Spectrum Estimation by Wavelet Thresholding of Multitaper Estimators

A. T. Walden[1], D. B. Percival[2] & E. J. McCoy[1]

Abstract – Current methods for power spectrum estimation by wavelet thresholding use the empirical wavelet coefficients derived from the log periodogram. Unfortunately, the periodogram is a very poor estimate when the true spectrum has a high dynamic range and/or is rapidly varying. Also, because the distribution of the log periodogram is markedly non-Gaussian, special wavelet-dependent thresholding schemes are needed. These difficulties can be bypassed by starting with a multitaper spectrum estimator. The logarithm of this estimator is close to Gaussian distributed if a moderate number (≥ 5) of tapers are used. In contrast to the log periodogram, log multitaper estimates are not approximately pairwise uncorrelated at the Fourier frequencies, but the form of the correlation can be accurately and simply approximated. For scale-independent thresholding the correlation acts in accordance with the wavelet shrinkage paradigm to suppress small scale 'noise spikes' while leaving informative coarse scale coefficients relatively unattenuated. This simple approach to spectrum estimation is demonstrated to work very well in practice. Additionally, the progression of the variance of wavelet coefficients with scale can be accurately calculated, allowing the use of scale-dependent thresholds. This more involved approach also works well in practice, but is not uniformly preferable to the scale-independent approach.

EDICS number: 3.1.1, 2.4.4

[1] Dept. of Mathematics, Imperial College of Science, Technology & Medicine, London

[2] Applied Physics Laboratory, University of Washington & MathSoft, Inc., Seattle, WA

Correspondence to Dr. Donald B. Percival, Applied Physics Laboratory, Box 355640, University of Washington, Seattle, WA 98195–5640; Phone: (206)–543–1368; Fax: (206)–543–6785; E-mail: dbp@apl.washington.edu

I. INTRODUCTION

Let $\{X_t\}$ be a discrete, real-valued stationary process with power spectrum (spectral density) S. Given a time series that can be regarded as a realization of a finite portion $X_1, \ldots, X_N$ of this process, we consider the problem of estimating S. A large number of nonparametric and parametric spectrum estimators has been studied as solutions to this problem (see, e.g., [16,18,19] and references therein), but, due to the rich variety of spectra found in practical applications, estimators are usually designed for a restricted subclass of spectra and can exhibit poor performance on spectra outside of that subclass. Recently there has been interest in tackling spectrum estimation by using wavelet thresholding techniques to produce a smooth estimate of the logarithm of the power spectrum. The appeal of wavelet thresholding is that it has been shown to have desirable minimax properties for function estimation over a broad classes of functions and to automatically adapt to the smoothness properties of the underlying function [3,4,9]. In the context of spectrum estimation, the basic idea, as discussed by Gao [6,7] and Moulin [14], consists of four basic steps:

(i) Calculate the logarithm of the periodogram $\frac{1}{N}\left|\sum_{t=1}^{N} X_t e^{-i2\pi ft}\right|^2$ at $f = j/N$, $j = 1, \ldots, N/2$.

(ii) Apply a standard, periodic, partial discrete wavelet transform (DWT) out to level q_0 to the log periodogram ordinates, where q_0 is specified in advance.

(iii) Apply a thresholding procedure to the resulting empirical wavelet coefficients (leaving the remaining empirical scaling coefficients entirely alone).

(iv) Invert the partial DWT, producing a hopefully smooth estimate of the log spectrum.

Clearly, without (iii), this procedure would leave the log periodogram unmodified. Thresholding is thus a key step and is based on representing the log periodogram as a 'signal' plus 'noise,' where the signal is the true spectrum. Hence, the procedure falls in the class of wavelet shrinkage estimates for noisy data: when the noise has a Gaussian distribution, Donoho and Johnstone [3] proposed a simple level-independent 'universal' threshold.

There are three main problems with the above procedure. First, the noise is log χ_2^2 distributed and hence markedly non-Gaussian so that one cannot use simple Gaussian-based thresholds in the wavelet-shrinkage scheme. This problem has been addressed in Gao [6,7] and Moulin [14] by developing special threshold levels that depend on scale and wavelet. Second, independent of wavelet thresholding, the periodogram can be a very poor spectrum estimate because of substantial bias due to sidelobe leakage. Third, use of the standard partial DWT out to level q_0 with the periodogram requires the restrictive assumption that the sample size N is a multiple of 2^{q_0}.

To eliminate problems due to the periodogram, we must start with a better log spectrum estimator. A well-known procedure for alleviating bias due to leakage in the periodogram is to apply a data taper (window) to the time series prior to computing the spectrum estimator, resulting in a so-called direct spectrum estimator. A direct spectrum estimator has better small sample bias properties than the periodogram, but, when subsequently smoothed across frequencies, has an asymptotic variance larger than that of a smoothed periodogram. The idea behind the multitaper spectrum estimator is to reduce this variance by computing a small number K of direct spectrum estimators, each with a different data taper, and then to average the K estimators together. If all K tapers are pairwise orthogonal and properly designed to prevent leakage, the resulting multitaper estimator will be superior to the periodogram in terms of reduced bias and variance, particularly for spectra with high dynamic range and/or rapid variations. Multitapering has been successfully used to study, e.g., the relationship between carbon dioxide and global temperature [11], turbulent plasma fluctuations [20] and heights of ocean waves [22]. Extensive discussion and background information on multitapering can be found in [21] and [18, Chapter 7].

In the present context, if we replace the log periodogram with the log of a multitaper spectrum estimator using K tapers, then the noise distribution is of log χ^2_{2K} form, which is approximately Gaussian for moderate choice of K. Hence this choice of spectrum estimator largely avoids the problem of the non-Gaussianity of the noise, but the noise will now be correlated at the Fourier frequencies. This paper studies the nature of this correlation and shows how it can be handled in both scale-independent and scale-dependent wavelet thresholding to produce attractive practical spectrum estimators (in particular, the restriction on the sample size is eliminated).

In Section II we discuss multitapering and point out that the log of the spectrum estimator has close to a Gaussian distribution for five or more tapers. The covariance of the logarithm of the multitaper spectrum estimator is also derived. In Section III, after establishing notation for the DWT, we consider its application to the log of the multitaper spectrum estimator. The concepts of wavelet shrinkage and scale-independent thresholding are discussed. Scale-dependent thresholding requires an understanding of the scale-dependent variance of the wavelet coefficients; we show that these variances are easy to compute and in fact increase with scale. Also in Section III a two-sided log multitaper spectrum estimator is introduced, and a summary of our proposed algorithm is given. Results of extensive simulation studies are reported in Section IV using four very different processes previously proposed in the literature as representative test cases. These show that wavelet thresholding of log multitaper spectrum estimates gives better results than previously reported

in [6,7,14] using the log periodogram. Section V gives our conclusions and some areas for future research.

II. MULTITAPERING

A. Basics

We assume for convenience that $\{X_t\}$ has zero mean. Given the sample $X_1, \ldots, X_N$, a multitaper spectrum estimator [21] utilizes several different data tapers, the kth of which is denoted $\{h_{t,k} : t = 1, \ldots, N\}$. These tapers are chosen to be orthonormal: $\sum_t h_{t,j} h_{t,k} = 0$ if $j \neq k$ and $= 1$ if $j = k$. The multitaper spectrum estimator is the average of K direct spectral estimators (eigenspectra) and hence takes the form

$$\hat{S}^{(mt)}(f) \equiv \frac{1}{K} \sum_{k=0}^{K-1} \hat{S}_k^{(mt)}(f) \text{ with } \hat{S}_k^{(mt)}(f) \equiv \left| \sum_{t=1}^{N} h_{t,k} X_t e^{-i2\pi ft} \right|^2 . \tag{1}$$

Each taper $\{h_{t,k}\}$ has an associated spectral window $\mathcal{H}_k(f) \equiv \left| \sum_{t=1}^{N} h_{t,k} \exp(-i2\pi ft) \right|^2$, and the overall spectral window $\overline{\mathcal{H}}(f)$ is the average of the individual $\mathcal{H}_k(f)$'s. If the spectrum is not rapidly varying over the effective bandwidth of $\overline{\mathcal{H}}(f)$, the eigenspectra are approximately uncorrelated, which in turn yields the approximation

$$v(f) \equiv \frac{\hat{S}^{(mt)}(f)}{S(f)} \sim \frac{\chi^2_{2K}}{2K}, \qquad 0 < f < 1/2 \tag{2}$$

(see, e.g., [18, p. 360]). A convenient set of easily computable orthonormal tapers is the set of the sine tapers, the kth of which is given by

$$h_{t,k} = \left\{ \frac{2}{N+1} \right\}^{1/2} \sin \left\{ \frac{(k+1)\pi t}{N+1} \right\}, \quad t = 1, \ldots, N. \tag{3}$$

These tapers were introduced by Riedel and Sidorenko [20] and are used henceforth.

Assuming (2), it follows from Bartlett and Kendall [1] that $E\{\log v(f)\} = \psi(K) - \log K$ and $\text{var}\{\log v(f)\} = \psi'(K)$, where $\psi(\cdot)$ and $\psi'(\cdot)$ denote, respectively, the digamma and trigamma functions. Comparison of the distribution of $\log v(f)$ with a normal distribution having the same mean and variance shows very good agreement for $K \geq 5$ [1]. Hence, provided K is at least 5, the random variable

$$\eta(f) \equiv \log v(f) - \psi(K) + \log K$$

is approximately Gaussian distributed with mean 0 and variance $\sigma_\eta^2 = \psi'(K)$. If we let

$$Y(f) \equiv \log \hat{S}^{(mt)}(f) - \psi(K) + \log K, \tag{4}$$

then we have

$$Y(f) = \log S(f) + \eta(f); \tag{5}$$

i.e., the log multitaper estimator (plus a known constant) can be written as a signal (the true log spectrum) plus approximately Gaussian noise with zero mean and known variance σ_η^2. If we now evaluate (5) over a grid of equally spaced frequencies, the resulting model is close to the model usually assumed for wavelet thresholding, with the important exception that the error terms need not be uncorrelated. To see how this correlation affects thresholding, we must first consider the covariance structure of $\eta(f)$ across frequencies.

B. Covariance of Log of Multitaper Spectrum Estimator

For a fixed f and ν such that $0 < f < 1/2$ and $0 < f + \nu < 1/2$, let us define

$$s_\eta(\nu) \equiv \mathrm{cov}\{\eta(f), \eta(f+\nu)\} = \mathrm{cov}\{\log v(f), \log v(f+\nu)\}.$$

If $\log v(f)$ and $\log v(f+\nu)$ were exactly bivariate normal, we would have

$$\mathrm{cov}\{\log v(f), \log v(f+\nu)\} = \log\left(1 + \mathrm{cov}\{v(f), v(f+\nu)\}\right)$$

[8, Section 3.7]; since $\log v(f)$ and $\log v(f+\nu)$ are approximately so, we can take the above to be a reasonable approximation, and hence $s_\eta(\nu) \approx \log(1 + s_v(\nu))$, where $s_v(\nu) \equiv \mathrm{cov}\{v(f), v(f+\nu)\}$. Under the assumption that $S(f) \approx S(f+\nu)$ for small ν, we have

$$s_v(\nu) \approx \frac{1}{K^2 S^2(f)} \sum_{k=0}^{K-1} \sum_{l=0}^{K-1} \mathrm{cov}\left\{\hat{S}_k^{(mt)}(f), \hat{S}_l^{(mt)}(f+\nu)\right\}.$$

Under the same assumption on S, Thomson [21, p. 1069] showed that

$$\mathrm{cov}\left\{\hat{S}_k^{(mt)}(f), \hat{S}_l^{(mt)}(f+\nu)\right\} \approx S^2(f) \left| \sum_{t=1}^{N} h_{t,k} h_{t,l} e^{i2\pi\nu t} \right|^2,$$

an approximation that neglects a frequency dependent term that is only significant for f close to 0 or $1/2$. Hence, we have

$$s_v(\nu) \approx \tilde{s}_v(\nu) \equiv \frac{1}{K^2} \sum_{k=0}^{K-1} \sum_{l=0}^{K-1} \left| \sum_{t=1}^{N} h_{t,k} h_{t,l} e^{i2\pi\nu t} \right|^2.$$

The above development thus gives $s_\eta(\nu) \approx \tilde{s}_\eta(\nu) \equiv \log(1 + \tilde{s}_v(\nu))$.

Figure 1 shows a plot of $\tilde{s}_\eta(\nu)$ versus ν for $K = 10$ sine tapers with $N = 2048$. We note that $\tilde{s}_\eta(\nu)$ is negligible for $\nu \geq 2W$, where $2W \equiv (K+1)/(N+1) \doteq 0.0054$ is the bandwidth of a

multitaper estimator using sine tapers; see [20] or [22]. The straight line shape of the autocovariance is maintained for other values of N and K. Recalling that $s_\eta(0) = \sigma_\eta^2$ and using $2W \approx (K+1)/N$, we can thus formulate a very simple and convenient model, namely,

$$s_\eta(\nu) = \begin{cases} \sigma_\eta^2 \left(1 - \frac{|\nu|N}{K+1}\right), & \text{if } |\nu| \le (K+1)/N; \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

III. WAVELET THRESHOLDING OF MULTITAPER ESTIMATORS

A. Basics of the Discrete Wavelet Transform

Let $\{h_{1,0}^{(D)}, \ldots, h_{1,L_1-1}^{(D)}\}$ denote the coefficients of a Daubechies compactly supported wavelet filter of even length L_1 [2, Ch. 6]. We associate this filter with unit scale, and we assume the normalization $\sum [h_{1,m}^{(D)}]^2 = 1$. For any $M = 2^q \ge L_1$, define $h_{1,m}^{(D)} = 0$ for $m = L_1, \ldots, M-1$, and let

$$H_{1,n}^{(D)} \equiv \sum_{m=0}^{M-1} h_{1,m}^{(D)} e^{-i2\pi mn/M}, \quad n = 0, \ldots, M-1,$$

be the discrete Fourier transform (DFT) of the zero padded wavelet filter. Let $\{g_{1,m}^{(D)}\}$ be the scaling filter, defined as $g_{1,m}^{(D)} \equiv (-1)^{m+1} h_{1,L_1-1-m}^{(D)}$ for $m = 0, \ldots, L_1 - 1$. Let $\{G_{1,n}^{(D)}\}$ denote the DFT of the zero-padded scaling filter. Define the length M periodized wavelet filter $\{h_{j,m}^{(D)}\}$ for scale 2^{j-1} as the inverse DFT of

$$H_{j,n}^{(D)} \equiv H_{1,2^{j-1}n \bmod M}^{(D)} \prod_{l=0}^{j-2} G_{1,2^l n \bmod M}^{(D)}, \quad n = 0, \ldots, M-1.$$

The discrete wavelet transform (DWT) of an $M = 2^q$ dimensional column vector $\mathbf{Y}$ is defined via an $M \times M$ dimensional matrix $\mathcal{W}$ whose rows contain circularly shifted versions of the wavelet filters $\{h_{j,m}^{(D)}\}$, $j = 1, 2, \ldots, q$. Application of the DWT matrix $\mathcal{W}$ to the vector $\mathbf{Y}$ gives $\mathbf{y} \equiv \mathcal{W}\mathbf{Y}$. We denote the elements of $\mathbf{y}$ as

$$\mathbf{y} = \left[y_{1,1},\, y_{1,2},\, \ldots,\, y_{1,\frac{M}{2}},\, y_{2,1},\, y_{2,2},\, \ldots,\, y_{2,\frac{M}{4}},\, \ldots,\, y_{q-1,1},\, y_{q-1,2},\, y_{q,1},\, y_{q+1} \right]^T$$

where the $y_{j,k}$'s are known as the wavelet coefficients, and the remaining term y_{q+1} is the so-called scaling coefficient (in fact this is equal to $\sqrt{M}$ times the sample mean of $\mathbf{Y}$ because all of the elements of the last row of $\mathcal{W}$ are equal to $1/\sqrt{M}$ [13]). Note that $\{y_{j,k} : k = 1, \ldots, M/2^j\}$ is the set of wavelet coefficients associated with scale 2^{j-1}.

In practice, the rows of the matrix $\mathcal{W}$ are not constructed explicitly, but rather the DWT is implemented via a 'pyramid' algorithm that computes the wavelet coefficients one scale at a time

starting with unit scale [12]. This algorithm allows the construction of partial DWTs, in which the wavelet coefficients are computed only for scales indexed by $j = 1, \ldots, q_0$ with $q_0 < q$. For a partial DWT, there are $M/2^{q_0}$ scaling coefficients, which are unaltered but required for computation of the inverse partial DWT.

B. The DWT of the Log Multitaper Spectrum

Let $2M = 2^{q+1}$ be a power of two greater than or equal to the sample size N. By applying a standard fast Fourier transform algorithm to the sequences

$$\{h_{1,k}X_1, \ldots, h_{N,k}X_N, \underbrace{0, \ldots, 0}_{2M-N \text{ zeros}}\}, \quad k = 0, \ldots, K-1,$$

we can readily compute the log multitaper spectrum ordinates $Y(f_n) = \log \hat{S}^{(mt)}(f_n) - \psi(K) + \log K$ at the $M = 2^q$ frequencies $f_n \equiv n/(2M), n = 0, \ldots, M-1$. This yields a sampled version of the model (5), which we can express in vector notation as

$$\mathbf{Y} \equiv \begin{bmatrix} Y(f_0) \\ \vdots \\ Y(f_{M-1}) \end{bmatrix} = \mathbf{S} + \mathbf{N} \equiv \begin{bmatrix} \log S(f_0) \\ \vdots \\ \log S(f_{M-1}) \end{bmatrix} + \begin{bmatrix} \eta(f_0) \\ \vdots \\ \eta(f_{M-1}) \end{bmatrix}. \tag{7}$$

We assume that $\mathbf{N}$ is multivariate Gaussian with zero mean and covariance matrix

$$\Sigma_{\mathbf{N}} \equiv \begin{bmatrix} s_\eta(f_0) & s_\eta(f_1) & \cdots & s_\eta(f_{\frac{M}{2}-1}) & s_\eta(f_{\frac{M}{2}}) & s_\eta(f_{\frac{M}{2}-1}) & \cdots & s_\eta(f_1) \\ s_\eta(f_1) & s_\eta(f_0) & \cdots & s_\eta(f_{\frac{M}{2}-2}) & s_\eta(f_{\frac{M}{2}-1}) & s_\eta(f_{\frac{M}{2}}) & \cdots & s_\eta(f_2) \\ s_\eta(f_2) & s_\eta(f_1) & \cdots & s_\eta(f_{\frac{M}{2}-3}) & s_\eta(f_{\frac{M}{2}-2}) & s_\eta(f_{\frac{M}{2}-1}) & \cdots & s_\eta(f_3) \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ s_\eta(f_1) & s_\eta(f_2) & \cdots & s_\eta(f_{\frac{M}{2}}) & s_\eta(f_{\frac{M}{2}-1}) & s_\eta(f_{\frac{M}{2}-2}) & \cdots & s_\eta(f_0) \end{bmatrix} \tag{8}$$

with $s_\eta(f_n)$ defined by (6). This circular matrix is symmetric and positive semidefinite. Its elements are in accordance with the discussion in Section II for $\eta(f_n)$'s with f_n sufficiently far from 0 and 1/2; for other frequencies, it provides a convenient mathematical structure that is put to good use in (9) below. The matrix $\Sigma_{\mathbf{N}}$ can be regarded as an approximation to what would be a more natural assumption for the covariance structure of $\mathbf{N}$, namely, a Toeplitz matrix $\Sigma'_{\mathbf{N}}$ whose first row is given by $s_\eta(f_n)$, $n = 0, \ldots, M-1$. In their study of wavelet shrinkage for correlated data, Johnstone and Silverman [9, pp. 322, 342] also assume a circular covariance matrix and cite arguments as to why it is a reasonable approximation to $\Sigma'_{\mathbf{N}}$. Because of (6), the matrix $\Sigma_{\mathbf{N}}$ is almost a banded Toeplitz matrix: it has $2K+1$ nonzero diagonals, but it also has $K(K+1)/2$ nonzero 'off-band' elements in both the upper right-hand and lower left-hand corners – these off-band elements are what distinguish $\Sigma_{\mathbf{N}}$ and $\Sigma'_{\mathbf{N}}$.

Note that the DWT of $\mathbf{Y}$ can be written as $\mathbf{y} \equiv \mathcal{W}\mathbf{Y} = \mathcal{W}\mathbf{S} + \mathcal{W}\mathbf{N} \equiv \mathbf{s} + \mathbf{n}$. On an element by element basis, we have $y_{j,k} = s_{j,k} + n_{j,k}$ and $y_{q+1} = s_{q+1} + n_{q+1}$.

C. Wavelet Shrinkage

Because the noise term $\mathbf{N}$ in model (7) is assumed to be Gaussian, it follows that the wavelet coefficients $\{y_{j,k}\}$ are Gaussian distributed, so we can make use of the hard and soft thresholding schemes discussed in [3] and [4]. A soft threshold function can be defined, with threshold T:

$$\delta_S(\omega, T) = \mathrm{sgn}(\omega)(|\omega| - T)_+, \text{ where } (|\omega| - T)_+ = \begin{cases} |\omega| - T, & \text{if } |\omega| > T; \\ 0, & \text{otherwise.} \end{cases}$$

The soft threshold shrinks or kills all the wavelet coefficients to which it is applied. As a result, soft thresholding produces smoother results than hard thresholding, with a rule defined by

$$\delta_H(\omega, T) = \omega\, 1\{|\omega| > T\}.$$

A compromise between soft and hard thresholding is a 'mid' threshold, defined by

$$\delta_M(\omega, T) = \mathrm{sgn}(\omega)(|\omega| - T)_M, \text{ where } (|\omega| - T)_M = \begin{cases} 2(|\omega| - T)_+, & \text{if } |\omega| < 2T; \\ |\omega|, & \text{otherwise.} \end{cases}$$

With such a threshold, large coefficients, i.e., those exceeding $2T$ in magnitude, are left untouched, those between T and $2T$ in magnitude are shrunk, while those less than T in magnitude are killed.

The wavelet coefficients $y_{j,k}$ could be thresholded using $\delta(\cdot,\cdot)$ set to $\delta_S(\cdot,\cdot)$, $\delta_H(\cdot,\cdot)$ or $\delta_M(\cdot,\cdot)$:

$$\hat{y}_{j,k} = \begin{cases} \delta(y_{j,k}, T), & \text{if } 1 \le j \le q_0; \\ y_{j,k}, & \text{if } j > q_0, \end{cases}$$

where $q_0 < q$ is some specified coarse resolution level.

D. Scale-independent Universal Thresholding

For either soft, mid or hard thresholding, a critical issue is to determine an appropriate threshold level T. If the signal component is in fact zero, then with high probability the combination of (zero) signal plus noise should not exceed the threshold level. A simple, but useful, approach is discussed in [3] and [4]. If the elements of $\mathbf{N}$ were independent and identically distributed $\mathcal{N}(0, \sigma_\eta^2)$ random variables, then the elements of their wavelet transform $\mathbf{n}$ would also be such. In this case the threshold is chosen as $T = T_M = \sigma_\eta\sqrt{(2\log M)}$. The justification for this so-called universal threshold T_M is that, if $Z_1, \ldots, Z_M$ is a sequence of independent and identically distributed $\mathcal{N}(0,1)$ random variables, then as $M \to \infty$

$$P_M \equiv Pr\left[\max_i |Z_i| > \sqrt{(2\log M)}\right] \to 0.$$

This means that, asymptotically, if the signal component is in fact zero, then the probability of a 'false alarm' will tend to zero, so that the combination of (zero) signal plus noise will not exceed the threshold level, and will hence be set to zero. Hastie and Tibshirani (in the discussion of [4], pp. 347–9) point out that such a threshold will frequently allow noise 'wiggles' into the reconstruction of smooth functions. Donoho *et al.* [4, p. 366] comment that such wiggles can be made rarer by changing T_M to $T_M = \sigma\sqrt{(3 \log M)}$.

The universal threshold is attractively simple, but is strictly suitable only for iid Gaussian $\mathbf{n}$. In model (7) we have assumed the noise $\mathbf{N}$ to be Gaussian and hence the corresponding wavelet coefficients $\mathbf{n}$ also have a Gaussian distribution; however, because the elements of $\mathbf{N}$ have covariance (8), the elements of $\mathbf{n}$ are also correlated. We need to see how the correlation of the noise $\mathbf{N}$ feeds through to the variance of the wavelet coefficients at each scale.

E. Scale-dependent Thresholding

Johnstone and Silverman [9] look at scale-dependent thresholding, which in our context would explicitly allow for a different variance for the $n_{j,k}$ at each scale, induced by the correlation in the noise. The current work provides a near-ideal application for the Johnstone–Silverman technique. As indicated below, we can easily compute the required scale-dependent variances from the parameters defining the multitaper estimator and the chosen wavelet filters. In most other statistical applications these variances must be estimated from the coefficients at each scale, which becomes problematic as the number of coefficients $\{n_{j,k}\}$ decreases with increasing level j. Moreover, we can derive insightful properties about the scale-dependent variances that enable us to better understand how scale-independent thresholding contrasts with the scale-dependent method.

Under the mild assumption that $L_1 \leq M$, scale-dependent variances can be easily computed as follows. Because by construction the covariance matrix $\Sigma_{\mathbf{N}}$ in (8) is circular, we have, e.g., Fuller [5, p. 151], that $\Sigma_{\mathbf{N}} = G^H D G$, where the kth row of G contains $\{e^{-i2\pi kn/M}/\sqrt{M} : n = 0, \ldots, M-1\}$; '$H$' denotes Hermitian transpose; and D is a diagonal matrix with diagonal elements $\{S_n : n = 0, \ldots, M-1\}$, which is the DFT of the first row of $\Sigma_{\mathbf{N}}$. Let $\mathbf{a}_{j,k}^T$ represent the row of $\mathcal{W}$ whose multiplication with $\mathbf{N}$ yields $n_{j,k}$, and let $\{A_n\}$ be its DFT. Because $|A_n| = |H_{j,n}^{(D)}|$, we have

$$\operatorname{var}\{n_{j,k}\} = \mathbf{a}^T \Sigma_{\mathbf{N}} \mathbf{a} = \mathbf{a}^H G^H D G \mathbf{a} = \frac{1}{M} \sum_{n=0}^{M-1} S_n \left| H_{j,n}^{(D)} \right|^2 \equiv \sigma_j^2, \tag{9}$$

which are the required scale-dependent variances (note that σ_j^2 does not depend on k).

Scale-dependent thresholding uses a different threshold at each scale, with the threshold at scale 2^{j-1} being given by $T_{M,j} = \sigma_j\sqrt{(2\log M)}$. To understand the effect of this type of thresholding, we worked out an explicit progression for the variances σ_j^2 under the special case of the Haar DWT and found that, with $M = N/2 = 2^q$ as before and with the mild condition $K < M/2$,

$$\sigma_1^2 < \sigma_2^2 < \cdots < \sigma_q^2 < \bar{\sigma}_{q+1}^2, \tag{10}$$

where $\bar{\sigma}_{q+1}^2$ is the variance of the single scaling coefficient n_{q+1} (see the appendix for a proof; for selected q and K, we have used (9) to demonstrate numerically that the above still holds for DWTs based on other Daubechies wavelets). For comparison with scale-dependent thresholding $T_M = \sigma_\eta\sqrt{(2\log M)}$, note that $\mathrm{tr}\{\Sigma_{\mathbf{N}}\} = \mathrm{tr}\{\mathcal{W}\Sigma_{\mathbf{N}}\mathcal{W}^T\}$ says that

$$\sigma_\eta^2 = \frac{\bar{\sigma}_{q+1}^2}{2^q} + \sum_{j=1}^{q}\frac{\sigma_j^2}{2^j},$$

so (10) along with an easy 'proof by contradiction' shows that $\sigma_1^2 < \sigma_\eta^2 < \bar{\sigma}_{q+1}^2$. Computations indicate that in fact $\sigma_2^2 < \sigma_\eta^2 < \sigma_3^2$ when $K = 5,\ldots,9$; $\sigma_\eta^2 \doteq \sigma_3^2$ when $K = 10$; and $\sigma_3^2 < \sigma_\eta^2 < \sigma_4^2$ when $K = 11,\ldots,41$ (this covers the range of K likely to be used in practical applications). This positioning of the scale-independent variance σ_η^2 with respect to the scale-dependent variances σ_j^2 becomes important in interpreting the results of our simulation study (see Section V).

To confirm the increase in σ_j^2 with j and, more importantly, to validate the statistical model defined by (6), (7) and (8), we conducted a simulation study in which the following steps were repeated 1000 times:

(i) a sample $X_1,\ldots,X_N$ from a specified Gaussian stationary process was generated, with $2M = N = 2048$;

(ii) the multitaper estimator was calculated using $K = 10$ sinusoidal tapers;

(iii) the DWT of $\mathbf{N} = \mathbf{Y} - \mathbf{S}$ was computed based upon wavelet filters of lengths $L_1 = 2, 4, 8$ and 16, yielding the wavelet coefficients $\{n_{j,k}\}$; and

(iv) the standard deviation of the wavelet coefficients $\{n_{j,k}\}$ at scales 1, 2, 4 and 8 was estimated from the square root of $\hat{\sigma}_j^2 \equiv 2^{j-q}\sum_{k=1}^{2^{q-j}} n_{j,k}^2$.

Two different processes were used, a white noise process and the autoregressive process of order 2, AR(2), used by Moulin [14] (see Section IV for details), but the results proved to be virtually identical, so we only report on the AR(2) case here. Box plots for the 1000 standard deviations for

each wavelet filter and each scale are plotted in Fig. 2 (the box plots shown there are defined as follows: a thin horizontal line is drawn through the box at the median of the data, the upper and lower ends of the box are at the upper and lower quartiles, and the vertical lines extend from the box to points within a standard range of the data, defined as 1.5 times the inter-quartile range). The variability of the estimates $\hat{\sigma}_j$ increases with scale, mainly because the number of wavelet coefficients from which the standard deviation is calculated reduces from $M/2 = 512$ for unit scale to $M/16 = 64$ for scale 8. The thick horizontal lines extending beyond each box plot indicate the value of σ_j derived from (9). The 'nominal' standard deviation $\sigma_\eta = \sqrt{[\psi'(10)]} \doteq 0.32$ (to be used in the scale-independent thresholding) is marked as a thin horizontal line spanning the width of each plot. This exceeds the sample median of the $\hat{\sigma}_j$'s for scales 1, 2 and 4, when using Daubechies' extremal phase wavelet with 4 coefficients, (D(4)), or her least asymmetric wavelets with 8 or 16 coefficients, (LA(8) and LA(16)). For the Haar wavelet, the nominal standard deviation exceeds the sample median of the $\hat{\sigma}_j$'s for scales 1 and 2. Fig. 2 thus nicely illustrates (10) and supports our contention that the approximations leading up to (6), (7) and (8) are reasonable.

F. Two-Sided Multitaper Spectrum Estimate

The theoretical results, which were compared to the empirical results in Fig. 2, were derived using the covariance matrix in (8), which was formulated using nonnegative frequencies less than 1/2. However, because $\hat{S}^{(mt)}(f)$ is an even periodic function with unit period, we could just as easily have used nonpositive frequencies greater than $-1/2$ or, equivalently, frequencies f_n satisfying $1/2 < f_n \leq 1$ so that (7) would involve $f_{M+1}, \ldots, f_{2M}$ rather than $f_0, \ldots, f_{M-1}$. There is an advantage to using all $2M$ frequencies $f_0, \ldots, f_{2M-1}$ because then we compute the DWT of a complete period of the log multitaper spectrum estimate, which better matches the circular nature of the DWT and hence avoids severe discontinuities at the boundaries. We note that Moulin [14] makes a similar argument and that standard lag window spectral estimates are similarly formed by circularly smoothing over a complete period of the periodogram. Accordingly, in practise we take the DWT of an expanded version of (7), namely,

$$\mathbf{Y} \equiv \begin{bmatrix} Y(f_0) \\ \vdots \\ Y(f_{2M-1}) \end{bmatrix} = \mathbf{S} + \mathbf{N} \equiv \begin{bmatrix} \log S(f_0) \\ \vdots \\ \log S(f_{2M-1}) \end{bmatrix} + \begin{bmatrix} \eta(f_0) \\ \vdots \\ \eta(f_{2M-1}) \end{bmatrix}. \tag{11}$$

Note that we still use T_M for thresholding: this level is appropriate for the nonnegative and nonpositive cases separately, and, because of the localized nature of the DWT for small j, the

DWT of (11) is essentially a combination of these two cases except for frequencies close to 0 or $\pm 1/2$, for which our theoretical development does not apply anyway. We also note that, in contrast to lag window estimates, the thresholded multitaper estimates need *not* be symmetric about $f = 1/2$, so we have chosen to average $\hat{S}^{(wtmt)}(f_n)$ and $\hat{S}^{(wtmt)}(f_{2M-n})$, where $\hat{S}^{(wtmt)}(f)$ is the wavelet-thresholded spectrum estimate, in all results discussed below.

G. Summary of Proposed Spectrum Estimation Algorithm

Given (a) $X_1, \ldots, X_N$, a time series that is a realization of a portion of a zero mean stationary process; (b) K, the number of tapers ; (c) M, a power of two satisfying $2M \equiv 2^{q+1} \geq N$; (d) q_0, a resolution level satisfying $q_0 \leq q$; (e) $\delta(\cdot, \cdot)$, one of the thresholding functions discussed in Section III.C; and (f) $\{h_{1,m}^{(D)}\}$, a Daubechies wavelet filter of length L_1; we construct our spectrum estimator $\hat{S}^{(wtmt)}$ in the following manner:

(i) *Calculate the logarithm of the multitaper estimate.* For $k = 1, \ldots, K$, we form the tapered time series $h_{1,k}X_1, \ldots, h_{N,k}X_N$ using the $h_{t,k}$'s defined in (3). We then append $2M - N$ zeros to the tapered series and use a standard 'power of two' fast Fourier transform algorithm to form the kth eigenspectrum

$$\hat{S}_k^{(mt)}(\tfrac{n}{2M}) = \left| \sum_{t=1}^{2M} h_{t,k} X_t e^{-i2\pi nt/2M} \right|^2 , \quad n = 0, \ldots, 2M-1$$

(cf. (1); here $h_{t,k}X_t \equiv 0$ for $t > N$). We then average the K eigenspectra to form the multitaper estimate $\hat{S}^{(mt)}(\frac{n}{2M})$ (cf. (1)). Next we form $Y(\frac{n}{2M}) \equiv \log_e \hat{S}^{(mt)}(\frac{n}{2M}) - \psi(K) + \log K$ and place these $2M$ values in the column vector $\mathbf{Y}$ (cf. (4) and (11)).

(ii) *Apply a partial DWT out to level q_0.* Let $\mathcal{W}_{q_0}$ represent the q_0th order partial DWT. By q_0 recursive applications of the pyramid algorithm described by Mallat [12], we obtain the DWT transform coefficients $\mathbf{y} \equiv \mathcal{W}_{q_0}\mathbf{Y}$. The elements of $\mathbf{y}$ to be thresholded are the wavelet coefficients for levels $j = 1, \ldots, q_0$, denoted as $y_{j,k}, k = 1, \ldots, M/2^{j-1}$. The remaining $M/2^{q_0-1}$ coefficients in $\mathbf{y}$, i.e., the level q_0 scaling coefficients, are left alone.

(iii) *Apply thresholding to the wavelet coefficients.*

(a) For scale-independent thresholding, we set $\hat{y}_{j,k} = \delta(y_{j,k}, \sigma_\eta \sqrt{[2 \log_e M]})$, where $\sigma_\eta \equiv \sqrt{\psi'(K)}$.

(b) For scale-dependent thresholding, we set $\hat{y}_{j,k} = \delta(y_{j,k}, \sigma_j \sqrt{[2 \log_e M]})$, where σ_j^2 is defined in (9) and involves the sequences $\{H_{j,n}^{(D)}\}$ and $\{S_j\}$. Computation of $\{H_{j,n}^{(D)}\}$ is discussed

in the first paragraph of Section III, while $\{S_j\}$ is the DFT of the first row of $\Sigma_{\mathbf{N}}$ in (8), which in turn involves $s_\eta(f_n)$ of (6) with $f_n \equiv n/(2M)$ and $\sigma_\eta^2 \equiv \psi'(K)$.

(iv) *Invert the partial DWT and produce the smoothed spectrum estimate.* We form $\hat{\mathbf{y}}$ by starting with $\mathbf{y}$ and then replacing each wavelet coefficient $y_{j,k}$ with the corresponding thresholded $\hat{y}_{j,k}$. We then compute $\widehat{\mathbf{Y}} \equiv \mathcal{W}_{q_0}^T \hat{\mathbf{y}}$ by q_0 recursive applications of the inverse pyramid algorithm [12]. With $\widehat{Y}_n$ denoting the nth element of $\widehat{\mathbf{Y}}$, we can now form our proposed spectrum estimate:

$$\hat{S}^{(wtmt)}(f_n) \equiv \begin{cases} e^{\widehat{Y}_0}, & n = 0 \text{ (i.e., } f_n = 0); \\ (e^{\widehat{Y}_n} + e^{\widehat{Y}_{2M-n}})/2, & n = 1, \ldots, M-1 \text{ (i.e., } f_n = \frac{1}{2M}, \ldots, \frac{M-1}{2M}); \text{ and} \\ e^{\widehat{Y}_M}, & n = M \text{ (i.e., } f_n = \frac{1}{2}). \end{cases}$$

If $\{X_t\}$ cannot be assumed to have zero mean (usually the case in practical applications), we replace X_t in the above by $X_t' - \bar{X}$, where $\bar{X}$ is the sample mean.

IV. SIMULATION STUDY

We tested our scale-independent and scale-dependent schemes on four different Gaussian stationary processes. In models A, B and D, the variance of the zero mean Gaussian white noise process $\{\epsilon_t\}$ is set to unity.

A. The AR(24) process $X_t = \sum_{n=1}^{24} \phi_{n,24} X_{t-n} + \epsilon_t$ used by Gao [6,7], where the $\phi_{n,24}$'s are specified by equating coefficients of z^n in

$$\prod_{l=1}^{12} \left(1 - \frac{2\cos(\alpha_l)}{A_l} z + \frac{z^2}{A_l^2}\right) = 1 - \sum_{n=1}^{24} \phi_{n,24} z^n$$

with

$$\alpha_l \equiv \begin{cases} 0.2l, & l = 1, \ldots, 5; \\ 0.2 + 0.2l, & l = 6, \ldots, 12; \end{cases} \text{ and } A_l \equiv \begin{cases} 1.005, & l = 3; \\ 1.03, & \text{otherwise} \end{cases}$$

(thus $\phi_{1,24} \doteq 2.521627, \phi_{2,24} \doteq -4.771540, \ldots, \phi_{24,24} \doteq -0.516712$). Each realization of this process (and also of the AR(2) process below) was produced by generating stationary start-up values [10].

B. The AR(2) process used by Moulin [14] specified by $X_t = \phi_{1,2} X_{t-1} + \phi_{2,2} X_{t-2} + \epsilon_t$, where $\phi_{1,2} \equiv 0.97\sqrt{2}$ and $\phi_{2,2} \equiv -(0.97)^2$.

C. The 'typical mobile radio communications' spectrum used by Moulin [14] specified by a superposition of two bandlimited, fading, mobile radio signals, a white background noise, and a narrow-band interference term with Gaussian spectrum. The overall power spectrum is

$$\begin{aligned} S(f) =& 10^{-3} + 0.2 \exp^{-(|f|-0.45)^2/(2 \cdot 10^{-6})} + \sqrt{[1 - (f/B_0)^2]} 1\{0 \le |f| \le B_0\} \\ &+ \sqrt{[1 - (\{|f| - f_0\}/B_0)^2]} 1\{f_0 - B_0 \le |f| \le f_0 + B_0\} \end{aligned}$$

with $f_0 = 0.3$ and $B_0 = 0.1$. To generate the sequences, let N' be any positive even integer. From a Gaussian white noise sequence $\{Z_1, \ldots, Z_{N'}\}$ construct $C_0 = Z_1$, $C_{N'/2} = Z_{N'}$, and $C_n = (Z_{2n} + iZ_{2n+1})/\sqrt{2}$, $n = 1, \ldots, (N'/2) - 1$. Then for $f_n = n/N'$ define $F_n = C_n\sqrt{[S(f_n)]}$, $n = 0, \ldots, N'/2$, and $F_n = F^*_{N'-n}$, $n > N'/2$. Finally, let

$$X_t = \frac{1}{\sqrt{N'}} \sum_{n=0}^{N'-1} F_n e^{i2\pi f_n t},$$

corresponding effectively to a discretization of the spectral representation formula. By sampling $N \ll N'$ values, with N' large, this approximate frequency domain method gives excellent synthesis [17]. In our simulations N' was set to $4N$.

D. The high order moving average (MA) process $X_t = \sum_{n=0}^{15000} \theta_n \, \epsilon_{t-n}$ used by Gao [6,7] for which $\theta_0 = 1$; $\theta_1 = \pi/4$; and $\theta_{n+1} = \sin(\pi n/2)/n$, $n = 1, 2, \ldots, 14999$. Note that, starting with θ_3, the odd numbered coefficients are exactly zero. Each realization of $X_1, \ldots, X_N$ requires generation of $N + 15000$ deviates from $\{\epsilon_t\}$. As will be seen the spectrum corresponding to this series has a sharply-defined peak and trough covering some 25 dB, and is difficult to estimate.

Except for the simulations reported in the second paragraphs of subsections B and C below, we used a $K = 10$ multitaper spectrum estimate, a series length of $N = 2048$, and the Daubechies LA(8) wavelet filter. The bandwidth of the multitaper estimator is $(K+1)/(N+1) \doteq 0.0054$, which is small relative to the widths of the peaks in models A, B and C, but is too large to faithfully capture the tip of the very sharp peak in model D. Other choices of wavelet filter (D(4) and LA(10)) produced very similar results.

For each model, scale-independent and scale-dependent soft, mid and hard thresholds were investigated, and different numbers of wavelet coefficients were left untouched by the thresholding, namely 64, 32 and 8, corresponding to setting q_0 to 5, 6 and 8, respectively. For each of 1000 repetitions over each choice of thresholding and value of q_0, the root mean square error (rmse)

$$\left[\frac{1}{N/2+1} \sum_{n=0}^{N/2} \{10 \log_{10} \hat{S}^{(wtmt)}(f_n) - 10 \log_{10} S(f_n)\}^2 \right]^{1/2} \tag{12}$$

was calculated, and the average and standard error of these rms errors over the 1000 simulations was recorded. Results are summarized in Fig. 3. Furthermore, the spectrum estimate with rmse closest to the average of the 1000 is plotted in Figs. 4–7 as a 'representative' estimate for the stated choice of model, threshold type and q_0.

A. The AR(24) Model

The spectrum of the AR(24) process has a dynamic range close to 90 dB. This process was used by Gao [7] in his study of wavelet thresholding of the log periodogram, but in fact the periodogram for this process is badly biased due to leakage at least for sample sizes $N \leq 2048$ and $f \geq 0.4$ (see Fig. 1 of [7]). In contrast, the multitaper scheme suppresses leakage and, when combined with wavelet shrinkage, produces a smoothed log spectrum estimate with no leakage, as shown in Fig. 4. This figure shows the representative estimates for (a) scale-independent, soft thresholding and (b) scale-dependent, hard thresholding, both with $q_0 = 5$ (these choices give the minimum average rmse across parameter combinations for both scale-independent and scale-dependent thresholding; see Fig. 3). Both estimates are good, but the scale-independent estimate (a) would probably be judged superior in the estimation of the three highest frequency peaks in the spectrum. The scale-dependent estimate (b) shows evidence of fine-scale noise coefficients unattenuated by the thresholding.

B. The AR(2) Model

Figure 5 shows the representative estimates for (a) scale-independent, soft thresholding and (b) scale-dependent, hard thresholding, both with $q_0 = 8$ (these choices again giving the minimum average rmse for both types of thresholding; see Fig. 3). Both estimates are quite good. The scale-dependent estimate (b) again shows evidence of fine-scale noise coefficients unattenuated by the thresholding, but also does slightly better in estimating the height of the peak in the spectrum.

By using his scale and wavelet-dependent thresholding scheme applied to the periodogram with $N = 512$, Moulin [14] obtained a mean square error 43% (using his RC3 wavelet) and 41% (using his RD6 wavelet) of that found by scale-independent universal thresholding of the periodogram. The multitaper approach ($N = 512$, $K = 10$, LA(8) wavelet filter coefficients, scale-independent soft thresholding or scale-dependent hard thresholding) gave a mean square error 32% of that found by scale-independent universal thresholding of the periodogram; hence the improvement using our scheme is larger than obtained by Moulin [14].

C. Mobile Radio Communications Model

Figure 6 shows the representative estimates for (a) scale-independent, soft thresholding and (b) scale-dependent, hard thresholding, both with $q_0 = 6$ (again corresponding to the minimum average rmse for both types of thresholding; see Fig. 3). Both the 30 dB rises and falls of the bandlimited part of the spectrum, and the narrow-band interference at $f = 0.45$ are well estimated

by both methods, with the scale-independent estimate (a) being somewhat better at capturing the height of the peak. Again, the scale-dependent estimate (b) shows evidence of fine-scale noise coefficients unattenuated by the thresholding.

By using his scale and wavelet-dependent thresholding scheme applied to the periodogram with $N = 256$, Moulin [14] obtained a mean square error 89% (using his RC3 wavelet) and 88% (using his RD6 wavelet) of that found by scale-independent universal thresholding of the periodogram. The multitaper approach ($N = 256$ with $K = 5$ because of the relatively shorter data length, LA(8) wavelet filter coefficients) gave a mean square error 49% (using scale-independent soft thresholding) and 56% (using scale-dependent hard thresholding), of that found by scale-independent universal thresholding of the periodogram. The improvement using our simpler scheme is again larger than obtained by Moulin [14].

D. High Order MA Model

The true spectrum of this model, shown by the light curves in Fig. 7, comprises a slow decay, a very sharp peak followed by a deep trough, and then a gradual increase. The minimum average rmse across parameter combinations occurs with $q_0 = 6$ for both (a) scale-independent, soft thresholding and (b) scale-dependent, hard thresholding (see Fig. 3). The representative spectrum estimates are shown in Fig. 7. Here the scale-dependent estimate (b) gives a better overall visual impression than the scale-independent estimate (a), even though the observed rmse's for the two estimates are virtually identical (see Fig. 3). The very sharp peak is too sharp to be properly resolved from a series of this length by either estimate. Again, both methods do much better than Gao's method [6,7].

V. CONCLUSIONS AND DISCUSSION

We have developed a technique for spectrum estimation based on wavelet thresholding of the log of multitaper spectrum estimates. In contrast to the log periodogram, the log multitaper estimator can have much better small sample bias properties and is close to Gaussian distributed provided a moderate number of tapers is used ($K \geq 5$). As quantified approximately by Equation (6), the log multitaper estimator is correlated across frequencies. When formulated as the 'signal plus noise' model (5), the resulting spectrum estimation problem falls into the variation on wavelet shrinkage studied recently by Johnstone and Silverman [9], who demonstrate that correlated noise can be handled by basing thresholds on scale-dependent variances. The structure of our model is such that these scale-dependent variances can be readily calculated using (9). Simulations of

four representative stationary processes show that hard thresholding is the best to use with scale-dependent thresholds and that our proposed estimator outperforms the log periodogram-based estimators proposed previously in the literature [6,7,14].

We also considered scale-independent thresholding. Because the scale-independent noise variance σ_η^2 falls above just the lowest 2 or 3 scale-dependent variances σ_j^2 (see the thin horizontal lines spanning the plots in Fig. 2), the correlation that multitapering induces acts in accordance with the underlying paradigm for wavelet shrinkage in that it actually helps suppress small scale 'noise spikes' while leaving informative large scale wavelet coefficients relatively unattenuated (this result suggests that the usual scale-independent wavelet shrinkage can be expected to work well in other 'signal plus correlated noise' scenarios in which the correlation of the noise decays in a manner similar to (6)). In our simulation study, scale-independent thresholding actually outperformed the scale-dependent thresholding in a number of cases (see Fig. 3, in which scale-independent results are indicated by dashed lines). The simulations also suggest that soft thresholding is the best scheme to use in combination with scale-independent universal thresholds.

The rmse of the spectrum estimators does vary somewhat with the choice of q_0, as shown in Fig. 3. However, provided that soft thresholding is used with scale-independent universal thresholds, and hard thresholding is used with scale-dependent universal thresholds, the variation of rmse with q_0 does not cause the corresponding spectrum estimates to be markedly different in their visual appearances. Thus choice of q_0 should not be too critical in practice, but this issue certainly deserves more research (e.g., cross-validation might be useful in picking q_0).

There are several lines of research that are left open and should be explored in the future. First, by itself the multitaper method is capable of producing excellent spectrum estimates in the sense of exhibiting all the important spectral features, but smoothing is required to achieve statistical consistency if the number of tapers K is held fixed as the sample size N increases. The purpose of wavelet shrinkage is to provide this smoothing, so it would be helpful to know that, as $N \to \infty$, we are guaranteed that our spectrum estimator converges in some sense to the true spectrum. The approach of Johnstone and Silverman [9] provides a pathway to proving consistency, but unfortunately we cannot evoke their main theorem directly because our covariance matrix (8) is that of a (circularized) noninvertible moving average process and hence fails to satisfy the bounded dependence condition that Johnstone and Silverman [9, p. 334] assume. Because of the good results we obtained in our simulations, we conjecture that it is possible to prove the Johnstone–Silverman theorem under a condition that allows for noninvertible covariance matrices. Second,

because thresholding is so critical to our scheme, it is important to look beyond simple universal thresholding and consider more recent ideas for setting threshold levels, including approaches based upon Stein's unbiased risk estimator (SURE) [3,9] and cross-validation [15]. Finally, Johnstone and Silverman [9] note that 'shift-invariant' wavelet denoising can be used to eliminate certain alignment effects with the ordinary DWT, so it would interesting to explore how this impacts our proposed spectrum estimator.

APPENDIX

Here we prove (10) for the special case of the Haar wavelet under the assumption that $K < M/2$. First we use (9) to develop an explicit expression for $\tilde{\sigma}_j^2 \equiv \sigma_j^2/\sigma_\eta^2$, which requires expressions for $\{S_n\}$ (the DFT of the first row of (8)) and for $\{|H_{j,n}^{(D)}|^2\}$. Using (6) and the fact $N = 2M$, we have

$$s_\eta(f_l) = \begin{cases} \sigma_\eta^2 \left(1 - \frac{|l|}{K+1}\right), & |l| \le K; \\ 0, & \text{otherwise;} \end{cases} \quad \text{so } S_n = \sigma_\eta^2 \sum_{l=-K}^{K} \left(1 - \frac{|l|}{K+1}\right) e^{-i2\pi ln/M}. \tag{13}$$

Hence we can write

$$\tilde{\sigma}_j^2 = \sum_{l=-K}^{K} \left(1 - \frac{|l|}{K+1}\right) A_l, \text{ where } A_l \equiv \frac{1}{M} \sum_{n=0}^{M-1} \left|H_{j,n}^{(D)}\right|^2 e^{i2\pi ln/M}.$$

For the Haar wavelet, $\{\left|H_{j,n}^{(D)}\right|^2\}$ is the DFT of the circular autocorrelation of

$$\underbrace{\frac{1}{2^{j/2}}, \ldots, \frac{1}{2^{j/2}}}_{2^{j-1} \text{ of these}}, \underbrace{-\frac{1}{2^{j/2}}, \ldots, -\frac{1}{2^{j/2}}}_{2^{j-1} \text{ of these}}, \underbrace{0, \ldots, 0}_{M-2^j \text{ of these}};$$

for $j < q$ and letting $\lambda_j \equiv 2^j$, the inverse DFT is given by

$$A_l = \begin{cases} 1 - \frac{3|l|}{\lambda_j}, & |l| = 0, \ldots, \lambda_j/2; \\ -\left(1 - \frac{|l|}{\lambda_j}\right), & |l| = \lambda_j/2, \ldots, \lambda_j - 1; \\ 0, & \text{otherwise;} \end{cases}$$

and, for $j = q$,

$$A_l = 1 - \frac{4|l|}{M}, \qquad |l| = 0, \ldots, \frac{M}{2}.$$

With $B \equiv K + 1$, we can obtain explicit expressions for $\tilde{\sigma}_j^2$, $j < q$, and for $\tilde{\sigma}_q^2$:

[i] $\tilde{\sigma}_j^2 = \frac{1 - B^2 + B\lambda_j}{\lambda_j}$ when $B \le \lambda_j/2$;
[ii] $\tilde{\sigma}_j^2 = \frac{6\lambda_j^2 B + 4\lambda_j - 6\lambda_j B^2 + 2B^3 - 2B - \lambda_j^3}{6B\lambda_j}$ when $\lambda_j/2 < B < \lambda_j$;
[iii] $\tilde{\sigma}_j^2 = \frac{2 + \lambda_j^2}{6B}$ when $B \ge \lambda_j$: and

[iv] $\tilde{\sigma}_q^2 = \frac{3BM-4B^2+4}{3M}$.

Our first claim is that $\tilde{\sigma}_j^2 < \tilde{\sigma}_{j+1}^2$ for $j = 1, \ldots, q-2$. Suppose first that $B = 2^p$ for integer p such that $1 \le p \le q-2$. The condition $B \le \lambda_j/2$ of case [i] then holds when $j \ge p+1$, while the condition $B > \lambda_j - 1$ of case [iii] holds when $j \le p$. Hence $\tilde{\sigma}_j^2 < \tilde{\sigma}_{j+1}^2$ for all $p+1 \le j \le q-2$ because case [i] implies

$$\frac{1-B^2+B\lambda_j}{\lambda_j} < \frac{1-B^2+B\lambda_{j+1}}{\lambda_{j+1}} \Longleftrightarrow 2-2B^2+2B\lambda_j < 1-B^2+2B\lambda_j \Longleftrightarrow 1 < B^2,$$

which holds since $B \ge 2$. Also $\tilde{\sigma}_j^2 < \tilde{\sigma}_{j+1}^2$ for all $j \le p-1$ because case [iii] implies

$$\frac{2+\lambda_j^2}{6B} < \frac{2+\lambda_{j+1}^2}{6B} = \frac{2+4\lambda_j^2}{6B} \Longleftrightarrow 0 < 3\lambda_j^2,$$

which holds since $\lambda_j \ge 1$. The claim thus holds for $B = 2^p$ if we can show that it holds for the case $j = p$; i.e., $\tilde{\sigma}_p^2 < \tilde{\sigma}_{p+1}^2$. From cases [iii] and [i] we have, respectively,

$$\tilde{\sigma}_p^2 = \frac{2+\lambda_p^2}{6B} = \frac{2+\lambda_p^2}{6\lambda_p} \text{ and } \tilde{\sigma}_{p+1}^2 = \frac{1-B^2+B\lambda_{p+1}}{\lambda_{p+1}} = \frac{1+\lambda_p^2}{2\lambda_p}.$$

The statement $\tilde{\sigma}_p^2 < \tilde{\sigma}_{p+1}^2$ is thus equivalent to

$$\frac{2+\lambda_p^2}{6\lambda_p} < \frac{1+\lambda_p^2}{2\lambda_p} \Longleftrightarrow 4+2\lambda_p^2 < 6+6\lambda_p^2 \Longleftrightarrow 0 < 1+2\lambda_p^2,$$

which is obviously true. This establishes the claim when B is a power of two.

If B is not a power of two, there exists an integer $p \ge 2$ such that $2^{p-1} < B < 2^p$ so that we can write $B = 2^{p-1} + J$, where $1 \le J \le 2^{p-1} - 1$; i.e., $\lambda_{p-1} < B \le \lambda_p - 1$ with $B = \lambda_{p-1} + J$ and $1 \le J \le \lambda_{p-1} - 1$. Since $B \le \lambda_{j-1}$ holds if $j \ge p+1$, it follows from the same 'case [i]' argument as before that $\tilde{\sigma}_j^2 < \tilde{\sigma}_{j+1}^2$ for all $j \ge p+1$; likewise, since $B > \lambda_j - 1$ holds for all $j \le p-1$, it follows from the same 'case [iii]' argument as before that $\tilde{\sigma}_j^2 < \tilde{\sigma}_{j+1}^2$ for all $j \le p-2$. The desired result thus holds if we can establish the $j = p-1$ and $j = p$ cases, i.e., that $\tilde{\sigma}_{p-1}^2 < \tilde{\sigma}_p^2 < \tilde{\sigma}_{p+1}^2$. From cases [i], [ii] and [iii] we have

$$\tilde{\sigma}_{p+1}^2 = \frac{1-B^2+4B\lambda_{p-1}}{4\lambda_{p-1}};\ \ \tilde{\sigma}_p^2 = \frac{12\lambda_{p-1}^2 B + 4\lambda_{p-1} - 6\lambda_{p-1}B^2 + B^3 - B - 4\lambda_{p-1}^3}{6B\lambda_{p-1}};$$

and $\tilde{\sigma}_{p-1}^2 = \frac{2+\lambda_{p-1}^2}{6B}$. Using $B = \lambda_{p-1} + J$, the inequality $\tilde{\sigma}_{p-1}^2 < \tilde{\sigma}_p^2$ is equivalent to

$$0 < 2\lambda_{p-1}^3 + 3\lambda_{p-1}J(\lambda_{p-1} - J) + \lambda_{p-1} + J(J^2 - 1).$$

Because $\lambda_{p-1} - J > 0$ and $J^2 - 1 \geq 0$, all terms on the right-hand side are nonnegative; moreover, since the first term is always greater than zero, the above expression is also such. We can now conclude that $\tilde{\sigma}_{p-1}^2 < \tilde{\sigma}_p^2$. Similarly, $\tilde{\sigma}_p^2 < \tilde{\sigma}_{p+1}^2$ is equivalent to

$$0 < 10J + 6\lambda_{p-1}(\lambda_{p-1}^2 - 1) + 18\lambda_{p-1}^2 J + 8\lambda_{p-1}J^2 + 10J^2(\lambda_{p-1} - J).$$

Because $\lambda_{p-1}^2 - 1 > 0$ and $\lambda_{p-1} - J > 0$, all terms on the right-hand side are positive, and hence the above expression is also such. This establishes our first claim.

Our second claim is that $\tilde{\sigma}_{q-1}^2 < \tilde{\sigma}_q^2$, which we establish by comparing the expression for $\tilde{\sigma}_q^2$ in [iv] with expressions for $\tilde{\sigma}_{q-1}^2$ in [i], [ii] and [iii] (note that $\lambda_{q-1} = M/2$):

[i] Here $B \leq M/4$ and $\tilde{\sigma}_{q-1}^2 = (2 - 2B^2 + BM)/M$, so $\tilde{\sigma}_{q-1}^2 < \tilde{\sigma}_q^2$ holds if

$$\frac{2 - 2B^2 + BM}{M} < \frac{3BM - 4B^2 + 4}{3M} \Longleftrightarrow B^2 > 1,$$

which is true because $B \geq 2$ always.

[ii] Here $M/4 < B < M/2$ and

$$\tilde{\sigma}_{q-1}^2 = \frac{12BM^2 + 16M - 24B^2M + 16B^3 - 16B - M^3}{24BM}.$$

Writing $B = \frac{M}{4} + J$, where $1 \leq J < M/4$, the claim $\tilde{\sigma}_{q-1}^2 < \tilde{\sigma}_q^2$ is true in this case if

$$12J^3 + M < \frac{M^3}{16} + \frac{3JM^2}{4} + 3J^2M + 12J,$$

which holds because (a) $12J^3 < 3JM^2/4$ since $J < M/4$ and (b) $M < 3J^2M$ since $J \geq 1$, thus establishing the claim in this case.

[iii] Here $B = M/2$ (the largest value it can attain because of the assumption $K < M/2$), so we have $\tilde{\sigma}_{q-1}^2 = (M^2 + 8)/12M$ and $\tilde{\sigma}_q^2 = (M^2 + 8)/6M$, from which it is obvious that $\tilde{\sigma}_{q-1}^2 < \tilde{\sigma}_q^2$ holds.

This establishes the second claim.

Finally, letting $\tilde{\sigma}_{q+1}^2 \equiv \bar{\sigma}_{q+1}^2/\sigma_\eta^2$, we claim that $\tilde{\sigma}_q^2 < \tilde{\sigma}_{q+1}^2$. An argument similar to that used to obtain (9) yields

$$\bar{\sigma}_{q+1}^2 = \frac{1}{M}\sum_{n=0}^{M-1} S_n \left|G_{q,n}^{(D)}\right|^2,$$

where $\{G_{q,n}^{(D)}\}$ is the DFT of the sequence of length M whose values are all equal to $1/\sqrt{M}$; i.e., $G_{q,0}^{(D)} = \sqrt{M}$ while $G_{q,n}^{(D)} = 0$ otherwise, thus yielding $\bar{\sigma}_{q+1}^2 = S_0$. From (13) we have

$$S_0 = \sigma_\eta^2 \left(1 + 2 \sum_{l=1}^{K} 1 - \frac{l}{K+1} \right) = \sigma_\eta^2 (K+1),$$

so $\tilde{\sigma}_{q+1}^2 = B$. Thus $\tilde{\sigma}_q^2 < \tilde{\sigma}_{q+1}^2$ is true if

$$\frac{3BM - 4B^2 + 4}{3M} < B \Longleftrightarrow B^2 > 1,$$

which is true because $B \geq 2$ always. This establishes the final claim and hence (10).

ACKNOWLEDGEMENTS

We would like to thank the referees for some very helpful comments that greatly improved this paper. The work of E. J. McCoy was supported through an EPSRC (UK) Grant, and that of D. B. Percival through an ONR ARL Grant (Applied Physics Laboratory, University of Washington) and an NSF Phase I SBIR Grant (MathSoft). The support is gratefully acknowledged.

REFERENCES

[1] M. S. Bartlett and D. G. Kendall, "The statistical analysis of variance-heterogeneity and the logarithmic transformation," *Suppl. J. R. Statist. Soc.,* vol. 8, pp. 128–38, 1946.

[2] I. Daubechies, *Ten Lectures on Wavelets.* Philadelphia: SIAM, 1992.

[3] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika,* vol. 81, pp. 425–55, 1994.

[4] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard, "Wavelet shrinkage: asymptopia? (with discussion)," *J. R. Statist. Soc. B,* vol. 57, pp. 301–69, 1995.

[5] W. A. Fuller, *Introduction to Statistical Time Series* (2nd Edition). New York: Wiley, 1996.

[6] H.–Y. Gao, "Wavelet estimation of spectral densities in time series analysis." PhD dissertation, Dept. of Statistics, University of California, Berkeley, 1993.

[7] H.–Y. Gao, "Choice of thresholds for wavelet shrinkage estimate of the spectrum," *J. Time Series Analysis*, vol. 18, pp. 231–51, 1997.

[8] C. W. J. Granger and M. Hatanaka, *Spectral Analysis of Economic Time Series.* Princeton: Princeton University Press, 1964.

[9] I. M. Johnstone and B. W. Silverman, "Wavelet threshold estimators for data with correlated noise," *J. R. Statist. Soc. B*, vol. 59, pp. 319–51, 1997.

[10] S. M. Kay, "Efficient generation of colored noise," *Proc. IEEE*, vol. 69, pp. 480–1, 1981.

[11] C. Kuo, C. Lindberg and D. J. Thomson, "Coherence established between atmospheric carbon dioxide and global temperature," *Nature,* vol. 343, pp. 709–14, 1990.

[12] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 11, pp. 674–93, 1989.

[13] E. J. McCoy and A. T. Walden, "Wavelet analysis and synthesis of stationary long-memory processes," *J. Comp. and Graph. Statist.,* vol. 5, pp. 26–56, 1996.

[14] P. Moulin, "Wavelet thresholding techniques for power spectrum estimation," *IEEE Trans. Signal Processing,* vol. 42, pp. 3126–36, 1994.

[15] G. P. Nason, "Wavelet shrinkage using cross-validation," *J. R. Statist. Soc. B*, vol. 58, pp. 463–79, 1996.

[16] A. Papoulis, *Probability, Random Variables, and Stochastic Processes.* New York: McGraw–Hill, 1991.

[17] D. B. Percival, "Simulating Gaussian random processes with specified spectra," *Computing Science and Statistics,* vol. 24, pp. 534–8, 1992.

[18] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications.* Cambridge: Cambridge University Press, 1993.

[19] M. B. Priestley, *Spectral Analysis and Time Series.* London: Academic Press, 1981.

[20] K. S. Riedel and A. Sidorenko, "Minimum biased multitaper spectral estimation," *IEEE Trans. Signal Processing,* vol. 43, pp. 188–95, 1995.

[21] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proc. IEEE,* vol. 70, pp. 1055–96, 1982.

[22] A. T. Walden, E. J. McCoy and D. B. Percival, "The effective bandwidth of a multitaper spectral estimator," *Biometrika,* vol. 82, pp. 201–14, 1995.

LIST OF FIGURES

Figure 1 – The autocovariance function $\tilde{s}_\eta(\nu)$ versus ν for $N = 2048$ and $K = 10$ sine tapers.

Figure 2 – Box plots of the estimated standard deviations $\hat{\sigma}_j$ of wavelet coefficients $n_{j,k}$ at scales 1, 2, 4 and 8 derived from the AR(2) process using different wavelet filters with $N = 2048$ and $K = 10$. The thick horizontal lines extending beyond each box plot indicate the value of σ_j derived from (9). The 'nominal' standard deviation $\sigma_\eta = \sqrt{[\psi'(10)]} \doteq 0.32$ (used in scale-independent thresholding) is marked as a thin horizontal line extending across each of the plots.

Figure 3 – Average value over 1000 simulations of the rmse given in Equation (12) for each of the four stationary processes described in Section IV. The average rmse is plotted for the scale-dependent method with the three types of thresholding, hard (thick solid line), mid (medium solid line), soft (thin solid line), and for the scale-independent method with the three types of thresholding, hard (thick dashed line), mid (medium dashed line), soft (thin dashed line). Three values of coarse resolution level, q_0, were considered, namely, 5, 6 and 8, corresponding to 64, 32 and 8 scaling coefficients left untouched by the thresholding. Series of length $N = 2048$ and the LA(8) wavelet filter were used.

Figure 4 – Estimated spectrum (thick curves) and true spectrum (thin) for the AR(24) process. The spectrum estimates are representative in that they have rmse closest to the average of the 1000 simulations. (a) Scale-independent, soft thresholding, and (b) scale-dependent, hard thresholding. $q_0 = 5$ in both cases. Series of length 2048 and LA(8) wavelet coefficients were used.

Figure 5 – As in Fig. 4, but for AR(2) process.

Figure 6 – As in Fig. 4, but for mobile radio communications process.

Figure 7 – As in Fig. 4, but for high order MA model.

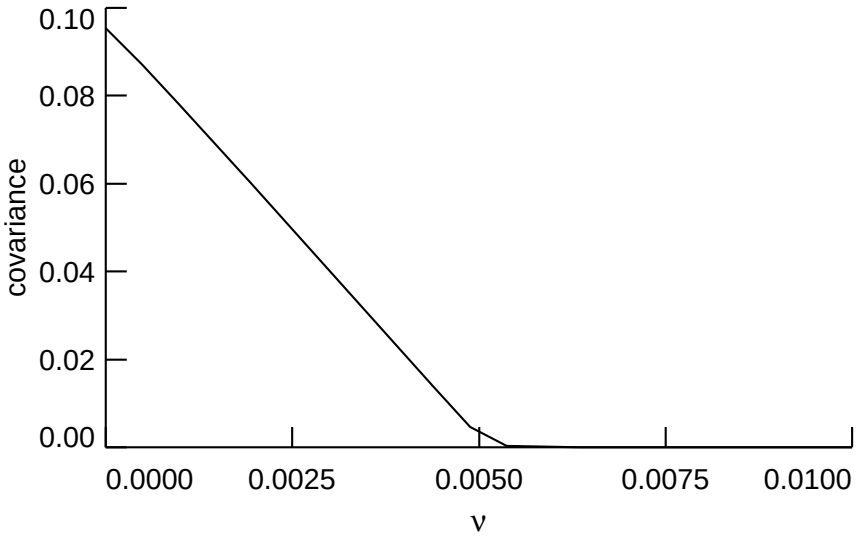
0.10
0.08
0.06
0.04
0.02
0.00
covariance
0.0000
0.0025
0.0050
0.0075
0.0100
ν

Figure 1.

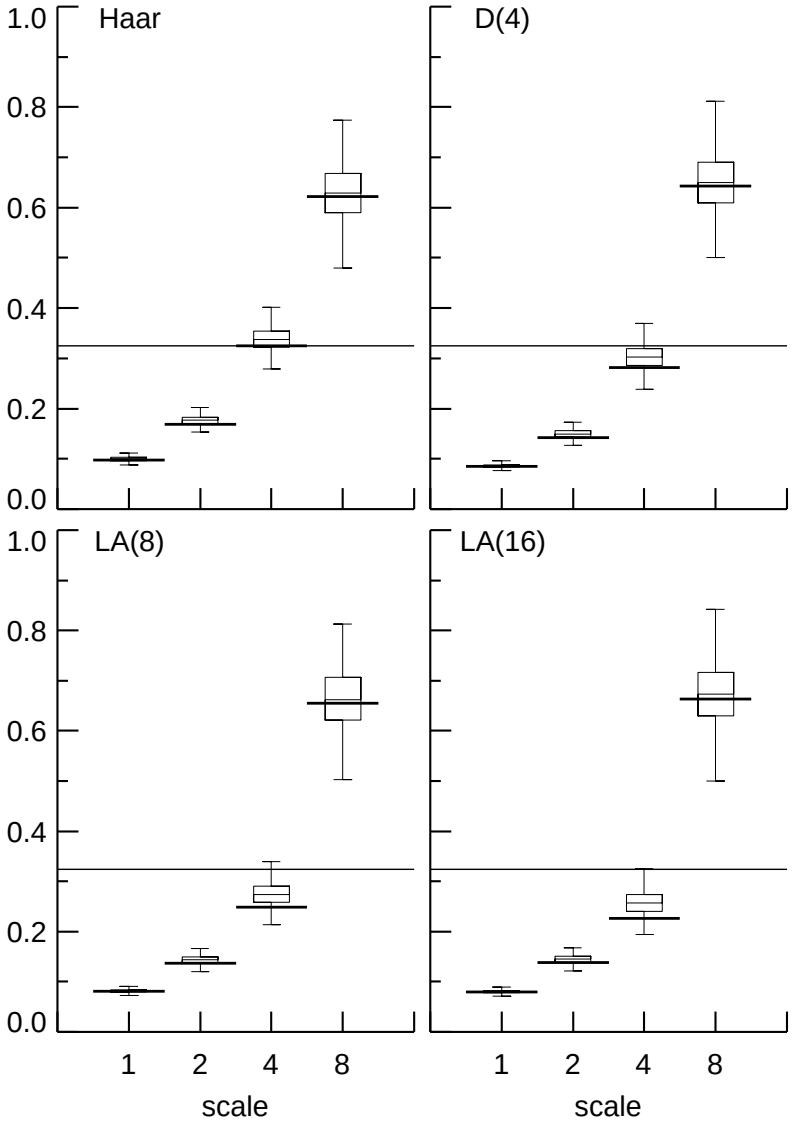

Figure 2.

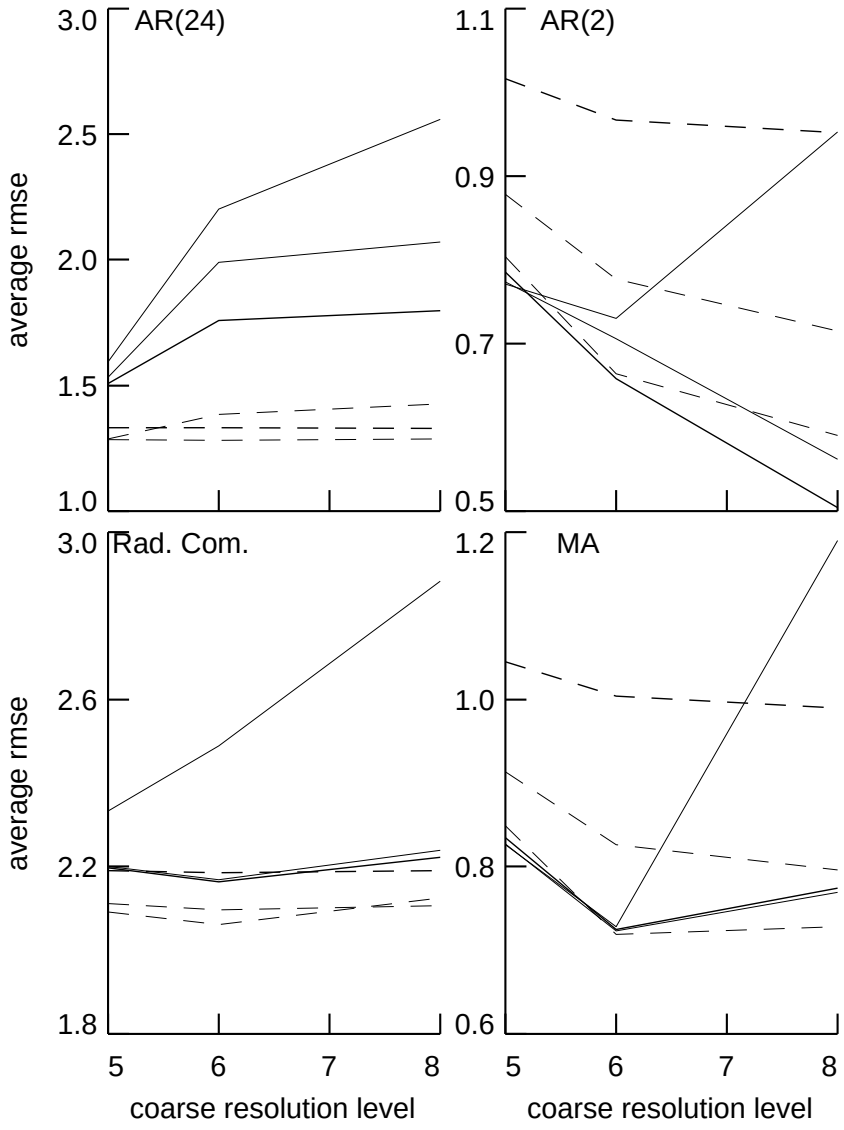
AR(24)
AR(2)
Rad. Com.
MA
average rmse
average rmse
3.0
2.5
2.0
1.5
1.0
1.1
0.9
0.7
0.5
3.0
2.6
2.2
1.8
1.2
1.0
0.8
0.6
5
6
7
8
coarse resolution level
coarse resolution level

Figure 3.

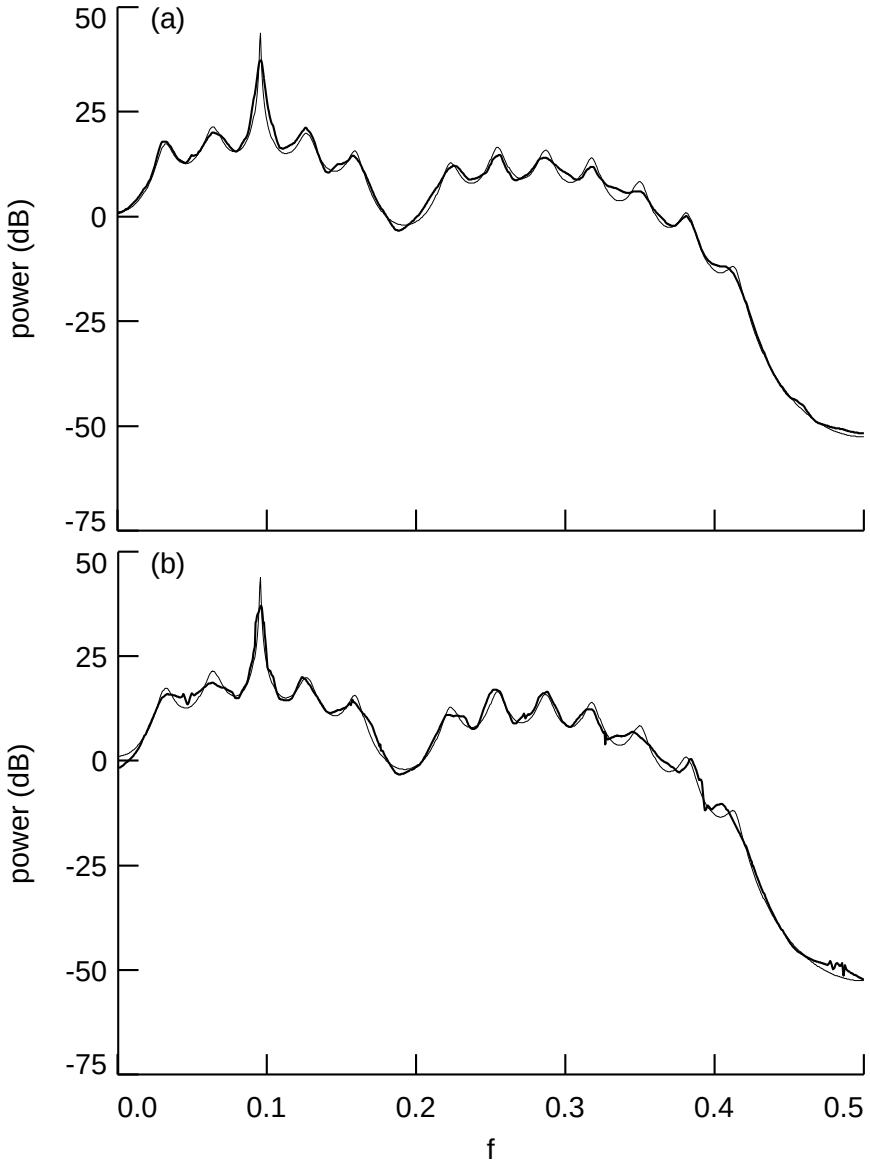

Figure 4.

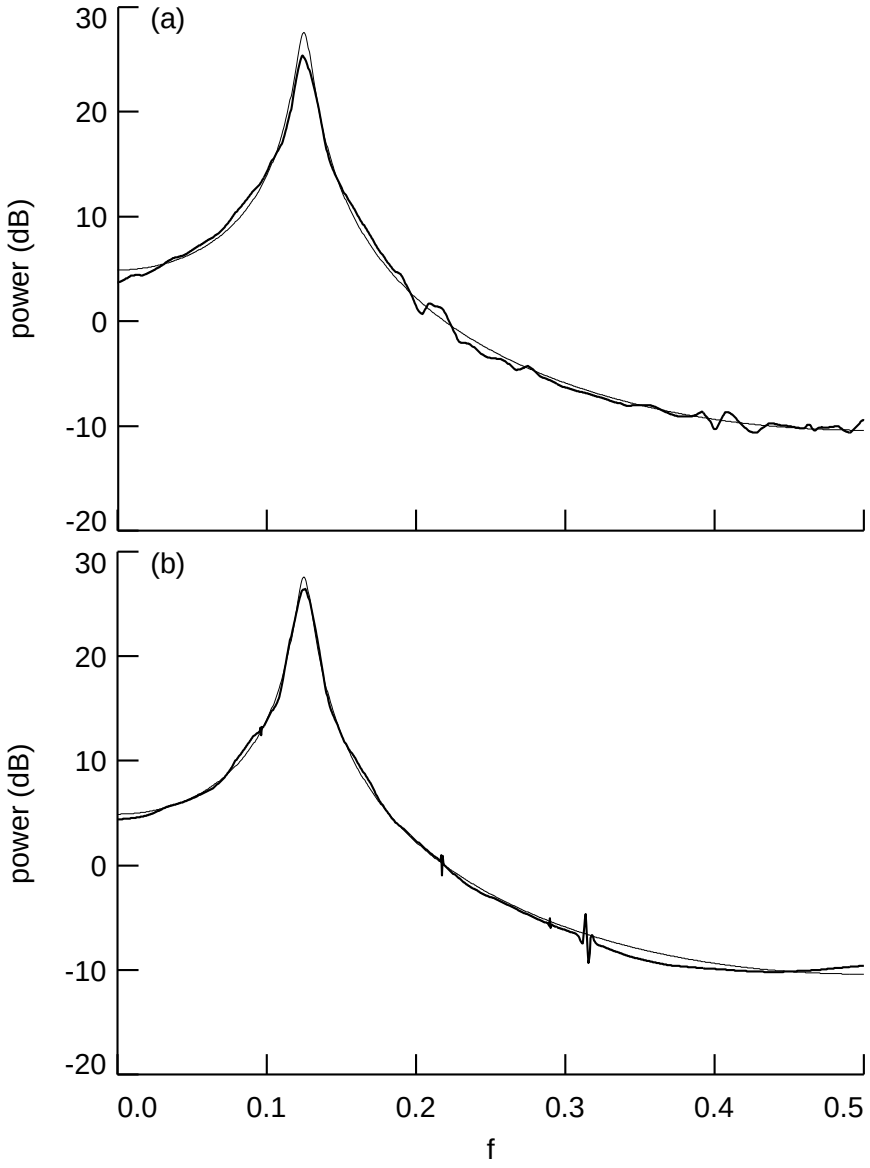

Figure 5.

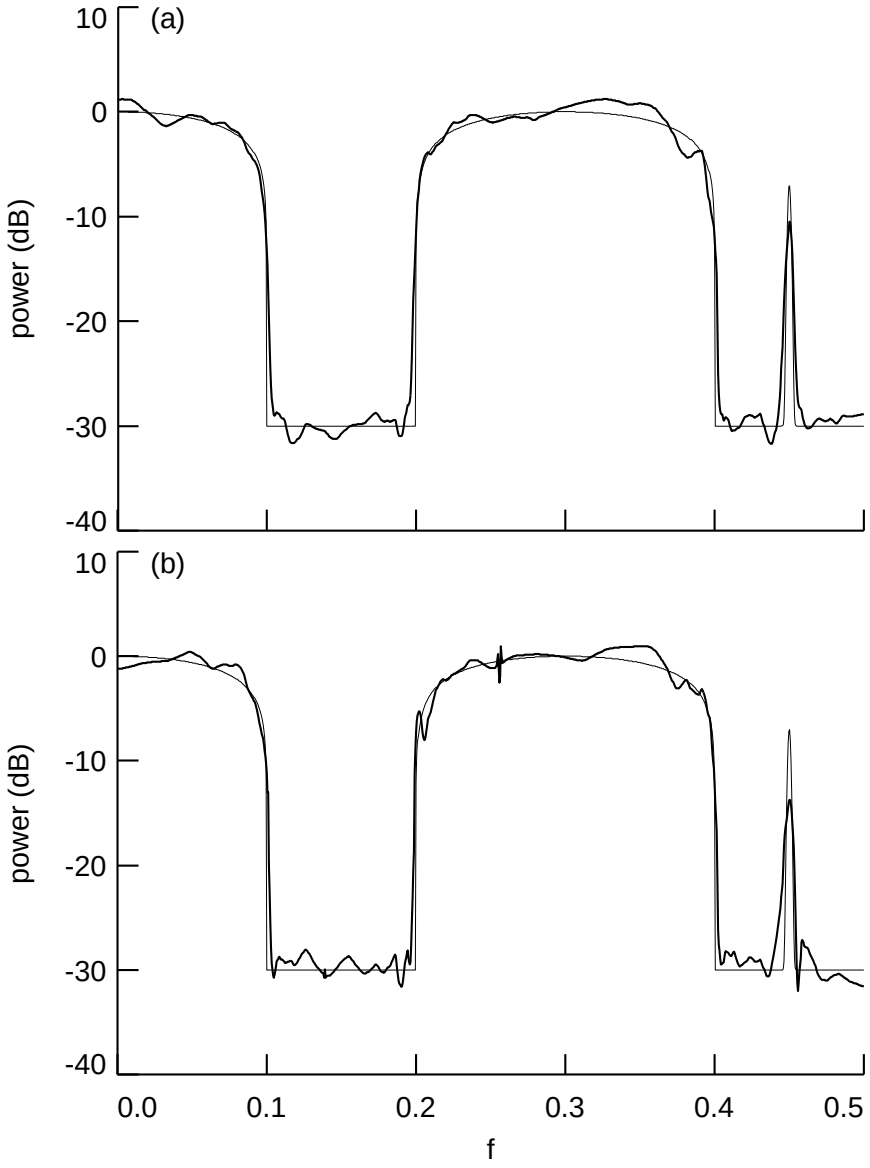

Figure 6.

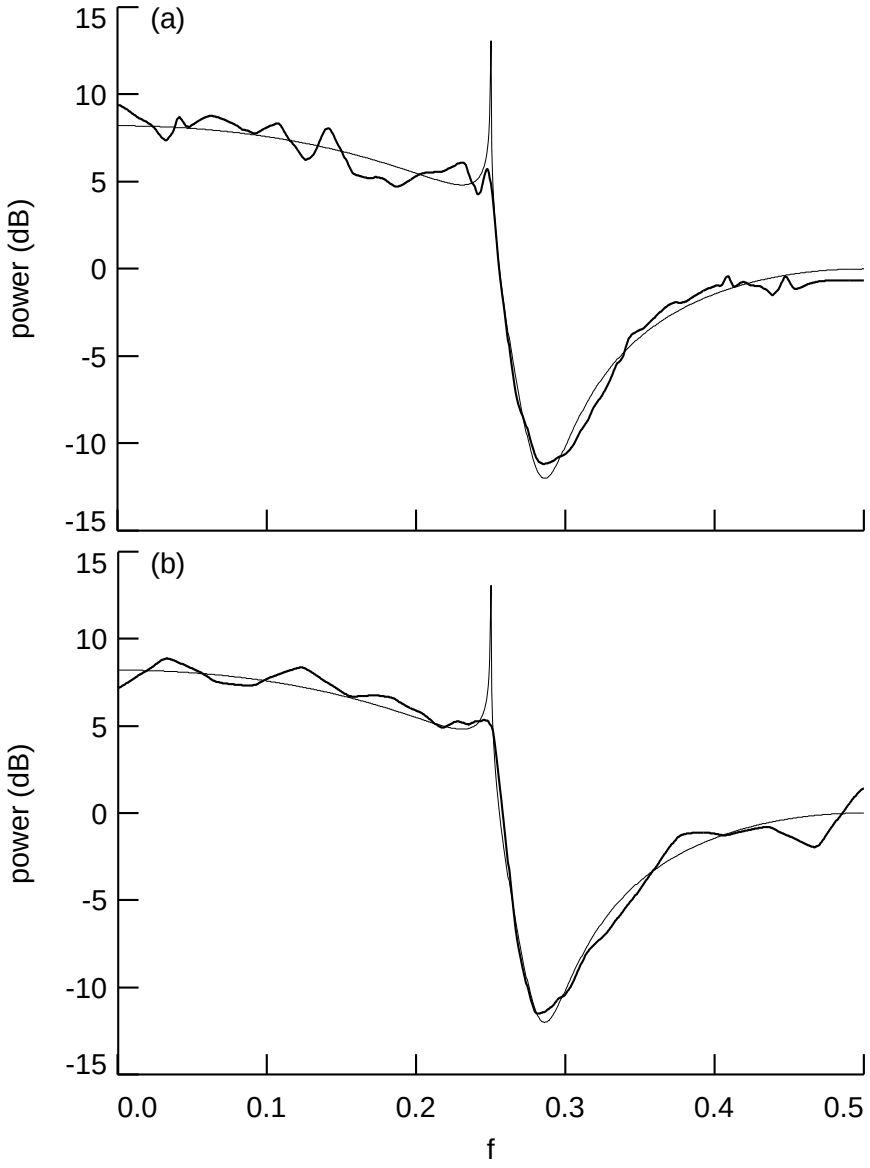

Figure 7.