not negligible probability at the edge of the parameter space and a very little probability for values of ρ smaller than zero. 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 level equal to 5. A chain was run for 5000 iterations. Metropolis steps on ρ were performed using $\sigma_\rho = 0.005$, a value resulted to be a good compromise between two needs: exploring the whole parameter space and obtaining not less then 40% of the sampled values accepted. The convergence was performed using the Heidelberger and Welch test (1983) implemented in CODA[4] software that suggested to discard the first 2000 iterations. A visual inspection of the reached convergence for the nuisance (but obviously crucial) parameters σ^2, λ and ρ can be assessed looking at Figure 8. The transient phase for these parameters seems to last for 4-500 iterations so that the choice of a burn-in of 2000 iterations can be considered conservative. In the same figure posterior density estimates are given. In order to reduce the correlation among simulated values we produced the autocorrelation plots of values selected every 2, 5, 10 iterations of the last 3000 and decided to take one simulated value every 10 iterations (see Figure 9).

Before evaluating the smoothing performance on the signal, it can be interesting to have a look at the shrinking effect on the wavelet coefficients. Figure 10 compares the empirical coefficients with the average over the simulated values of the d's. As it was expected, shrinkage has very moderate effect on large coefficients while reduces the smaller to values very close to zero. In the same figure an histogram of the 217 coefficients that would have been set to zero using the *SureShrink* method (Donoho and Johnstone (1994, 1995a)) is given.

After averaging over the simulated values of the d's, the IWT can be applied. Figures 11a) and 11b) show the noisy and smoothed signals superimposed to the original. A substantial smoothing has been performed. For a more precise idea of the noise reduction, Figures 11c) and 11d) show histograms of the the original data subtracted from the corrupted and the denoised.

[4]Copyright MRC Biostatistics Unit 1995

9 Concluding Remarks

We have explored the correlation structure of wavelet coefficients proposing a recursive algorithm to calculate within- and across-level covariances. Focusing on Bayesian wavelet shrinkage, we have used these findings to drive the specification of the model leading to a parsimonious solution that depends only on two parameters. Inference on these parameters has been made with the aid of MCMC methods. The results of Section 4 can be seen in the more general context of modeling wavelet coefficients which is implied in the non parametric estimation of densities and regression functions involving wavelets.

Some possible different implementations of the model could be explored. Notably, under study is the possibility to specify two noise variance parameters for scaling and wavelet coefficients: this solution could allow to verify the existence of a different level 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-level covariances on the distance among coefficients: some possibilities are under investigation.

Acknowledgments

The first author is grateful to Brani Vidakovic, Duke University, for introducing her to the Bayesian approach to the wavelet shrinkage, to Antonio Moro, University of Florence, for the continuous support and to Bernard Silverman and Guy Nason, Bristol University, for stimulating and helpful suggestions. Discussions with Alan B. Evans, University of Kent at Canterbury, are gratefully acknowledged. This work was partially supported by CNR, Consiglio Nazionale delle Ricerche, and Fondi ex 40% "Contributi attuali della metodologia statistica all'econometria". M. Vannucci is a research fellow at the University of Kent at Canterbury, supported under EPSRC Grant Number GK/K73343. Programs have been written in Matlab[5]. The toolbox *Wavbox 4.3* (Copyright Carl Taswell, (1995)) was used to calculate the wavelet transforms.

APPENDIX A

We state here some results about 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, $N = 1, 2, \ldots$.

In the case $N = 1$ it is possible to prove that the eigenvalues of the matrix $\Sigma_{j_1+1,\phi}(\rho)$ are (see for example Basilevsky (1983) page 223) $\lambda_k = 1 + 2\rho\cos(\frac{k\pi}{n+1}),\ k =$

[5]The MathWorks, Inc, Natick, Mass.,USA

$1, \ldots, n$ where n is the dimension of the matrix. Thus, the lowest eigenvalue is λ_n if $\rho > 0$ or λ_1 if $\rho < 0$ and $\Sigma_{j_1+1,\phi}(\rho)$ is positive definite for each n if $|\rho| < 1/2$.

In the general case $N > 1$, the eigenvalues have the analytical expression

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

but to locate the lowest eigenvalue for this expression is far harder. Numerically, we found that, given N and n, the minimum eigenvalue, as a function of ρ, $\rho \geq 0$, is a monotonic decreasing function. 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 for ρ negative.

References

Abramovich, F., Sapatinas, T. and Silverman, B.W. (1998). Wavelet thresholding via a bayesian approach. *Journal of the Royal Statistical Society, Series* B; (to appear).

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

Chan, R.S. and Geyer, C.J. (1994). Discussion of the paper " markov chain for exploring posterior distribution", by l.tierney. *Annals of Statistics*, **22,** 1747–1758.

Chib, S. and Greenberg, E. (1994). Understanding the metropolis-hastings algorithm. *American Statistician*, **49,** 327–335.

Chipman, H.A., Kolaczyk, E.D. and McCulloch, R.E. (1997). Adaptive bayesian wavelet shrinkage. *Journal of the American Statistical Association*, **92,**; (to appear).

Chui, K. (1992). *An Introduction to Wavelets.* Academic Press.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1995). Multiple shrinkage and subset selection in wavelets. Technical Report DP 95-37. ISDS, Duke University.

Crouse, M.S., Nowak, R.D. and Baraniuk, R.G. (1998). Wavelet-based signal processing using hidden markov models. *IEEE Transactions on Signal Processing (Special Issue on Wavelets and Filterbanks)*; (to appear).

Daubechies, I. (1992). *Ten Lectures on Wavelets*, volume 61. SIAM, CBMS-NSF Conference Series.

Dijkerman, R.W. and Mazumdar, R.R. (1994). On the correlation structure of the wavelet coefficients of fractional brownian motion. *IEEE Transactions on Information Theory*, **40(5),** 1609–1612.

Donoho, D. (1993). Nonlinear wavelet methods for recovery of signals, densities and spectra from indirect and noisy data. In *Proceedings of Symposia in Applied Mathematics*, volume 47, pp. 173–205. American Mathematical Society.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaption via wavelet shrinkage. *Biometrika*, **81,** 425–455.

Donoho, D. and Johnstone, I. (1995a). Adapting to unknown smoothness by wavelet shrinkage. *Journal of the American Statistical Association*, **90**.

Donoho, D. and Johnstone, I. (1997). Minimax estimation via wavelet shrinkage. *Annals of Statistics*, **25**.

Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995b). Wavelet shrinkage: Asymptopia? *J. R. Statis. Soc.*, **57(2),** 301–369.

Flandrin, P. (1992). Wavelet analysis and synthesis of fractional brownian motion. *IEEE Transactions on Information Theory*, **38(2)**.

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, **31,** 1109–1144.

Johnstone, I.M. and Silverman, B.W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series* B, **59(2),** 319–353.

Mallat, S.G. (1989a). Multiresolution approximations and wavelet orthonormal bases of $l_2(r)$. *Transactions of the American Mathematical Society*, **315(1)**.

Mallat, S.G. (1989b). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11(7),** 674–693.

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

Müller, P. and Vidakovic, B. (1995). Bayesian inference with wavelets: Density estimations. Technical Report DP 95-33. ISDS, Duke University.

Nason, G.P. (1994). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society, Series* B, **58,** 463–479.

Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society, Series* B, **55,** 3–23.

Strang, G. (1989). Wavelets and dilation equations: a brief introduction. *SIAM Review*, **31(4),** 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). Springer Verlag.

Tewfik, A.H. and Kim, M. (1992). Correlation structure of the discrete wavelet coefficients of fractional brownian motion. *IEEE Transactions on Information Theory*, **38(2)**.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, **22,** 1701–1728.

Vannucci, M. (1996). On the application of wavelets in statistics. Ph.D. Thesis. Dipartimento di Statistica G.Parenti, University of Florence, Italy. In italian.

Vidakovic, B. (1994). Nonlinear wavelet shrinkage with bayes rules and bayes factor. Technical Report DP 94-24. ISDS, Duke University.

Vidakovic, B. and Müller, P. (1995). Wavelet shrinkage with affine bayes rules with applications. Technical Report DP 95-34. ISDS, Duke University.

Wang, Y. (1996). Function estimation via wavelet shrinkage for long memory data. *Annals of Statistics*, **24(2),** 466–484.

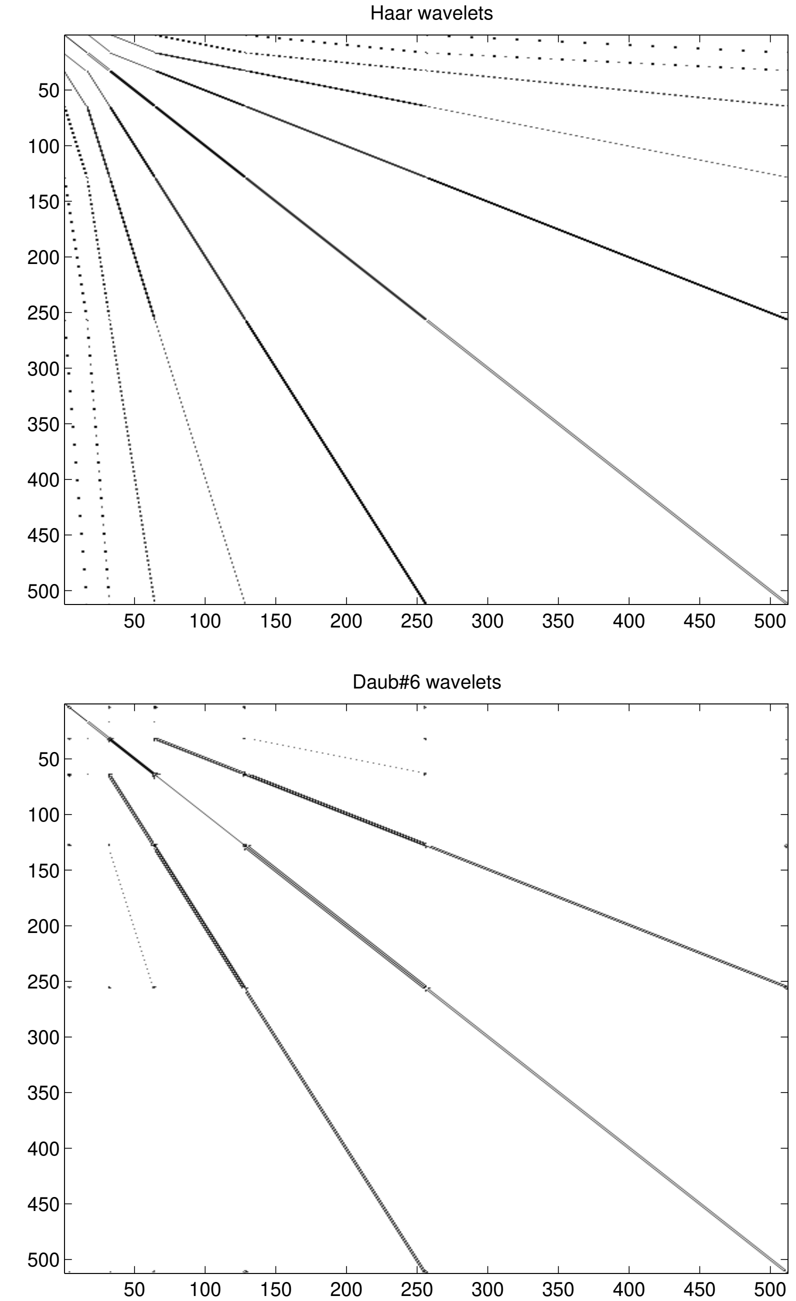

Figure 3: Wavelet coefficients' covariance matrices $\Sigma(\lambda, \rho)$ with $\lambda = 1$ and $\rho = 0.5$ for different wavelet families. Highest gray-scales values of the images correspond to largest entries of the matrices. Levels are represented from coarser to finer.

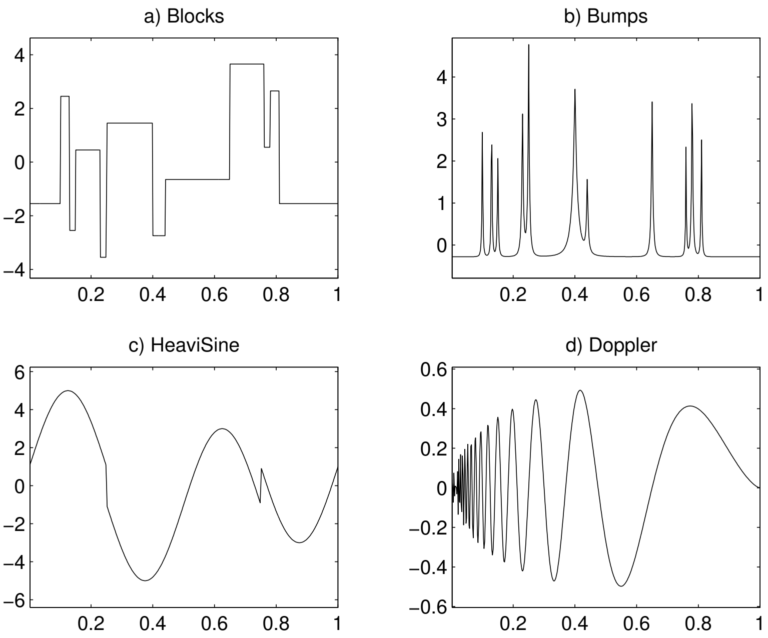

Figure 4: The four test signals.

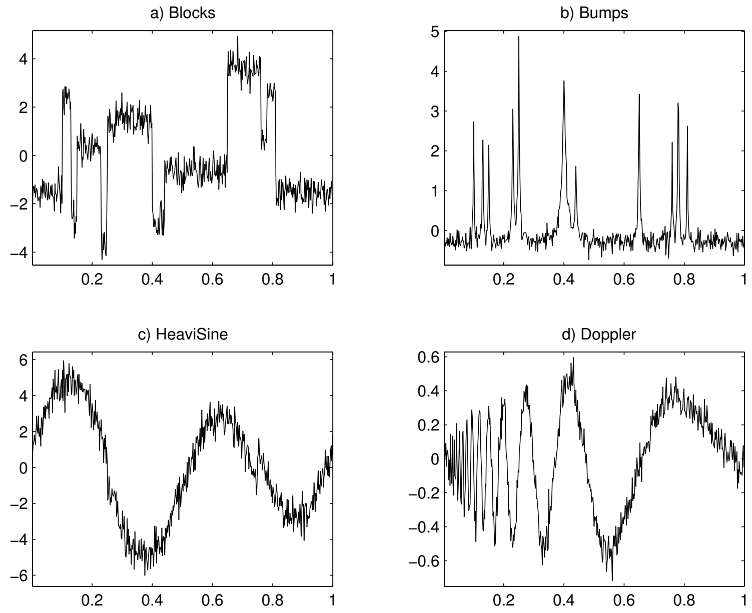

Figure 5: The four signals corrupted by a Gaussian white noise with signal-to-noise ratio $SNR = sd(f)/\sigma$ equal to 5.

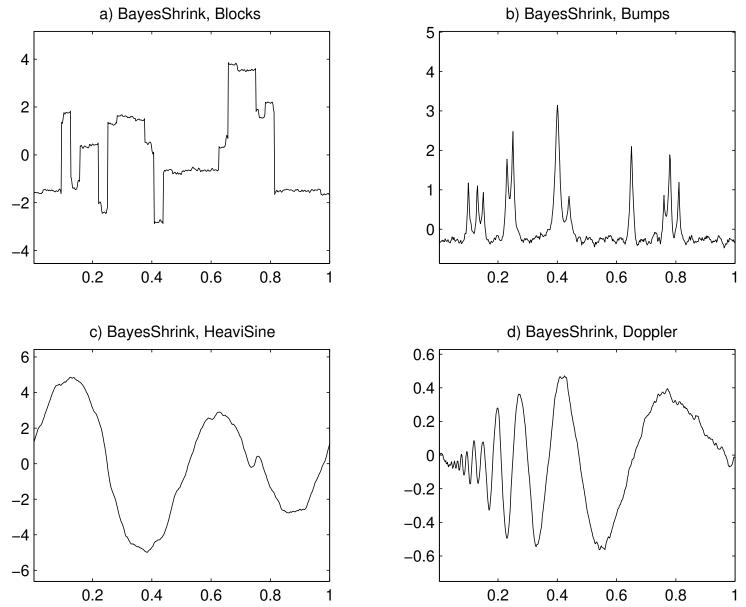

Figure 6: The four signals reconstructed using the $V\&M$ model with the proposed recursive specification of the matrix Σ.

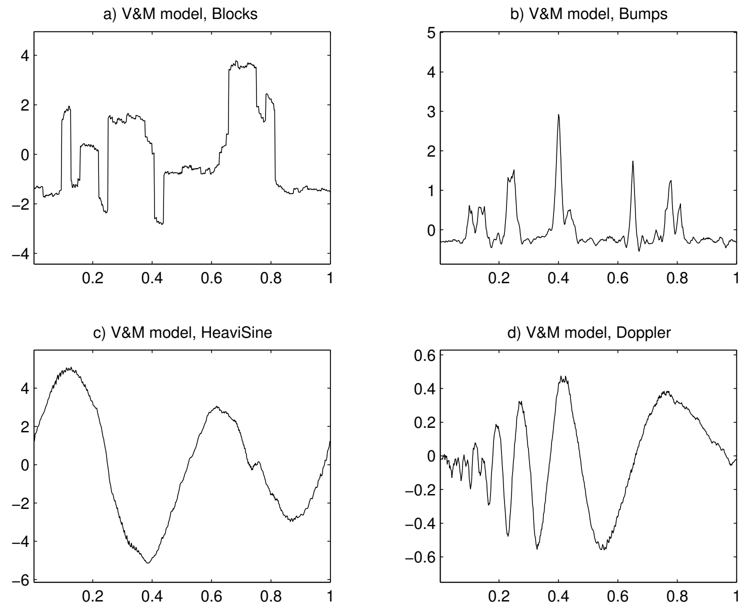

Figure 7: The four signals reconstructed using the $V\&M$ model in the original formulation.

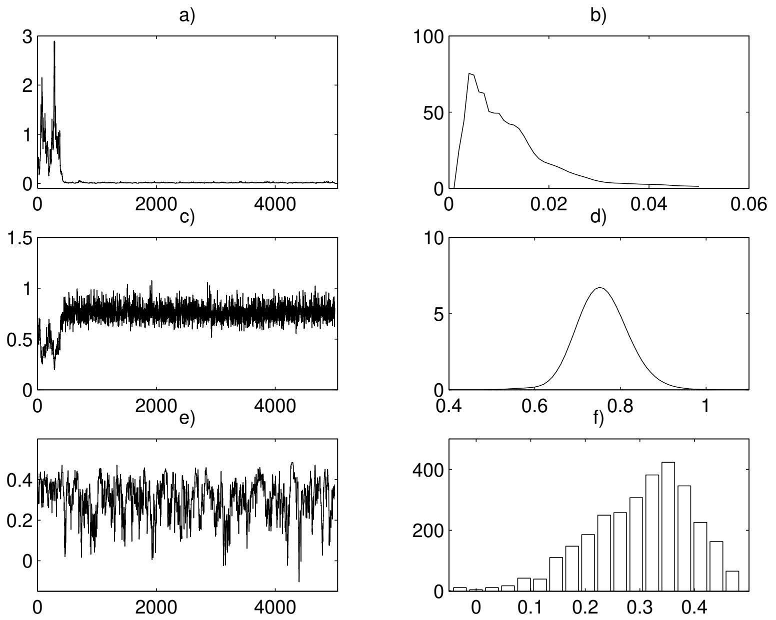

Figure 8: Blocks signal. *First column*: 5000 simulated values of the parameters λ (plot a)), σ^2 (plot c)) and ρ (plot e)). *Second column*: Corresponding posterior density estimates.

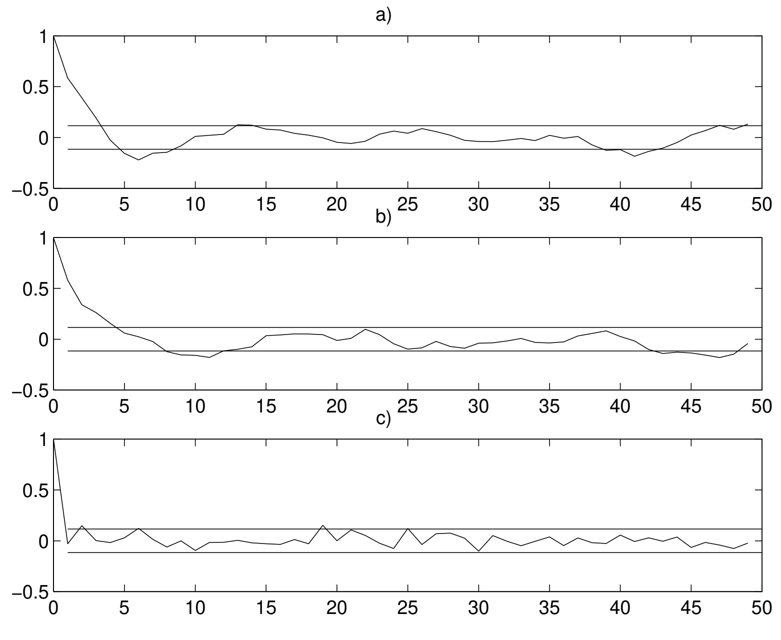

Figure 9: Blocks signal. Autocorrelation functions with thinning interval equal 10 of the last 3000 simulated values of λ (plot a)), ρ (plot b)) and σ^2 (plot c)).

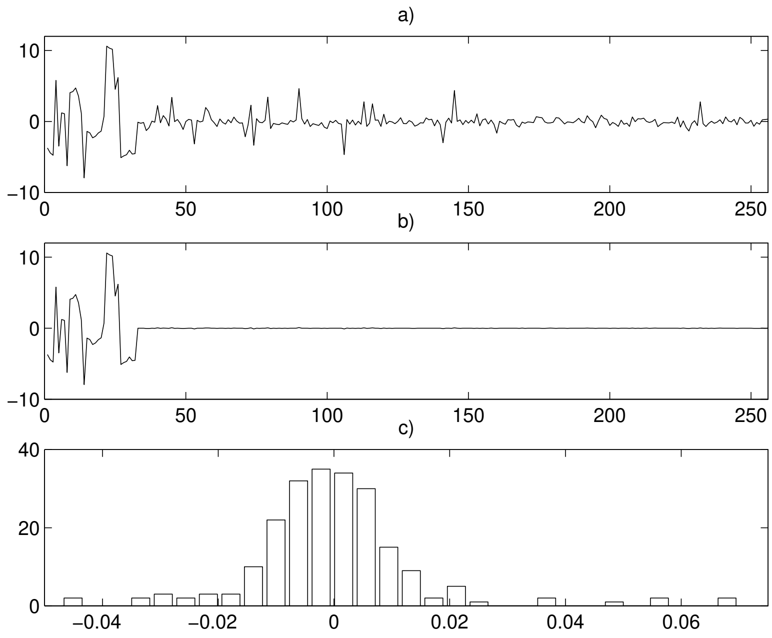

Figure 10: Blocks signal. Plot a): Empirical coefficients. Plot b): Means of the simulated values of the d's. Plot c): Histogram of the simulated values of the 217 coefficients that would have been set to zero with *SureShrink*.

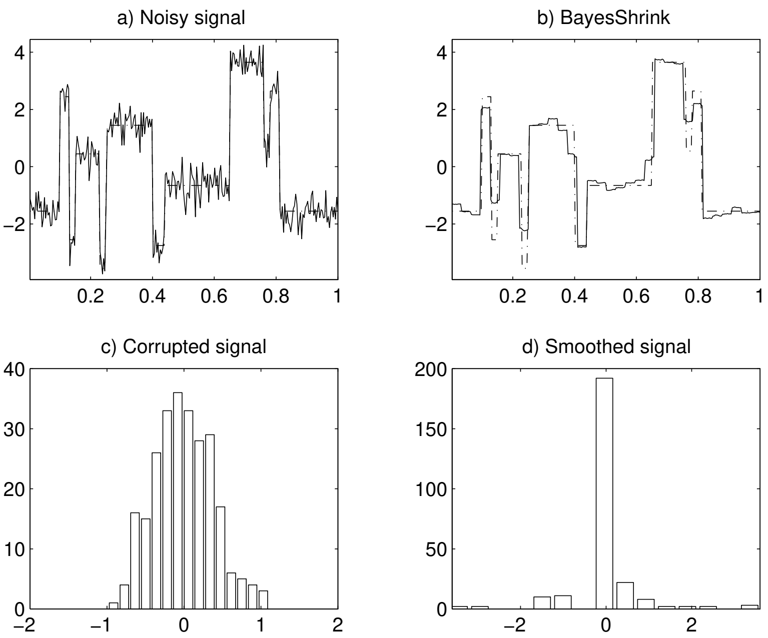

Figure 11: Blocks signal. Plot a): Noisy signal superimposed to the original. Plot b): Smoothed signal superimposed to the original. Plot c): Histogram of the original data subtracted from the corrupted. Plot d): Histogram of the original data subtracted from the smoothed.