J. R. Statist. Soc. B (1997)
59, *No.* 2, *pp.* 319–351

Wavelet Threshold Estimators for Data with Correlated Noise

By IAIN M. JOHNSTONE and BERNARD W. SILVERMAN†

Stanford University, USA *Bristol University, UK*

[Received January 1995. Final revision June 1996]

SUMMARY

Wavelet threshold estimators for data with stationary correlated noise are constructed by applying a level-dependent soft threshold to the coefficients in the wavelet transform. A variety of threshold choices is proposed, including one based on an unbiased estimate of mean-squared error. The practical performance of the method is demonstrated on examples, including data from a neurophysiological context. The theoretical properties of the estimators are investigated by comparing them with an ideal but unattainable 'bench-mark', that can be considered in the wavelet context as the risk obtained by ideal spatial adaptivity, and more generally is obtained by the use of an 'oracle' that provides information that is not actually available in the data. It is shown that the level-dependent threshold estimator performs well relative to the bench-mark risk, and that its minimax behaviour cannot be improved on in order of magnitude by any other estimator. The wavelet domain structure of both short- and long-range dependent noise is considered, and in both cases it is shown that the estimators have near optimal behaviour simultaneously in a wide range of function classes, adapting automatically to the regularity properties of the underlying model. The proofs of the main results are obtained by considering a more general multivariate normal decision theoretic problem.

Keywords: ADAPTIVE ESTIMATION; DECISION THEORY; ION CHANNELS; LEVEL-DEPENDENT THRESHOLDING; LONG-RANGE DEPENDENCE; MINIMAX ESTIMATION; NON-LINEAR ESTIMATORS; NONPARAMETRIC REGRESSION; ORACLE INEQUALITY; WAVELET TRANSFORM

1. INTRODUCTION

1.1. *Wavelet Threshold Estimates for White Noise*

Suppose that we are given n samples from a function f observed with noise:

$$Y_i = f(t_i) + e_i, \qquad i = 1, \ldots, n, \tag{1}$$

with $t_i = (i-1)/n$ and e_i drawn from some noise process. In the case where e_i is white noise, wavelet thresholding methods have been the subject of considerable attention: see, for example, Donoho and Johnstone (1994) who showed that an estimator based on thresholded wavelet expansions has desirable minimax properties.

The extension of most smoothing methods to deal with correlated data has not always been entirely straightforward, and often in the nonparametric smoothing literature the case of correlated data is not dealt with in any detail. In this paper, we consider problems in which the errors e_i may have very general correlation structure. We derive methods and results that are applicable to noise processes with both short- and long-range dependence.

†*Address for correspondence*: School of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK.
E-mail: B.W.Silverman@bristol.ac.uk

© 1997 Royal Statistical Society 0035–9246/97/59319

1.2. *Level-dependent Thresholding*

The simplest wavelet thresholding methods work by performing a wavelet transformation of the given data and then using the same threshold to deal with all the coefficients in the expansion (or at least all coefficients above a 'primary resolution level' below which no thresholding at all is carried out). However, if the noise in the data is stationary and correlated, then the variance of the wavelet coefficients will depend on the level in the wavelet decomposition but will be constant within each level. Therefore it will be natural to deal differently with the coefficients at each level, and to use a level-dependent approach.

Even though the original data may be highly autocorrelated, the wavelet transform will very often yield an array of coefficients that exhibits much less dependence. This is even the case where the noise exhibits long-range dependence, e.g. having an autocorrelation function that decays as $t^{-\alpha}$ for some α in the interval (0, 1). We illustrate this by consideration of a particular example (in Section 2.2) and later in the paper by a more general theoretical discussion (in Section 6.3). This 'decorrelating' feature of the wavelet transform gives intuitive support to an approach where each wavelet coefficient is thresholded separately.

1.3. *Outline of the Paper*

We proceed in two complementary ways. On the one hand, we exercise the level-dependent thresholding principle on some examples in Section 3, mainly concentrating on some data generated in a neurophysiological context. Our discussion incorporates some other recent methodological developments such as the use of a translation invariant transform and the use of an unbiased risk threshold estimate. In preparation, Section 2 briefly reviews relevant aspects of Donoho and Johnstone (1994) and properties of the wavelet transform when e_i are drawn from a stationary process.

On the other hand, we carry out a detailed theoretical investigation which aims to show that the variants of level-dependent thresholding presented in Sections 2 and 3 have good large sample adaptivity properties under very mild conditions on the form of the correlation. In Section 4, the performance of the estimator relative to an unattainable 'bench-mark' risk is quantified. The bench-mark risk is that of an idealized diagonal projection estimate, where individual coefficients of the wavelet transform are 'kept' or 'killed' according to instructions from an 'oracle'. This risk can be considered in its own right as a measure of the sparsity, or economy of representation, of the wavelet expansion of the signal f. Estimates constructed using level-dependent thresholding are shown to have risk at most $O(\log n)$ times the bench-mark. Further, Section 5 shows that, under very mild conditions, the behaviour of the threshold estimator relative to the bench-mark risk cannot be improved on by *any* estimator based on the given data Y.

Section 6 investigates models for both short- and long-range dependent noise and indicates the way in which standard models correspond to properties of the array of wavelet coefficients. These models are used in Section 7 to translate the performance of the level-dependent thresholded wavelet estimators relative to the bench-mark risk into a near optimality property simultaneously over a wide range of function classes. This indicates that the estimators adapt automatically to the smoothness properties of the underlying function.

Some concluding remarks are made in Section 8, and the more technical details and proofs are given in Appendix A. Although our main interest is in their consequences for wavelet estimators, the key theoretical results of the paper are obtained in the context of a more general problem in multivariate normal decision theory, concerned with the estimation of an n-vector θ of parameters from a vector observation $X \sim N(\theta, V)$.

For further reading on the subject of wavelets and their applications in statistics, and in particular of the role of minimax results in the area, the reader is referred to Donoho *et al*. (1995) together with its published discussion and numerous references. For a discussion focused specifically on wavelet estimators of signals in stationary correlated noise, see Brillinger (1994, 1996).

2. WAVELET TRANSFORMS AND ESTIMATORS

2.1. *Basic Definitions and Notation*

We first establish some notation and recall the definition of the Donoho–Johnstone estimator for the white noise case. Let $\mathcal{W}$ be a periodic discrete wavelet transform operator, and let Y be the n-vector of observations $Y_1, \ldots, Y_n$. We suppose that $n = 2^J$ for some J. Write

$$w_{jk} = (\mathcal{W}Y)_{jk} \qquad j = 0, 1, \ldots, J-1, \quad k = 1, \ldots, 2^j \tag{2}$$

with the remaining element labelled w_{-1}. Let $\theta = \mathcal{W}\mathbf{f}$ be the wavelet transform of the signal $\mathbf{f} = (f\{(i-1)/n\})_{i=1}^n$ and $\mathbf{z} = \mathcal{W}\mathbf{e}$ be the wavelet transform of the noise.

To construct the estimator, define η_S to be the *soft threshold function*

$$\eta_S(w, \lambda) = \mathrm{sgn}(w)(|w| - \lambda)_+. \tag{3}$$

Suppose that the e_i in model (1) are independent identically distributed (IID) $N(0, \sigma^2)$ random variables, for some known σ^2. The Donoho–Johnstone estimator is then constructed by soft thresholding the wavelet coefficients w_{jk} at threshold λ, e.g. $\lambda = \sigma\sqrt{(2 \log n)}$, and then transforming back. Thus we define $\hat{\theta}$ by

$$\hat{\theta}_{jk} = \eta_S(w_{jk}, \lambda) \tag{4}$$

and the estimator $\hat{\mathbf{f}}$ by

$$\hat{\mathbf{f}} = \mathcal{W}^T\hat{\theta}. \tag{5}$$

In practice the transformations $\mathcal{W}$ and the inverse transform $\mathcal{W}^T$ are carried out by a fast $O(n)$ algorithm. For more details see, for example, Mallat (1989) or Nason and Silverman (1994). Some remarks about the choice of the factor $\sqrt{(2 \log n)}$ in the threshold will be made in Section 2.3.1 later.

Two further comments are in order. In applications we may wish to replace the soft threshold function η_S by a *hard* threshold

$$\eta_H(w, \lambda) = w\, I\{|w| \geqslant \lambda\},$$

or some compromise between the two. Hard thresholding generally preserves peak heights better, but at some cost in visual smoothness. The theoretical results of this paper focus on soft thresholding, but analogues of all mean-squared error results

could be developed for hard thresholding; see, for example, Donoho and Johnstone (1994).

Secondly, in practice, thresholding is usually restricted to levels j above some user-specified primary resolution level L, below which it is felt that signal predominates over noise.

2.2. *Wavelet Coefficients of Coloured Noise*

Before discussing the extension of the thresholding idea to deal with correlated noise, it will be useful to consider informally the effect of the wavelet transform on a correlated noise process. Suppose therefore that the errors e_i have a multivariate normal distribution with mean 0 and covariance matrix Γ_n. We shall assume that the errors are stationary so that Γ_n has entries $r^{(n)}_{|s-t|}$, say. For simplicity in our main discussion, we shall also assume periodicity where necessary, so that Γ_n is a circulant matrix; there are then no boundary effects in the distribution of the wavelet coefficients of the noise process. For reasons given in Section 8, the stationarity assumption is not essential, and it is also straightforward to extend both the methods and the results of the paper to the non-periodic case.

Let V_n be the covariance matrix of the vector $\mathbf{z}$, the wavelet transform of the error vector $\mathbf{e}$, so that V_n is obtained from Γ_n by the orthogonal transform

$$V_n = \mathcal{W}\Gamma_n\mathcal{W}^{\mathrm{T}}. \tag{6}$$

The filters used to construct the discrete wavelet transform are time invariant. Hence (neglecting boundary effects) within each level the distribution of the z_{jk} will be stationary, and the variance of z_{jk} will depend only on the level j. We write $\sigma_j^2 = \mathrm{var}(z_{jk})$ for each j.

The properties of the wavelet transform have two heuristic consequences, which are by no means essential for the development of our practical method or for its theoretical justification, but they will be illuminating in the subsequent discussion. Firstly, for many models likely to be of relevance in practice, the autocorrelation of the z_{jk} within each level dies away rapidly. Qualitatively, this is a consequence of the fact that wavelets are 'almost eigenfunctions' of many operators; see Frazier *et al.* (1991) and Meyer (1990). Secondly, because the mapping from the original series to the wavelet coefficients on any particular level is essentially a band pass filter, there will tend to be little or no correlation between the wavelet coefficients at different levels. (Some further details are provided for the reader's convenience in Appendix A.)

To illustrate the various properties on an example, 2048 points were simulated from a mean-zero long-range dependent process with spectral density approximately proportional to frequency$^{-0.9}$. These were regarded as observations taken at time points $j/2048$, $j = 1, \ldots, 2048$. A plot of the generated series is given in Fig. 1(a), and the discrete wavelet transform of the series is plotted in Fig. 1(c). In Fig. 1, the wavelet coefficients on each scale have been scaled separately; the sample standard deviations of the coefficients on each level are plotted in Fig. 1(d), and it can be seen that the log-variances decrease roughly linearly as the level increases.

It appears visually from Fig. 1(c) that on each level the wavelet coefficients are approximately white noise. To provide vindication of this, and to illustrate the lack of correlation between levels, see Fig. 2. This illustrates the autocorrelation function

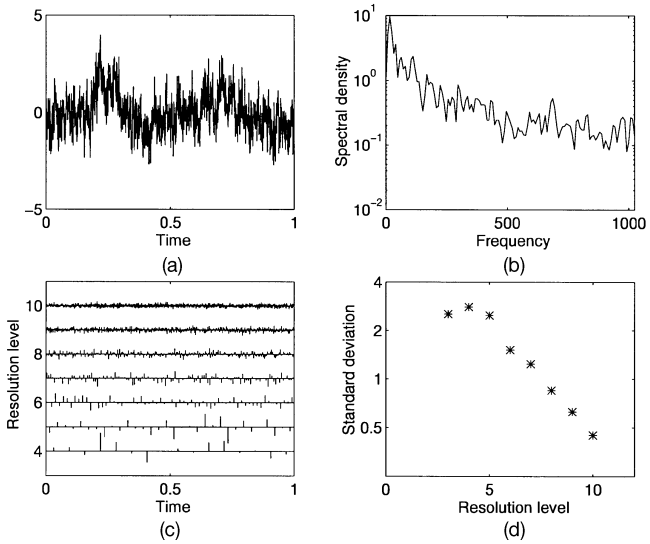

Fig. 1. (a) Generated process exhibiting long-range correlation; (b) estimated spectral density of the data shown in (a); (c) discrete wavelet transform of the series using Daubechies's nearly symmetric wavelet of order 8, with coefficients scaled relative to the largest coefficient at all levels; (d) sample standard deviations (plotted on a logarithmic scale) of the various levels of the discrete wavelet transform shown in (c)

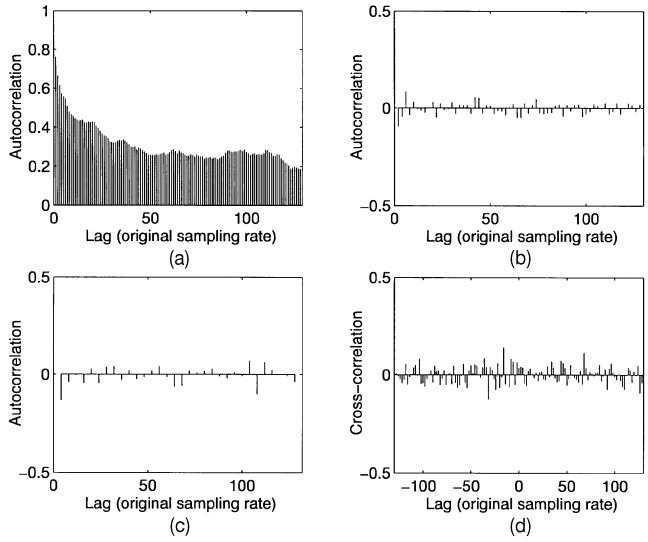

Fig. 2. (a) Sample autocorrelation function of series illustrated in Fig. 1(a); (b) sample autocorrelation function of wavelet coefficients at level 10 of the discrete wavelet transform of the same series; (c) sample autocorrelation function of wavelet coefficients at level 9; (d) sample cross-correlations of coefficients between these two levels

of the original series and of the wavelet coefficients at levels 9 and 10, as well as the cross-correlation function between these two levels. (The autocorrelations at lag 0 have been omitted.) It can be seen that the long-range correlations in the original series are eliminated by the wavelet transform, and that the wavelet coefficients form series with virtually no autocorrelation or cross-correlation. Note that the timescale on which the various autocorrelations and cross-correlations are calculated is always that of the original series.

2.3. *Coloured Noise Estimators*

We can now set out some proposed estimators for data with coloured noise. In view of the discussion in Section 2.2, a natural extension of the usual wavelet thresholding method is to apply *level-dependent thresholding* to the transformed data $\mathbf{w}$. Let λ_j be a sequence of thresholds to be applied to the coefficients at level j, and define $\hat{\theta}$ to be the estimator

$$\hat{\theta}_{jk} = \eta(w_{jk}, \sigma_j \lambda_j).$$

Here η denotes soft or hard thresholding, or some compromise between the two. We write $\hat{\theta}$ for the corresponding estimator of θ, and set

$$\hat{\mathbf{f}} = \mathcal{W}^{\mathrm{T}} \hat{\theta}.$$

Under this formulation, allowing signal at low levels through without thresholding corresponds to setting $\lambda_j = 0$ for the relevant j. At higher levels, where there is a considerable number of coefficients at each level and the signal θ_{jk} can be assumed to be sparse, the noise variance σ_j^2 at each level can be estimated from the data. One possibility is to use a robust estimator such as

$$\hat{\sigma}_j^2 = \mathrm{MAD}\{w_{jk}, k = 1, \ldots, 2^j\}/0.6745. \tag{7}$$

where MAD denotes median absolute deviation from 0 and the factor 0.6745 is chosen for calibration with the Gaussian distribution. Other estimates are of course possible, e.g. the *mean* absolute deviation. We do not dwell on the estimation of the variance; we assume for the rest of this section that it has been carried out, and we treat σ_j^2 as known.

In most of our subsequent discussion, we measure loss in the L^2-sense and define the risk measure of an estimator by $R(\hat{\theta}, \theta) = E\|\hat{\theta} - \theta\|^2$, where the norm is the usual Euclidean norm. Since the discrete wavelet transform is orthogonal, the risk of an estimator will be the same as that of its discrete wavelet transform and so risk results obtained in the wavelet domain carry over directly to the original 'time' domain.

We now list several options that are available in constructing a thresholding estimator. Naturally, no one prescription will be best for all settings. However, all exploit the spatial adaptivity that is inherent in wavelet bases.

2.3.1. *'Universal' threshold*

A conservative choice of threshold that is attractive from certain theoretical perspectives is

$$\lambda_j = \sqrt{(2 \log n)}. \tag{8}$$

A threshold proportional to $\sqrt{(2 \log n)}$ is 'conservative', for the following reason. If $Z_1, \ldots, Z_n$ are normally distributed random variables with mean 0 and variances σ_i^2, then

$$P\left\{ \max_{1 \leqslant i \leqslant n} |Z_i/\sigma_i| > \sqrt{(2 \log n)} \right\} \to 0 \qquad \text{as } n \to \infty \tag{9}$$

whether or not the variables are independent. In all cases, the expected number of $|Z_i/\sigma_i|$ that exceed $\sqrt{(2 \log n)}$ tends to 0, and hence the probability in expression (9) tends to 0. Under independence, this probability is easily bounded above by $(4\pi \log n)^{-1/2}$. The probability in the correlated case is bounded above by that in the independence case, using a result of Khatri (1967) and Sidák (1968).

Thus, in the limit, no pure noise variables will be let through the threshold. For a very wide range of values of n (certainly $64 = 2^6 \leqslant n \leqslant 2^{19}$) the expected number of $|Z_i|$ that exceed the threshold will be between 0.075 and 0.125, so only in about a tenth of realizations will *any* pure noise variables exceed the threshold. In this sense the thresholding method gives a 'noise-free' reconstruction.

2.3.2. *Unbiased risk estimate*

The conservative properties of equation (8) come at the price of high threshold levels: in terms of L^2-loss, better performance is obtained with smaller thresholds. In the IID Gaussian error setting, Donoho and Johnstone (1995) showed that the method of Stein (1981) leads to an unbiased estimate (labelled SURE, for Stein unbiased risk estimate) for the mean-squared error of soft thresholding for each possible threshold choice λ. A data-based threshold choice can then be obtained simply by minimizing the estimate with respect to λ over the range $[0, \sigma\sqrt{(2 \log n)}]$.

It is widely known, from work of Hart (1991) and others, that it is in general dangerous to apply to correlated data unbiased risk criteria that are derived under the assumption of independent errors. However, a special feature of thresholding estimates turns out to imply in our setting that the risk estimate remains unbiased, even in the presence of correlation. To be specific, suppose that $X \sim N_d(\theta, V)$. Stein's method shows that the mean-squared error of an estimator $\hat{\theta} = X + g(X)$ may be written

$$\begin{aligned} E\|X + g(X) - \theta\|^2 &= E[\operatorname{tr}(V) + \|g(X)\|^2 + 2 \operatorname{tr}\{V \operatorname{Dg}(x)\}] \\ &= E\{U(t; X)\}, \end{aligned} \tag{10}$$

say, where $\operatorname{Dg}(X)$ denotes the $d \times d$ matrix with entries $\partial g_i/\partial x_j(X)$. In the case of soft thresholding at t, the kth component of g is

$$g_k(x) = \begin{cases} -t & x_k > t, \\ -x_k & |x_k| \leqslant t, \\ t & x_k < -t. \end{cases}$$

The key point is that thresholding operates co-ordinatewise, so that g_k is a function of x_k alone, and the matrix Dg in equation (10) is therefore diagonal.

If the covariance matrix V is homoscedastic, $\sigma_{kk} \equiv \sigma^2$, and so can be estimated by $\hat{\sigma}^2$ as defined in equation (7). The unbiased risk criterion is an estimate of $U(t; X)$ obtained by substituting the properties of g and this estimate of σ_{kk}, to obtain

$$\hat{U}(t) = \hat{\sigma}^2 d + \sum_k (x_k^2 \wedge t^2) - 2\hat{\sigma}^2 I\{|x_k| \leqslant t\}, \tag{11}$$

which is identical with that used in the IID case! We therefore propose to take

$$\hat{t}(x) = \underset{0 \leqslant t \leqslant \sigma\sqrt{(2 \log d)}}{\arg\min} \{\hat{U}(t)\}. \tag{12}$$

As explained in Donoho and Johnstone (1995) this minimization can easily be accomplished in $O(d \log d)$ time.

In the wavelet thresholding setting, we apply this prescription separately on each level to the coefficients $w_j = \{w_{jk}, k = 1, \ldots, 2^j\}$. The stationarity assumption implies the homoscedasticity condition needed in the derivation of equation (11). We then set

$$\lambda_j = \sigma_j \hat{t}(w_j/\sigma_j), \qquad L \leqslant j \leqslant J - 1.$$

2.3.3. *Translation invariant estimates*

The estimator that we have described is not translation invariant—there is an arbitrariness to the dyadic boundaries implicit in the wavelet filter cascade which becomes particularly evident in cases with sporadic discontinuities and low signal-to-noise ratios.

Coifman and Donoho (1995) have described a translation invariant wavelet denoising algorithm. In outline, if S represents a (circular) shift operator of one time unit, we can compute, in principle, all translated fits $S^{-k} \circ \hat{f} \circ S^k$, where $\hat{f}$ is an ordinary wavelet threshold estimate. A translation invariant estimate is obtained by averaging all the n translated fits. Although this method would at first sight appear to be $O(n^2)$, Coifman and Donoho (1995) observed (as have other researchers) that the translation invariant wavelet transform can be computed and inverted by a variant of the wavelet packet algorithm in $O(n \log n)$ time. The transform itself is a table of $\log_2 n$ rows by n columns, and the denoising is accomplished by thresholding the entries in this table before inverting. It is then natural, in our correlated data setting, to choose separate thresholds on each of the $\log_2 n$ levels (above some base level L). Of course, these thresholds may be chosen using either prescription above.

3. EXAMPLES

We illustrate the thresholding methods described in Section 2.3 on two examples. Software implementing these methods and scripts reproducing the figures will be available in releases .800 and later of the library *WaveLab* of MATLAB-based routines for wavelet and related time frequency–timescale analyses available from `http://stat.stanford.edu` on the World Wide Web.

3.1. *Doppler Signal*

This artificial function of spatially varying frequency was used by Donoho and Johnstone (1994) to illustrate the spatial adaptivity of wavelet shrinkage in the presence of IID Gaussian noise. Here, the signal is sampled at 2048 equally spaced points $t_i = i/2048$ for $i = 1, \ldots, 2048$. Noise $\{e_i, i = 1, \ldots, 2048\}$ generated from two Gaussian processes is added. The first is an AR(2) process $e_i =$

$(4/3)e_{i-1} - (8/9)e_{i-2} + w_i$ driven by IID Gaussian noise $\{w_i\}$, and scaled so that $\{e_i\}$ has unit variance. The second is the $1/f^{0.9}$ type of noise used earlier in Fig. 1, again normalized to unit variance. For clarity, only portions of the processes are shown in Figs 3(a) and 3(b). Figs 3(c) and 3(d) show that the distribution of noise over scales is quite different in the two cases. The reconstructions shown in Figs 3(e) and 3(f) were obtained using soft thresholding, scales estimated from equation (7) and 'universal thresholds' from equation (8) at levels $L = 6$ and above. The estimators are broadly

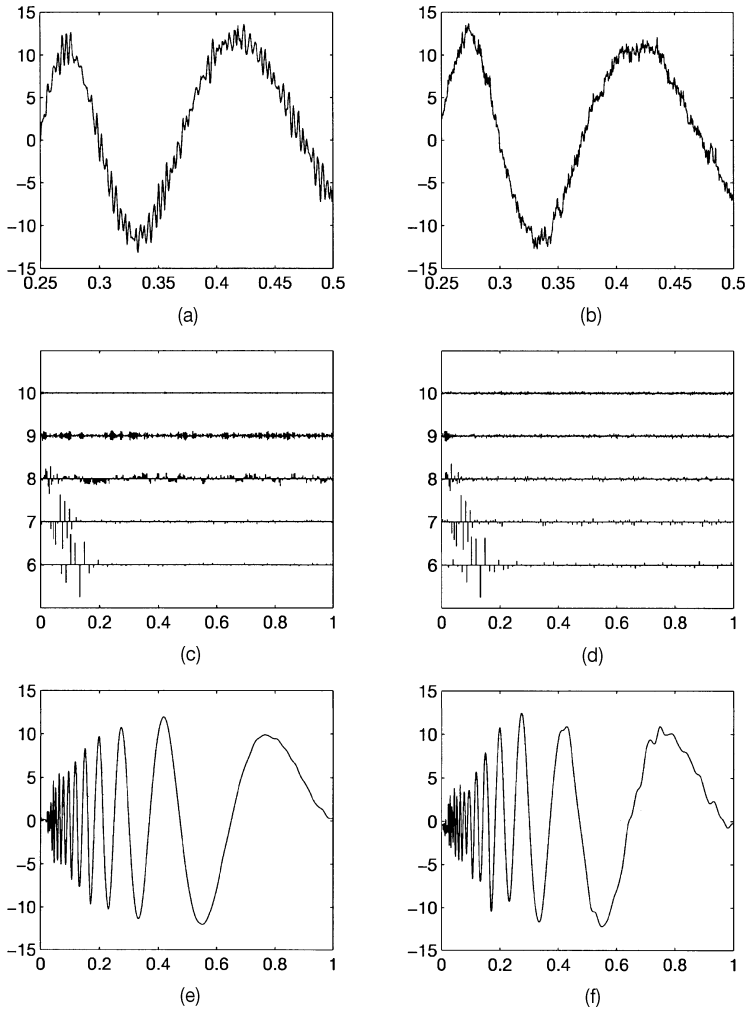

Fig. 3. (a), (b) Portion of the Doppler signal with addition of AR(2) and $1/f^{0.9}$ noise respectively (the full signal has $t_i = i/n \in [0, 1]$, for $i = 1, \ldots, n = 2048$); (c), (d) wavelet coefficients of the two full noisy signals; (e), (f) reconstructions obtained using $\lambda_j = \hat{\sigma}_j\sqrt{(2 \log n)}$, soft thresholding and Daubechies's nearly symmetric wavelet of order 8

comparable with those obtained in the IID case, and in particular show the spatial adaptation achieved by thresholding in the wavelet domain.

3.2. *Example from Neurophysiology*

An important problem in molecular physiology is the detection and measurement of the picoamp currents that flow through the single membrane channels that control movement in and out of cells. Bob Eisenberg, a physiologist, has encouraged the application of newer signal processing techniques to this problem; see Eisenberg (1994). In collaboration with his colleague Rick Levis, he has made available some generated data intended to represent most of the relevant challenges in processing such single-channel data. Fig. 4(a) shows part of an extract of length 4096 from the data file of 100 000 points provided by Eisenberg and Levis. The data consist of a step function switching between values 0 ('off') and 1 ('on') at random, with average period about 125 points, and measured in the presence of additive correlated noise. The noise contains white and so-called 'f'- and 'f^2'-components in strengths known to be representative of laboratory data and is then passed through a digital finite impulse response Gaussian filter intended to simulate a laboratory antialiasing filter.

This generated example differs in kind from the simulated data discussed in Section 3.1 in that its underlying model is carefully selected by practitioners directly involved in routine collection and analysis of real data. The obvious advantage of

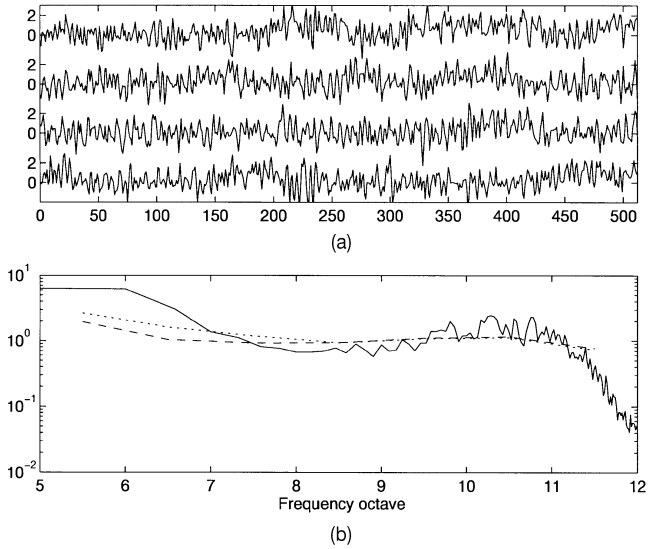

Fig. 4. (a) Subset of 2048 data points (2025–4072) from a record of 100 000 points provided by R. Eisenberg and R. Levis, with successive segments of length 512 shown from bottom to top; (b) estimate of the power spectral density (———) of the observed data y (a sequence of length 4096, being data points 1001–5096), with both amplitude and frequency shown on logarithmic scales, frequency coded according to frequency octaves, right-hand end point 50 kHz; median absolute deviation (– – –) and standard deviation (·····) of the Haar wavelet coefficients of the observed data, plotted against resolution level (plus 0.5) on the same vertical scale

using a generated data set rather than an actual data set obtained in practice is that for a 'real' data set the 'truth' is not known, and so it is impossible to quantify the performance of the various methods.

Our initial analysis will be based on a segment of length 4096. We find that these data $Y_i = \theta_i + e_i$ have a rather low signal-to-noise ratio $\rho = \mathrm{SD}(\theta)/\mathrm{SD}(e) = 0.48/0.88 = 0.55$. Fig. 4(b) shows the power spectral density of the data with frequency (as well as amplitude) plotted on a logarithmic scale by octave. The evidently coloured spectrum suggests that wavelet coefficients at different resolution levels will have different scales, and this is confirmed by the scales also shown on Fig. 4(b). The scale estimates $\hat{\sigma}_j$ were estimated at each level as in equation (7) from the ordinary (decimated) wavelet transform, using the median absolute deviation from the mean. Because of the very high sampling rate, the stationarity assumption is reasonable here, both in these generated data and also in the biological system, because the data considered represent only a fraction of a second. Normal probability plots support the Gaussian assumption.

Naïve application of the wavelet thresholding prescription (8) with thresholds $\hat{\sigma}_j\sqrt{(2 \log 4096)}$ produces a poor estimate, presumably because of the low signal-to-noise ratio and the non-translation-invariant structure.

The broken curve in Fig. 5(a) shows the result of applying *translation invariant* denoising with the Haar wavelet and (hard) thresholds $\hat{\sigma}_j\sqrt{(2 \log n)}$, setting the base level $L = 6$. The conservative threshold multiplier $\sqrt{(2 \log n)}$ has clearly led to some oversmoothing as the price for noise suppression. Smaller thresholds may be more appropriate in seeking the best trade-off of noise and resolution in this particular setting.

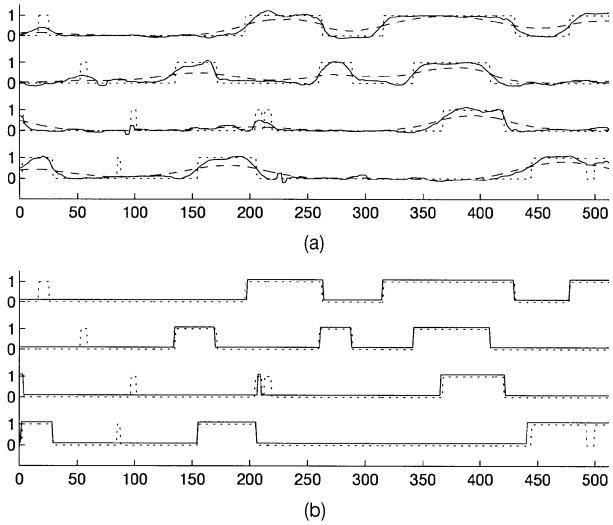

Fig. 5. (a) True on–off step function (·····), translation invariant reconstruction using Haar wavelets, $n = 4096$ and universal thresholds (– – – –) and similar translation invariant reconstruction using smaller thresholds as described in the text (———); (b) true on–off step function (·····) and estimate (———) produced by rounding the full curve in (a) to the nearer of the values 0 and 1, raised (for visual clarity) by 0.1 units relative to the dotted curve

The full curve in Fig. 5(a) shows the corresponding reconstruction using the *smaller* thresholds chosen by the Stein unbiased risk criterion (11) in bands $j = 6, 7, 8$. Evidently the fit is somewhat noisier but is more faithful to the underlying on–off signal. The unbiased risk estimate was not applied to the top three levels ($j = 9, 10, 11$): the signal has little power in these frequency bands and so the protection from noise provided by the conservative property (9) was considered valuable here.

Since the underlying cell current is known to be an on–off step function, it is natural to threshold the fits obtained above at level 0.5 to produce an estimate that also takes values 0 and 1. When processed in this way the SURE threshold estimate leads to the full curve in Fig. 5(b). The dotted curve in both parts of Fig. 5 is the true input signal. The on–off estimate has a 2.6% error rate (105 classification errors in 4096 points) on the selected data.

Table 1 compares the error rates of the various methods discussed on 10 successive segments of length 4096 from the original data record. Also shown is the error rate of the current special purpose method developed specifically by Eisenberg and Levis for this problem. Although translation invariant level-dependent wavelet thresholding is not yet quite as good as the special purpose filtering method, it is encouraging that relatively straightforward adaptation of the general wavelet thresholding prescription achieves reasonable results in this special context. The special purpose method has no promise of working on any other kind of data, whereas the methodology that we propose is part of a very generally applicable toolkit. It can be seen that both the special purpose method and the wavelet method perform better on the average of 10 successive runs than they do on the individual run considered in detail above; clearly that particular run is a relatively difficult one to reconstruct, whichever method is used. The approach is not limited to 0–1 currents, of course; indeed there are cases (Fredkin *et al.*, 1995) where neither the number of states nor their current levels are known *a priori*. Thus, future work in collaboration with the physiologists seems likely to yield further improvements in the methods.

4. ORACLE INEQUALITIES AND IDEAL RISKS

4.1. *Preliminary Remarks*

The remainder of the paper is devoted to theoretical developments to demonstrate that the large sample spatial adaptivity properties of wavelet shrinkage carry over to a variety of correlated data settings.

TABLE 1
Comparison of various methods for the neurophysiology data†

Method	*Errors*	*%*
(a) Pointwise classification of raw data	1180.3	28.8
(b) Wavelet thresholding, thresholds $\hat{\sigma}_j\sqrt{(2\log n)}$ at levels $j \geqslant 6$	643.1	15.7
(c) As (b), but using translation invariant wavelet transform	350.7	8.6
(d) As (c), but using SURE thresholds at levels $j = 6, 7, 8$	95.9	2.3
(e) Special purpose algorithm provided by Eisenberg and Levis	82.2	2.0

†The performances shown are the averages over 10 successive segments each of length 4096. Method (a) estimates θ_i to be 0 if $Y_i < 0.5$ and 1 otherwise.

We begin with an outline and discussion of the agenda. In this section, we present an oracle inequality bounding the risk of thresholding from above in terms of ideal risk: this is a simple finite sample result and makes no *a priori* assumptions about the unknown signal.

The remaining sections discuss asymptotic results. First, in Section 5, we describe a lower bound result, still with no *a priori* assumptions on the signal, asserting that the logarithmic risk inflation factor in the oracle inequality cannot be improved asymptotically.

Secondly, in Section 7, we show how the oracle inequality, now in conjunction with various smoothness assumptions on the signal, leads easily to adaptive near minimaxity results. To prepare for this, we describe in Section 6 a family of models encompassing short- and long-range dependence.

For the upper bound results of this section and Section 7, the discussion will focus on the 'universal' threshold (8), since the presentation and proofs are simplest here. However, similar conclusions will apply to the thresholds based on the unbiased risk criterion and to the translation invariant versions also introduced in Section 2. To make this point for the SURE thresholds a theorem is stated without proof in Section 7 that is in certain respects stronger than the result that we prove in detail. As for the translation invariant version, it is easy to see from Jensen's inequality that the averaging used in the translation invariant construction can only improve the resulting rates of convergence, but again we do not go into details.

Taken together, the results that demonstrate that wavelet shrinkage estimators based on thresholding, broadly interpreted, have desirable, provable, spatial adaptivity properties that are not shared by many other methods, linear methods in particular. At the same time, we do not, of course, assert that wavelet methods are uniquely optimal, nor indeed that any one variant is exactly optimal in a specific situation with additional relevant side information.

4.2. *Ideal Risks*

We begin with a general multivariate normal model. Suppose that observations $X_i, i = 1, \ldots, n$, satisfy

$$X_i = \theta_i + z_i$$

where the z_i have a multivariate normal distribution with mean 0 and covariance matrix V. Define $\sigma_i^2 = V_{ii} = \mathrm{var}(X_i)$ and

$$\overline{\sigma^2} = n^{-1} \sum_1^n \sigma_i^2 = n^{-1} \mathrm{tr}(V). \tag{13}$$

We shall assume that the variances σ_i^2 are known.

Suppose that it is of interest to estimate the vector of coefficients θ_i. A problem of this kind is obtained from the function estimation problem by applying the discrete wavelet transform; the X_i are then the empirical wavelet coefficients w_{jk}.

It is of particular interest to consider situations in which the θ-vector contains only a few elements that are substantially different from 0. This is because a wide variety of 'regular' functions have wavelet transforms of this kind, not only those that are smooth in conventional senses (e.g. having small integrated squared mth derivative)

but also those that are smooth between possible discontinuities, and those that have inhomogeneous smoothness properties; see Donoho and Johnstone (1994) for further details in the case of IID errors. With this in mind, we could consider the effect of having an *oracle* that told us which of the θ_i were near 0 and consider the best estimation rules that could be obtained in the presence of this additional information. Of course such rules cannot be used in practice, but we might hope that they could be mimicked well by suitable spatially adaptive estimators.

We shall in fact consider oracles among all *diagonal projections* of the form $\hat{\theta}_i = \delta_i X_i$, $\delta_i = 0$ or $\delta_i = 1$, so that the additional information provided by the oracle would tell us whether to 'keep' or 'kill' co-ordinate i. We let $R(\mathrm{DP}, \theta)$ be the risk that would be obtained for the ideal choice of sequence $\{\delta_i\}$. Then

$$\begin{aligned} R(\mathrm{DP}, \theta) &= \min_{\{\delta_i\}} \left\{ \sum_i E(\delta_i X_i - \theta_i)^2 \right\} = \sum_i \min\left\{\theta_i^2, E(X_i - \theta_i)^2\right\} \\ &= \sum_i (\theta_i^2 \wedge \sigma_i^2) \end{aligned} \tag{14}$$

where $\delta_i = I[\theta_i^2 \geqslant E(X_i - \theta_i)^2] = I[\theta_i^2 \geqslant \sigma_i^2]$ and so the 'ideal' diagonal projection would be obtained by setting

$$\delta_i = I[|\theta_i| > \sigma_i]. \tag{15}$$

If the oracle were able to tell us which θ_i were numerically larger than their corresponding noise standard deviation σ_i we could use the values given in equation (15) and attain the risk

$$R(\mathrm{DP}, \theta) = \sum_i (\theta_i^2 \wedge \sigma_i^2). \tag{16}$$

Of course, we cannot actually use this diagonal projection in practice, but the risk (16) can be used to construct a reference bench-mark against which to judge the behaviour of estimates that can be realized from observed data. For technical reasons, the bench-mark risk will be $\overline{\sigma^2} + \Sigma_i(\theta_i^2 \wedge \sigma_i^2)$; the additional $\overline{\sigma^2}$ is the average mean-square error in estimating a single parameter unbiasedly, and for all but the sparsest signals will be small compared with $R(\mathrm{DP}, \theta)$.

We set the threshold to $\lambda_n \sigma_i$, where $\lambda_n = \sqrt{(2 \log n)}$, and define

$$\hat{\theta}_i = \eta_S(X_i, \lambda_n \sigma_i). \tag{17}$$

First consider X_i as a scalar observation from an $N(\theta_i, \sigma_i^2)$ density. Donoho and Johnstone (1994) showed that

$$E(\hat{\theta}_i - \theta_i)^2 \leqslant (1 + 2 \log n)(n^{-1}\sigma_i^2 + \theta_i^2 \wedge \sigma_i^2).$$

Even though the components X_i are correlated, we may simply sum over co-ordinates to obtain the following upper bound for the risk of this estimator compared with the unattainable bench-mark risk, as follows.

Theorem 1. Suppose that $X \sim N_n(\theta, V)$ with $\sigma_i^2 = V_{ii}$ and define the soft threshold estimate $\hat{\theta}$ by equation (17). Then

$$E\|\hat{\theta} - \theta\|^2 \leqslant (1 + 2 \log n)\left\{\overline{\sigma^2} + \sum_i (\theta_i^2 \wedge \sigma_i^2)\right\} \qquad \text{for all } \theta \in \mathbb{R}^n. \tag{18}$$

Result (18) shows that, for all possible θ, the estimator $\hat{\theta}$ comes within a factor $1 + 2 \log n$ of achieving the bench-mark risk $\overline{\sigma^2} + R(\mathrm{DP}, \theta)$. The result can be used to show that wavelet threshold estimators can yield nearly optimal estimators over wide ranges of function classes for the unknown function of interest. Further details are given in Section 7 below.

Although the bench-mark risk has been derived by reference to an ideal diagonal projection, it can equally be considered in its own right as a measure of the compressibility of the signal θ_i relative to the given noise process. For example, suppose that $V = \sigma^2 I$, and that the signal f is a piecewise polynomial with a fixed number of break points, the polynomial pieces being of suitably low degree relative to the wavelet transform being used; for details see Donoho and Johnstone (1994). Then the number of non-zero elements in the wavelet transform θ of f will be of the order $\log n$ as $n \to \infty$ and the bench-mark risk therefore of the order $\sigma^2 \log n$. The average squared error of the function estimate $\hat{f}$ will satisfy

$$n^{-1} \sum_{i=1}^{n} \{\hat{f}(t_i) - f(t_i)\}^2 = n^{-1} \|\hat{\theta} - \theta\|^2.$$

Therefore results (18) and (19) will show that, measuring risk in an averaged mean-square error sense, the wavelet threshold estimator will have risk $O(n^{-1} \log^2 n)$, and that this risk is within a factor $\log n$ of that obtainable by an 'ideal' wavelet diagonal projection estimator. The risk compares extremely favourably with the corresponding rate $O(n^{-1/2})$ for non-adaptive linear methods.

For the white noise case $V = \sigma^2 I$, theorem 2 of Donoho and Johnstone (1994) shows that the behaviour indicated in inequality (18) cannot essentially be improved. They showed that

$$\frac{1}{2 \log n} \inf_{\tilde{\theta}} \sup_{\theta \in \mathbb{R}^n} \left\{ \frac{E\|\tilde{\theta} - \theta\|^2}{\sigma^2 + \sum_{i=1}^{n} (\theta_i^2 \wedge \sigma^2)} \right\} \to 1 \tag{19}$$

as $n \to \infty$, where for each n the infimum is taken over *all* estimators $\tilde{\theta}$ that depend on $X_1, \ldots, X_n$, not merely threshold estimators. The threshold estimator $\hat{\theta}$ thus asymptotically attains the best possible behaviour of any estimator relative to the bench-mark risk. We give an extension of this result to the case of general Gaussian noise in Section 5. A different extension, in the direction of IID non-Gaussian errors in the wavelet domain, has been recently developed by Averkamp and Houdré (1996).

5. LOWER BOUND ON ESTIMATION RISK

In this section we set out a theorem that demonstrates that, under very mild conditions, the performance attained by $\hat{\theta}$ in theorem 1 is in a minimax sense the best possible up to a constant: the logarithmic rate up to which $\hat{\theta}$ mimics the oracle cannot

be improved. When applied to the wavelet context, the conditions of the theorem will encompass both short-range and long-range dependent noise processes.

5.1. *Main Theorem*

We first define additional notation. Assume that V is invertible, and write σ^{ij} for the ij-element of V^{-1}. Define $\tau_i^2 = 1/\sigma^{ii}$. Then by standard multivariate normal theory

$$\tau_i^2 = \mathrm{var}(X_i | X_j, j \neq i)$$

and so $\tau_i^2 \leqslant \sigma_i^2$ for all i, with equality when the X_i are independent. Set

$$\overline{\tau^2} = n^{-1} \sum_1^n \tau_i^2. \tag{20}$$

Our result will be proved in the framework of a sequence of problems with increasing n. The covariance matrix of the noise at sample size n will be written V_n, but the consequent dependence of the quantities σ_i^2, τ_i^2, $\overline{\sigma^2}$ and $\overline{\tau^2}$ will not be made explicit in the notation. The theorem will be proved under the assumption that there are constants β and C_1, with $1 < \beta \leqslant 2$, such that, for all n,

$$n^{-1} \sum_{i=1}^n \sigma_i^{2\beta} \Big/ \left(n^{-1} \sum_{i=1}^n \sigma_i^2 \right)^{\beta} \leqslant C_1 \tag{21}$$

and that, for some finite constant C_2,

$$\overline{\sigma^2}/\overline{\tau^2} \leqslant C_2. \tag{22}$$

We can consider assumption (21) as being a *reverse Hölder* condition and assumption (22) as an overall *bounded dependence* condition. These assumptions will be discussed further below, but we first state the main result.

Theorem 2. Assume that the n-vector of observations X has an $N(\theta, V_n)$ distribution and that V_n is such that assumptions (21) and (22) are satisfied. Let $\tilde{\Theta}_n$ be the set of all estimators of θ in $\mathbb{R}^n$ based on $X_1, \ldots, X_n$. Then

$$\liminf_{n \to \infty} \left(\frac{1}{2 \log n} \frac{\overline{\sigma^2}}{\overline{\tau^2}} \right) \inf_{\tilde{\theta} \in \tilde{\Theta}_n} \sup_{\theta \in \mathbb{R}^n} \left\{ \frac{E \|\tilde{\theta} - \theta\|^2}{\overline{\sigma^2} + \sum_{i=1}^n (\theta_i^2 \wedge \sigma_i^2)} \right\} \geqslant 1. \tag{23}$$

We remark that the quantity $\overline{\sigma^2}/\overline{\tau^2}$ is a measure of the correlations in V_n. The infimum is taken over *all* estimators $\tilde{\theta}$, and the theorem illustrates that, provided $\overline{\sigma^2}/\overline{\tau^2}$ remains bounded, at most a constant multiple of loss is incurred by restricting attention to the apparently very restrictive class of pointwise threshold estimators with thresholds $\sigma_i \sqrt{(2 \log n)}$.

As in Donoho and Johnstone (1994), the key idea of the proof is to bound the minimax risk in inequality (23) by the Bayes risk relative to a certain prior on θ. However, the details of the argument are different in several respects because of the more general set-up. The proof is given in Appendix A. We conjecture that it may be

possible to prove the theorem under even milder conditions than those which we have required, but as yet we have been unable to do so.

In the context of wavelet threshold estimators, the theorem will be applied to a vector $X = (w_{jk})$ that has been obtained by wavelet transformation of the original observation vector Y, and the matrix V_n will be the orthogonal transform (6) of the covariance matrix Γ_n of the original observations. In considering the validity of the conditions (21) and (22) under short-range dependence, only the orthogonality of this transformation will be used. For long-range dependence, we shall appeal to the detailed properties of the wavelet transform of the noise process.

5.2. *Reverse Hölder condition (21)*

In the short-range dependence case, assume that Γ_n is a circulant matrix. Then

$$\overline{\sigma^2} = n^{-1} \operatorname{tr}(V_n) = n^{-1} \operatorname{tr}(\Gamma_n) = r_0^{(n)} = \operatorname{var}(Y_i). \tag{24}$$

Since the average of any convex function of the diagonal elements of a symmetric matrix is dominated by the average of the function of the eigenvalues, we have

$$n^{-1} \sum_{i=1}^{n} \sigma_i^4 \leqslant n^{-1} \operatorname{tr}(V_n^2) = n^{-1} \operatorname{tr}(\Gamma_n^2) = \sum_{j=0}^{n-1} (r_j^{(n)})^2.$$

Thus condition (21), with $\beta = 2$, will be satisfied if the sum of the squares of the autocorrelations of the original noise sequence,

$$(r_0^{(n)})^{-2} \sum_{j=0}^{n-1} (r_j^{(n)})^2,$$

remains bounded as $n \to \infty$, a natural and very mild condition.

Turning to the long-range dependence case as discussed in Section 6.3, choose any β such that $1 < \beta < \min(\alpha^{-1}, 2)$. We then have

$$2^{-J} \sum \sigma_i^{2\beta} = 2^{-J} \sum_{j=1}^{J} 2^j \times 2^{-\alpha\beta j} \asymp 2^{-\alpha\beta J}$$

as $n \to \infty$, and

$$\overline{\sigma^2} = 2^{-J} \sum_{j=1}^{J} 2^j \times 2^{-\alpha j} \asymp 2^{-\alpha J}.$$

Thus it follows at once that condition (21) is satisfied.

5.3. *Ratio $\overline{\sigma^2}/\overline{\tau^2}$*

For the long-range dependence case, condition (35) below states that, in the wavelet transform, each individual $\sigma_i^2/\tau_i^2 \leqslant c_0^{-1}$, so the quantity $\overline{\sigma^2}/\overline{\tau^2}$ is also bounded by c_0^{-1}.

For the short-range case, since the arithmetic mean of the quantities $(\sigma^{ii})^{-1}$ is necessarily greater than their harmonic mean, we have

$$\overline{\tau^2} = n^{-1} \sum (\sigma^{ii})^{-1} \geqslant \left(n^{-1} \sum \sigma^{ii} \right)^{-1} = \{ n^{-1} \operatorname{tr}(V_n^{-1}) \}^{-1} = \{ n^{-1} \operatorname{tr}(\Gamma_n^{-1}) \}^{-1}$$

since V_n is an orthogonal transform of Γ_n. Since Γ_n^{-1} is circulant, each element $\operatorname{var}(Y_i | Y_j,\ j \neq i)$ of its diagonal will be equal to $\{ n^{-1} \operatorname{tr}(\Gamma_n^{-1}) \}^{-1}$. Therefore, using expression (24), we shall have

$$\overline{\sigma^2} / \overline{\tau^2} \leqslant \operatorname{var}(Y_i) / \operatorname{var}(Y_i | Y_j,\ j \neq i) = \kappa_n, \tag{25}$$

say, the ratio of the variance of the original process to the residual variance of any particular observation about its (linear) predictor based on all the other observations.

The boundedness of κ_n as $n \to \infty$ is a very mild condition for the short-range dependence case. In practice the ratio $\overline{\sigma^2}/\overline{\tau^2}$ may be substantially closer to 1 than κ_n is, because the wavelet transform will often reduce the correlation, and so the factor κ_n may be pessimistic. In particular, if the discrete wavelets are approximately eigenvectors of the covariance matrix Γ_n then $\overline{\sigma^2}/\overline{\tau^2}$ will be close to 1.

6. MODELS FOR SHORT- AND LONG-RANGE DEPENDENCE

6.1. *Basic Framework*

Let $\{e_t,\ t \in \mathbb{Z}\}$ be a stationary Gaussian process with covariance sequence $r_k = \operatorname{cov}(e_t, e_{t+k})$. In this section we indicate ways in which our n-sample regression model with stationary errors,

$$Y_j = f(j/n) + e_j \qquad j = 1, \ldots, n, \tag{26}$$

can be approximated in large samples by one of a simple set of models of self-similar form that are sufficiently flexible to accommodate both short- and long-range dependence. These approximations elucidate the heuristic remarks made in Section 2.2 earlier.

The framework that we use is one in which the sampling rate n on the interval $[0, 1]$ is large. For $t \in [0, 1]$ we construct the observation process

$$Y_n(t) = n^{-1} \sum_{j=1}^{[nt]} Y_j = n^{-1} \sum_{j=1}^{[nt]} f(j/n) + n^{-1} \sum_{j=1}^{[nt]} e_j.$$

We define $F_n(t)$ to be the cumulative 'signal' process $n^{-1} \Sigma_{j=1}^{[nt]} f(j/n)$. Our approximations concern the difference between the processes $Y_n(t)$ and $F_n(t)$. The function f may depend on n, but any such dependence will not be made explicit in the notation. Although a more general theory would be possible, we isolate two cases of interest.

6.2. *Short-range Dependence*

Suppose that $\Sigma_{-\infty}^{\infty} |r_k| < \infty$. Set $\tau^2 = \Sigma_{-\infty}^{\infty} r_k$. In this case, the usual $\sqrt{n}$-normalization of the error partial sum process $\Sigma_1^{[nt]} e_j$ converges weakly to (scaled) standard Brownian motion $B(t)$ on $[0, 1]$, so that

$$n^{1/2}\{Y_n(t) - F_n(t)\} \Rightarrow \tau B(t) \qquad t \in [0, 1]. \tag{27}$$

This follows, for example, from lemma 5.1 of Taqqu (1975).

Make the calibration $\epsilon = \tau n^{-1/2}$. Then, in the sense made precise by expression (27), we can approximate the observation process $Y_n(t)$ by $Y(t)$ for $t \in [0, 1]$, where

$$Y(t) = F(t) + \epsilon B(t), \tag{28}$$

and

$$F(t) = \int_0^t f(s)\,\mathrm{d}s.$$

Let $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$ be a wavelet basis on $\mathbb{R}$ derived from a suitable wavelet ψ of compact support (with corresponding scaling function ϕ). Let $\{\tilde{\psi}_\lambda\}$ be the corresponding wavelet basis for [0, 1], e.g. as constructed by Cohen *et al.* (1992). Here the index λ runs over a set Λ defined by pairs (j, k), $j \geqslant j_0$, $k = 1, \ldots, 2^j$, for the wavelet functions and $(j_0 - 1, k), k = 1, \ldots, 2^{j_0}$, for the scaling functions.

By forming the inner products

$$y_\lambda = \int \tilde{\psi}_\lambda\,\mathrm{d}Y,$$

$$\theta_\lambda = \int \tilde{\psi}_\lambda f$$

and

$$z_\lambda^0 = \int \tilde{\psi}_\lambda\,\mathrm{d}B$$

we obtain an equivalent sequence space form of model (28):

$$y_\lambda = \theta_\lambda + \epsilon z_\lambda^0. \tag{29}$$

Because of edge effects near 0 and 1, the zero-mean Gaussian variables z_λ^0 associated with a given level j have a covariance matrix Γ_j which is not exactly spherical, satisfying instead

$$\sigma_a^2 I \leqslant \Gamma_j \leqslant \sigma_b^2 I \tag{30}$$

where the inequalities are in the sense of non-negative definite matrices and, crucially, the constants σ_a^2 and σ_b^2 *do not depend on j*, but only on the wavelet basis chosen. Thus, up to the absolute degree of approximation in the variance structure implied by inequality (30), we may replace the sequence space model (29) by

$$y_\lambda = \theta_\lambda + \epsilon z_\lambda, \qquad z_\lambda \overset{\text{IID}}{\sim} N(0, 1), \quad \lambda \in \Lambda. \tag{31}$$

6.3. *Long-range Dependence*

Now let us suppose that the autocovariance function of the errors $\{e_t\}$ decays slowly according to the model $r_k \sim Ak^{-\alpha}$, for $0 < \alpha < 1$; compare Beran (1994). Set

$$\tau^2 = 2A/(1-\alpha)(2-\alpha)$$

and $H = 1 - \alpha/2 \in (\frac{1}{2}, 1)$.

Define *fractional Brownian motion* $B_H(t)$ to be a zero-mean Gaussian process on $\mathbb{R}$ with covariance function $r(s, t)$ given by

$$r(s,t) = \frac{V_H}{2}(|s|^{2H} + |t|^{2H} - |t-s|^{2H}), \qquad s, t \in \mathbb{R},$$

where

$$V_H = \mathrm{var}\{B_H(1)\} = \frac{-\Gamma(2-2H)\cos(\pi H)}{\pi H(2H-1)}.$$

Lemma 5.1 of Taqqu (1975) now shows that the error partial sum process, normalized by n^H, converges weakly on the interval [0, 1] to $\tau B_H(t)$, so that

$$n^{\alpha/2}(Y_n - F_n)(t) \Rightarrow \tau B_H(t), \qquad t \in [0,1]. \tag{32}$$

Thus, setting $\epsilon = \tau^{1/\alpha} n^{-1/2}$, we can now approximate the observation process $Y_n(t)$ by $Y(t)$, where, for t in [0, 1],

$$Y(t) = F(t) + \epsilon^{\alpha} B_H(t). \tag{33}$$

As in the short-range case, we may form an equivalent sequence space model in terms of wavelet coefficients. To avoid annoying but inconsequential end effects, we may argue as in the short-range case that a nearly equivalent model (i.e. involving an approximation of the variance structure that is valid up to absolute multiplicative constants) may be obtained as

$$y_\lambda = \theta_\lambda + \epsilon^{\alpha}\gamma_j z_\lambda, \qquad \lambda \in \Lambda, \tag{34}$$

where

$$\gamma_j^2 = \mathrm{var}\left\{ \int \psi_\lambda \, \mathrm{d}B_H \right\} = 2^{-j(1-\alpha)}\tau^2$$

and

$$z_\lambda = \gamma_j^{-1} \int \psi_\lambda \, \mathrm{d}B_H.$$

The noise variables z_λ, which all have variance 1, are not uncorrelated but can be shown to be of 'bounded dependence' in the sense that, for all λ,

$$0 < c_0 \leqslant \mathrm{var}(z_\lambda | z_\mu, \mu \neq \lambda) \leqslant 1. \tag{35}$$

We can think of the initial segments $\{y_{jk}: j < J = \log_2 n, k = 1, \ldots, 2^j\}$ in model (34) with $\epsilon = \tau^{1/\alpha} n^{-1/2}$ as being analogous to the empirical coefficients w_{jk} in equation (2). Although this is not literally correct, we can use this identification to transfer intuition from the asymptotic models to empirical data.

It is simpler to do rates of convergence calculations in the approximating models (31) and (34). By some general decision theoretic and wavelet theoretic machinery

(see Donoho (1992), for example) we expect that these results can be carried over to the original regression model (26), but for reasons of space and complexity we omit the details. Instead we use the approximating models to indicate how the 'oracle inequality' theorem 1 yields very simply that our $\sigma_j\sqrt{(2 \log n)}$ thresholds provide the optimal rates of convergence (up to a logarithmic factor) in both short- and long-range dependence settings.

7. SMOOTHNESS, DEPENDENCE AND RATES OF CONVERGENCE

7.1. *Sequence Space Noise and Function Models*

In this section we shall consider the sequence space model

$$y_\lambda = \theta_\lambda + \epsilon^\alpha \gamma_j z_\lambda, \qquad \lambda \in \Lambda, \tag{36}$$

for $0 < \alpha \leqslant 1$. The z_λ will be $N(0, 1)$ random variables, not necessarily independent. The index set Λ is the set of possible $\lambda = (j, k)$. It will be assumed that the parameters α and τ are known—since the latter is a simple scale parameter, we shall set $\tau = 1$ without further loss of generality. We shall therefore have $\epsilon = n^{-1/2}$, $\gamma_j^2 = 2^{-j(1-\alpha)}$ and $\sigma_j^2 = \epsilon^{2\alpha}\gamma_j^2$. This model encompasses both the long-range dependence approximation (34) and, by setting $\alpha = 1$, the short-range dependence approximation (31).

We shall consider results for a broad range of function classes for the regression function f, corresponding to sequence space models for its coefficients θ_λ. A flexible scale of functional classes is given by the Besov family, which is specified in sequence space form as follows. Set $\|\theta_j\|_p^p = \Sigma_{k=1}^{2^j} |\theta_{jk}|^p$ and

$$b_{p,q}^\sigma(C) = \{(\theta_{jk}): \sum_{j=0}^{\infty} 2^{jsq} \|\theta_j\|_p^q \leqslant C^q\}, \qquad s = \sigma + \tfrac{1}{2} - 1/p.$$

For a fuller discussion of these spaces and the important roles of the indices (σ, p, q) see Frazier *et al.* (1991) and Donoho and Johnstone (1997). Here we note simply that σ is a smoothness parameter, corresponding to the number of derivatives that the function f has in L_p. The case $p = q = \infty$ corresponds to Hölder smoothness, defined by the uniform condition $|D^m f(x) - D^m f(y)| \leqslant C_0 |x - y|^\delta$, where $\sigma = m + \delta$ with $\delta \in (0, 1]$.

7.2. *Simple Adaptivity Result*

We now set $n = \epsilon^{-2} = 2^J$ and consider level-dependent threshold estimators of the form

$$\hat{\theta}_\lambda = \begin{cases} \eta_S\{y_\lambda, \sigma_j\sqrt{(2 \log n)}\} & j < J, \\ 0 & j \geqslant J. \end{cases} \tag{37}$$

Theorem 1 can be applied to $\hat{\theta}$ to describe its worst case convergence properties over a large variety of function classes. Indeed, the following theorem shows that $\hat{\theta}$ has a certain near optimality property simultaneously over all the classes. In practice, we are unlikely to know exactly which function class is the most appropriate description of *a priori* knowledge, and so it is precisely this robustness of near optimality that is a useful property for an estimator to have. The proof of the theorem is given in

Appendix A.3 later. It depends on elementary manipulations of the sequence space characterization of both the 'signal' and 'noise' processes.

Theorem 3. Let $0 < \alpha \leqslant 1$, and define $r(\sigma, \alpha) = 2\sigma\alpha/(2\sigma + \alpha)$. Set $n = 2^J$. Suppose that, for each index λ in Λ,

$$y_\lambda \sim N(\theta_\lambda, 2^{-j(1-\alpha)}n^{-\alpha}), \tag{38}$$

not necessarily independently. If $0 < p, q \leqslant \infty$ and $\sigma > 1/p$, under model (38),

$$\sup_{\theta \in b^\sigma_{p,q}(C)} (E\|\hat{\theta} - \theta\|^2) \leqslant C_0 \cdot C^{2(1-r/\alpha)} \cdot \log n \cdot n^{-r(\sigma,\alpha)}.$$

Remark 1. For $\alpha < 1$, the rate $n^{-r(\sigma,\alpha)}$ is the minimax rate of convergence for *all* estimates under the long-range dependence model (38) over the Besov smoothness spaces $b^\sigma_{p,q}(C)$; see Wang (1996). Thus, the simple level-dependent thresholding strategy mimics this rate to within a logarithmic term without having to know the parameters σ, p, q and C of the function space.

Remark 2. An important reason for considering the full set of Besov classes $b^\sigma_{p,q}$ is to demonstrate the advantages of even simple non-linear thresholding over linear methods. Indeed, Wang (1996) showed that linear estimators, such as the kernel methods studied by Hall and Hart (1990) and Csörgö and Mielniczuk (1993), cannot attain the optimal rates of convergence over Besov classes with $p < 2$. In contrast, theorem 3 holds for all $p \in (0, \infty]$. This is the same as occurs already with IID errors: for that setting it is discussed in greater detail in Donoho and Johnstone (1997).

Remark 3. In the special case of short-range dependence, $\alpha = 1$, the mean-squared error bound yields the familiar rate: $C_0 \log n \cdot n^{-2\sigma/(2\sigma+1)}$. Thus wavelet thresholding comes within a logarithmic term of the minimax rate quite generally, and regardless of whether the model involves short- or long-range dependence.

Remark 4. The logarithmic term in the mean-squared error bound comes from the high threshold choices used (involving $\sqrt{(2 \log n)}$. Lower choices of threshold can be specified to obtain the exactly optimal rate for given (σ, p, q, C), as is done implicitly by Wang (1996). Of course, these choices will depend on (σ, p, q, C). Donoho and Johnstone (1995) showed that *data-based* threshold choices based on unbiased estimates of risk could yield exactly correct rates of convergence in the white noise model and an extension of this result to our correlated data settings is stated below.

7.3. *Estimation of* α

An obvious defect of estimate (37) is that it depends on unknown parameters—the degree of dependence α and a scale parameter. We indicate briefly here that it is possible to estimate these from the data without losing the near minimaxity properties. For this, we use model (34), but with a parameterization in which $\gamma_j^2 = 2^{-(j+1)(1-\alpha)}\tau^2$.

There are many references on the estimation of α (see for example Beran (1994)), but we shall for convenience use a crude estimator based only on the top two levels of the wavelet decomposition. Write $\sigma_j^2 = \mathrm{var}(y_{jk}) = \epsilon^{2\alpha}\gamma_j^2$, and set

$$\hat{\alpha} = (1 - \log(\hat{\sigma}^2_{J-2}/\hat{\sigma}^2_{J-1}))_+.$$

A variety of scale estimates $\hat{\sigma}_j^2$ would be possible, including the median absolute deviation estimators considered in Section 3.2. For definiteness, and to obtain a simpler proof, we use here $\hat{\sigma}_j^2 = 2^{-j} \Sigma_1^{2^j} y_{jk}^2$, the bias of this estimator being negligible at high levels j because of the smoothness constraints on θ. A simple estimate of the scale parameter τ is then $\hat{\tau} = \hat{\sigma}_{J-1}\sqrt{n}$.

Let $\tilde{\theta}$ be the estimator obtained from estimate (37) by substituting $\hat{\alpha}$ and $\hat{\gamma}_j = 2^{-(j+1)(1-\hat{\alpha})/2}\hat{\tau}$. We emphasize that the specification of $\tilde{\theta}$ no longer depends on any unknown parameters.

Theorem 4. Let model (34) hold, with the parameterization given above. Suppose that $0 < \alpha \leqslant 1$, $0 < \tau < \infty$, $0 < p, q \leqslant \infty$, $\sigma > 1/p$. Then

$$\sup_{\theta \in b_{p,q}^{\sigma}(C)} (E\|\tilde{\theta} - \theta\|^2) \leqslant C_1 \cdot C^{2(1-r/\alpha)} \cdot \log n \cdot n^{-r(\sigma,\alpha)}.$$

Hence the wavelet thresholding estimator $\tilde{\theta}$ is nearly asymptotically minimax regardless of the parameters (σ, p, q, C) describing the function class, and the parameters (τ, α) governing the observations. This can be seen as a broad robustness property of wavelet methods which derives ultimately from the spatial and frequency localization properties of the wavelet basis itself. The proof, which is omitted, builds on the arguments for theorem 2 and properties of stochastic thresholds.

7.4. *SURE Gets the Thresholds Right*

A stronger adaptivity result is possible for an estimator using thresholds chosen via the unbiased risk method of Section 2.3.2. We shall merely state the result here, omitting the lengthy proof (Johnstone and Silverman, 1996).

In the physiology example, at resolution levels where noise greatly dominated signal, the thresholds were set at the high 'universal' level. As in Donoho and Johnstone (1995), we formalize this (admittedly crudely) in terms of a *pretest* for the absence of significant signal. Suppose that $X \sim N_d(\theta, V)$ with variances $\sigma_{kk} = 1$. Then the pretest compares an unbiased estimate of $\|\theta\|^2$, namely $s_d^2 = d^{-1} \Sigma_1^d x_k^2 - 1$, with a threshold γ_d:

$$\tilde{t}(x) = \begin{cases} \sqrt{(2 \log d)} & s_d^2 \leqslant \gamma_d, \\ \hat{t}(x) & s_d^2 > \gamma_d. \end{cases}$$

Thus the unbiased risk choice $\hat{t}$ of equation (12) is chosen only when the pretest rejects.

We consider the sequence model (34), and soft threshold estimators of the form

$$\hat{\theta}_\lambda^* = \eta_S(y_\lambda, \sigma_j \lambda_j)$$

$$\lambda_j = \begin{cases} 0 & j \leqslant L, \\ \tilde{t}(y_j/\sigma_j) & j \geqslant L. \end{cases}$$

If the parameters $(\sigma, p, q, C, \alpha)$ were all known, then the best possible estimation error of any threshold choice over the class $b_{p,q}^{\sigma}(C)$ is given by the minimax threshold risk

$$R^*_{T,\alpha}\{\epsilon; b^\sigma_{p,q}(C)\} = \inf_{(t_j)} \sup_{b^\sigma_{p,q}(C)} (E\|\hat{\theta}_{(t_j)} - \theta\|^2),$$

where $\hat{\theta}_{(t_j)}$ stands for the estimator $(\eta(y_{j,k}, t_j))_{jk}$. From results of Wang (1996), it is known that the minimax threshold risk in model (34) is of the same order in ϵ as the minimax risk over *all* estimators, i.e. there is no great loss of efficiency due to coordinatewise thresholding, and, over $b^\sigma_{p,q}(C)$, we have $R_{T,\alpha}(n^{-1/2}) \asymp n^{-r(\sigma,\alpha)}$.

Against this background, we have the following result for the estimator $\hat{\theta}^*$ using the choice of thresholds based on the unbiased risk criterion.

Theorem 5. Let $\eta, \gamma > 0$ be small and prespecified, and let the pretest threshold $\gamma_d = d^{-\gamma}$. Then, for $(p, q) \in [1, \infty)$, $C \in (0, \infty)$ and $\sigma > \sigma_0(\alpha, \eta, \gamma)$, as $n = \epsilon^{-1/2} \to \infty$,

$$\sup_{\theta \in b^\sigma_{p,q}(C)} (E\|\hat{\theta}^* - \theta\|^2) \leqslant R^*_{T,\alpha}\{n^{-1/2}; b^\sigma_{p,q}(C)\}\{1 + o(1)\}.$$

This theorem says that the unbiased risk choice gets the thresholds right asymptotically: without needing to know (σ, p, q, C), and over a wide range of α, the estimator does as well as if these parameters were known and used to set optimal thresholds explicitly. Note especially that the extra logarithmic term present in theorem 3 has been removed, owing to the lower thresholds chosen by the data-based rule.

8. SOME FURTHER REMARKS

Before setting out some of the detailed proofs in Appendix A, we make a few general remarks. Some comments about our perception of the value of the theoretical results may be in order. One may certainly be sceptical about the specific applicability of particular dependence models, and of results showing good adaptivity only in an asymptotic context. We view these models and asymptotic contexts as an important series of 'tests' against which to judge a given methodology. Some of these tests may be relatively artificial, but if the estimator 'passes' them then we may feel that we have gained a greater understanding of the applicability and potential of the methodology. The neurophysiological example certainly demonstrates that there is scope for further development, improvement and modification in practice. However, we are encouraged by the way that wavelet methods easily cope with the whole problem of adaptive estimation in correlated noise, especially under a wide range of possible forms of correlation.

We briefly comment on the case where the original data are stationary, but the covariance is not periodic. The remarks made in Section 4.6 of Donoho and Johnstone (1994) carry over to this case; the boundaries are dealt with by an appropriate preconditioning transformation of the data, affecting only a small number of data points near the boundaries, followed by a boundary-corrected version of the discrete wavelet transform. The effect on the theorems of the paper is only to introduce additional constants that do not depend on n, and the overall rate conclusions are unaffected.

A possible alternative procedure when dealing with correlated data (with known correlation structure) is to use a prewhitening transformation, which here might be followed by a wavelet transformation, thresholding (using the fixed $\sqrt{(2 \log n)}$ threshold) and backtransformation. This method would have the advantage that

wavelet thresholding is applied to a version of the data with homoscedastic uncorrelated noise. However, the wavelet decomposition of the signal in the original domain may have sparsity properties that are lost in the prewhitening transformation, and the advantages of using wavelet shrinkage on the prewhitened data would then be diminished. An overall comparison with the approach of this paper is an interesting topic for future research.

The results of this paper can be carried over to certain inverse problem settings. Suppose that A is an $n \times n$ invertible (but possibly ill-conditioned) matrix. Data $\tilde{X} \sim N(A\theta, \tilde{V})$ may be converted via the transformation $X = A^{-1}\tilde{X}$ to the form $X \sim N(\theta, V)$ considered in Sections 4 and 5 (with $V = A^{-1}\tilde{V}A^{-\mathrm{T}}$): note that thresholding is performed on the components of the inverted data X rather than in the domain of the observations $\tilde{X}$. If $\tilde{V}$ is known (for example if $A\theta$ is observed subject to white noise of known variance) then the variances of the individual components of X are also known, and so each component can be thresholded at a level that is proportional to its own standard deviation. The 'wavelet–vaguelette decomposition' of Donoho (1995) is essentially obtained by setting θ to be the wavelet transform of a function of interest in an inverse problem; X represents the coefficients of the data after expressing the operator in the wavelet basis for signals θ.

In this paper we have concentrated on estimation of a trend or signal in the presence of stationary correlated noise. This is thus a (nonparametric) first-moment setting. Other researchers (Gao, 1993; Moulin, 1994; Neumann, 1996) have considered the application of wavelet shrinkage to the second-moment setting of estimation of the spectral density of a stationary process of *zero mean*. These papers and ours may therefore be seen as addressing complementary problems.

There has been much recent interest in the estimation of 'spectra' of non-stationary processes that are somehow 'slowly varying' or 'locally stationary'; see, for example, Dahlhaus (1997), von Sachs and Schneider (1996), Adak (1995), Neumann and von Sachs (1997) and Mallat *et al.* (1997). Some approaches attempt an explicit segmentation of the process into intervals of approximate stationarity. Of course, for trend estimation in such situations, we could simply apply the methods of this paper on each such interval. More generally, an extension of the methods of this paper to near stationary settings would be of considerable interest.

ACKNOWLEDGEMENTS

This work was supported in part by grants from the Science and Engineering Research Council (now the Engineering and Physical Sciences Research Council), National Science Foundation (DMS 92-09130 and 95-05151) and National Institutes of Health (CA 59039-18 and 72028-01). We are indebted to Bob Eisenberg and Rick Levis for drawing our attention to the ion channel example and providing the test data. The work was facilitated by visits by BWS to Stanford and IMJ to Bristol. Helpful comments of Felix Abramovich, Guy Nason, John Rice, Rainer von Sachs and the reviewers are gratefully acknowledged.

APPENDIX A: PROOFS AND DETAILS

A.1. *Wavelet Transform of Stationary Processes*

We collect here, for the convenience of the reader, a few more heuristic details on the effect

of the wavelet transform on stationary processes. We claim no novelty for this material; for related results see, for example, Flandrin (1994), Houdré (1992) and Istas (1992).

The wavelet transform at level j, or scale 2^j, is given by convolution with a scaled version of the basic wavelet ψ: $W_j f = \psi_j * f$, where $\psi_j(t) = 2^{j/2}\psi(2^j t)$. In the Fourier domain, using circumflexes to denote Fourier transforms, the wavelet transform is then represented by multiplication: $\widehat{W_j f} = \widehat{\psi_j}\,\hat{f}$. The support of $\hat{\psi}$ is approximately concentrated in the frequency band $\pm[\pi, 2\pi]$, and so the support of $\widehat{\psi_j}$ is correspondingly located near $\pm[2^j\pi, 2^{j+1}\pi]$.

Now apply these considerations to the orthogonal increments representation of a (continuous time) stationary process

$$Y(t) = \int \exp(\mathrm{i}t\xi)\sqrt{g(\xi)}\,\mathrm{d}Z(\xi)$$

where $g(\xi)$ is the spectral density function of Y. The multiplier property shows that the jth-level wavelet transform of Y is given by

$$W_j(t) = \int \exp(\mathrm{i}t\xi)\,\widehat{\psi_j}(\xi)\sqrt{g(\xi)}\,\mathrm{d}Z(\xi).$$

We can now read off the covariance properties of the W_j-processes:

$$E\,W_j(t)\,\overline{W_{j'}(t)} = \int \exp\{\mathrm{i}\xi(t-t')\}\,\widehat{\psi_j}(\xi)\,\overline{\widehat{\psi_{j'}}(\xi)}\,g(\xi)\,\mathrm{d}\xi. \tag{39}$$

If the basic wavelet ψ generates an orthonormal basis, it may be shown that $\Sigma_j\,|\widehat{\psi_j}(\xi)|^2 \equiv 1$ for all ξ. Combined with the near disjoint support properties of $\widehat{\psi_j}$, this shows that the correlation between W_j and $W_{j'}$ decreases as $|j - j'|$ increases.

Within a given level j, equation (39) shows that $W_j(t)$ is stationary, roughly carrying the spectral information of Y in the frequency band $\pm[2^j\pi, 2^{j+1}\pi]$. If ψ has M vanishing moments, then $\hat{\psi}$ has M vanishing derivatives at $\xi = 0$, and this may be used to show a polynomial rate of decay (depending on $2M$) of the autocorrelations of $t \to W_j(t)$.

A.2. *Proof of Theorem 2*

As in Donoho and Johnstone (1994), the key idea of the proof is to bound the minimax risk in inequality (23) by the Bayes risk relative to a certain prior on θ. The details of the argument are different in several respects because of the more general set-up.

A.2.1. *Specifying a suitable prior*

We begin the proof by defining some additional notation. For any $\theta \in \mathbb{R}^n$ write

$$p(\theta) = 1 + (\overline{\sigma^2})^{-1} \sum_{i=1}^{n} (\theta_i^2 \wedge \sigma_i^2). \tag{40}$$

Define the modified loss function

$$\tilde{L}_n(\hat{\theta}, \theta) = (\overline{\sigma^2})^{-1}\,p(\theta)^{-1} \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2 \tag{41}$$

so that theorem 1 can be written

$$\sup_{\theta \in \mathbb{R}^n} \{E_\theta \tilde{L}_n(\hat{\theta}, \theta)\} \leqslant 2 \log n + 1.$$

Now we construct a prior $\pi_n(\theta)$ on θ and consider the Bayes risk with respect to this prior. As in Donoho and Johnstone (1994), choose $a \gg 0$. Define μ_n by

$$\phi(a + \mu_n) = \frac{\log n}{n} \phi(a).$$

Then $\mu_n \sim \sqrt{(2 \log n)}$ as $n \to 0$. Let $F[\epsilon, \mu]$ be the three-point distribution that places mass $\frac{1}{2}\epsilon$ on each of $\pm\mu$ and mass $1 - \epsilon$ on 0. The prior π_n is then defined by setting the components θ_i to be independent $F[n^{-1} \log n, \mu_n \tau_i]$.

Under this prior, we shall consider Bayes rules and risks relative both to the familiar squared error loss and also to the modified loss function $\tilde{L}_n$ as defined in equation (41). Write $\hat{\theta}^{\mathrm{b}}$ for the Bayes estimator $E(\theta|X)$ relative to squared error loss and $\tilde{\theta}^{\mathrm{b}}$ for the Bayes estimator relative to the loss function $\tilde{L}_n$. Let the respective Bayes risks be

$$\rho_n(\pi_n) = \sum_{i=1}^{n} E(\hat{\theta}_i^{\mathrm{b}} - \theta_i)^2$$

and

$$\tilde{\rho}_n(\pi_n) = E \tilde{L}_n(\tilde{\theta}_i^{\mathrm{b}}, \theta_i),$$

in both cases taking the expectations both over the prior and over the distribution of X.

A.2.2. *Some key lemmas*

The first lemma considers the Bayes risk relative to squared error loss.

Lemma 1. Suppose that θ has prior π_n, and that $X \sim N(\theta, V_n)$. The Bayes risk $\rho_n(\pi_n)$ of estimating θ from the observation X satisfies

$$\liminf_{n \to \infty} \left\{ \frac{\rho_n(\pi_n)}{\tau^2 \mu_n^2 \log n} \right\} \geqslant \Phi(a),$$

where Φ is the standard normal distribution function.

Proof. To prove the lemma, set

$$\xi_i = X_i + \tau_i^2 \sum_{j \neq i} \sigma^{ij}(X_j - \theta_j) = \theta_i + \left\{ V_n^{-1}(X - \theta) \right\}_i \Big/ \sigma^{ii}.$$

Since $X - \theta \sim N(0, V_n)$ independently of θ, we have, independently of θ,

$$V_n^{-1}(X - \theta) \sim N(0, V_n^{-1} V_n V_n^{-1}) = N(0, V_n^{-1})$$

and so, independently of θ_i,

$$\xi_i - \theta_i \sim N(0, \sigma^{ii})/\sigma^{ii} = N\{0, (\sigma^{ii})^{-1}\} = N(0, \tau_i^2).$$

Thus the joint distribution of (θ_i, ξ_i) is the same as that in a scalar problem where $\theta_i \sim F[n^{-1} \log n, \mu_n \tau_i]$ and $\xi_i \sim N(\theta_i, \tau_i^2)$. Given $\epsilon > 0$, Donoho and Johnstone (1994) showed that the Bayes risk $E\{\mathrm{var}(\theta_i|\xi_i)\}$ for this problem satisfies

$$E\{\mathrm{var}(\theta_i|\xi_i)\} \geqslant (1-\epsilon)\mu_n^2\tau_i^2 n^{-1}\Phi(a)\ \log\ n \tag{42}$$

for all sufficiently large n, independently of τ_i^2. (They actually showed that the Bayes risk in the scalar problem of estimating $\zeta \sim F[n^{-1}\ \log\ n, \mu_n]$ from a single observation $v \sim N(\zeta, 1)$ is asymptotic to $\mu_n^2 n^{-1}\ \Phi(a)\ \log\ n$ as $n \to \infty$.)

Consider the distribution of θ_i conditional on X and on θ_j for $j \neq i$. By straightforward manipulations, making use of the independence of the prior,

$$\log f(X, \theta) = -\tfrac{1}{2}(\theta_i - \xi_i)^2/\tau_i^2 + \log f(\theta_i) + \text{terms independent of } \theta_i$$

and so the distribution of θ_i conditional on $\{X, \theta_j \text{ for } j \neq i\}$ is the same as that conditional on ξ_i.

The Bayes risk then satisfies

$$\begin{aligned}
\rho_n(\pi_n) = \sum_i E(\hat{\theta}_i^{\mathrm{b}} - \theta_i)^2 \ &= \ \sum_i E\,\mathrm{var}(\theta_i|X) \\
&\geqslant \sum_i E\ \mathrm{var}(\theta_i|X, \theta_j \text{ for } j \neq i) \ = \ \sum_i E\,\mathrm{var}(\theta_i|\xi_i) \\
&\geqslant (1-\epsilon)\mu_n^2\,\overline{\tau^2}\ \Phi(a)\ \log\ n \qquad \text{for all sufficiently large } n,
\end{aligned}$$

applying inequality (42). This completes the proof of the lemma. □

The next lemma gives the form of the Bayes rule for π_n relative to the modified loss $\tilde{L}_n$.

Lemma 2. Defining the function p as in equation (40) above, the Bayes rule for π_n relative to the modified loss $\tilde{L}_n$ is given by

$$\tilde{\theta}^{\mathrm{b}}(X) = E\{\theta p(\theta)^{-1}|X\}/E\{p(\theta)^{-1}|X\}$$

and satisfies

$$\|\tilde{\theta}^{\mathrm{b}}\|^2 \leqslant \mu_n^2\,\overline{\sigma^2}\ E\{p(\theta)|X\}.$$

Proof. Use the notation E^X to denote an expectation conditional on X. The posterior modified risk of any estimator $\tilde{\theta}(X)$ is

$$\begin{aligned}
E^X \tilde{L}_n(\tilde{\theta}, \theta) &= E^X\{\|\tilde{\theta} - \theta\|^2\ p(\theta)^{-1}\} \\
&= \|\tilde{\theta}\|^2 E^X\ p(\theta)^{-1} - 2\sum_i \tilde{\theta}_i\ E^X\{\theta_i\ p(\theta)^{-1}\} + E^X\{\|\theta\|^2\ p(\theta)^{-1}\}.
\end{aligned}$$

Hence the Bayes estimator relative to $\tilde{L}_n$ is given by

$$\tilde{\theta}^{\mathrm{b}}(X) = E^X\{\theta\ p(\theta)^{-1}\}/E^X\ p(\theta)^{-1}$$

as required, proving the first part of the lemma.

For every $\theta \in \mathrm{supp}(\pi_n)$, we then have

$$\|\theta\|^2 = \sum_{i=1}^{n} \mu_n^2 \tau_i^2 \, I[\theta_i \neq 0] = \mu_n^2 \sum_{i=1}^{n} (\sigma_i^2 \wedge \tau_i^2) \, I[\theta_i \neq 0]$$
$$\leqslant \mu_n^2 \sum_{i=1}^{n} (\sigma_i^2 \wedge \mu_n^2 \tau_i^2) \, I[\theta_i \neq 0] = \mu_n^2 \sum_{i=1}^{n} (\sigma_i^2 \wedge \theta_i^2)$$
$$= \mu_n^2 \, \overline{\sigma^2} \{p(\theta) - 1\} < \mu_n^2 \, \overline{\sigma^2} \, p(\theta). \tag{43}$$

Applying the Cauchy–Schwarz inequality and then inequality (43), we obtain

$$\|\tilde{\theta}^{\mathrm{b}}\|^2 = \|E^X\{\theta \, p(\theta)^{-1}\}\|^2 / \{E^X p(\theta)^{-1}\}^2$$
$$\leqslant E^X\{\|\theta\|^2 \, p(\theta)^{-1}\} / E^X p(\theta)^{-1} < \mu_n^2 \, \overline{\sigma^2} \, \{E^X p(\theta)^{-1}\}^{-1}.$$

Now apply Jensen's inequality to complete the proof of the lemma. □

The next lemma gives bounds on the moments of $p(\theta)$ under the prior that will be useful subsequently.

Lemma 3. On the prior π_n, letting β be the constant in condition (21) above,

$$E \, p(\theta) \leqslant 1 + \log n$$

and

$$E|p(\theta) - E \, p(\theta)|^{\beta} < C_3 \log n.$$

For any fixed $\eta > 0$ let A_n be the event $\{p(\theta) \leqslant 1 + (1 + \eta) \log n\}$. Then

$$P(A_n) \to 1 \qquad \text{as } n \to \infty.$$

Proof. We have

$$p(\theta) = 1 + \overline{\sigma^2}^{-1} \sum_{i=1}^{n} (\mu_n^2 \tau_i^2 \wedge \sigma_i^2) I_i$$

where the I_i are the independent Bernoulli($n^{-1} \log n$) random variables $I[\theta_i \neq 0]$. Hence

$$E \, p(\theta) \leqslant 1 + (\overline{\sigma^2})^{-1} \sum \sigma_i^2 \, P(\theta_i \neq 0) = 1 + \log n.$$

We now use the fact that, for $1 < \beta \leqslant 2$, for any independent random variables ξ_i with mean 0

$$E|\sum_{i=1}^{n} \xi_i|^{\beta} \leqslant A_{\beta} \sum_{i=1}^{n} E|\xi_i|^{\beta}$$

for some constant A_{β} depending only on β. This is a consequence of the Marcinkiewicz–Zygmund inequality and the fact that, for positive x_i, $|\sum_i x_i|^{\beta/2} \leqslant \sum_i |x_i|^{\beta/2}$; see, for example, theorem 10.3.2 and exercise 10.2.5 of Chow and Teicher (1978). We now have

$$E|p(\theta) - E\,p(\theta)|^{\beta} = (\overline{\sigma^2})^{-\beta} E|\sum_{i=1}^{n} (\mu_n^2 \tau_i^2 \wedge \sigma_i^2)(I_i - n^{-1} \log n)|^{\beta}$$
$$\leqslant A_{\beta} (\overline{\sigma^2})^{-\beta} \sum_{i=1}^{n} \sigma_i^{2\beta} E|I_i - n^{-1} \log n|^{\beta}$$
$$\leqslant 2A_{\beta} n C_1 n^{-1} \log n = C_3 \log n$$

as required, applying condition (21) above. The last part of the lemma follows by the βth-moment version of Chebyshev's inequality. □

The next lemma gives an important technical bound on the behaviour of the modified loss Bayes estimator.

Lemma 4. Defining the event A_n as in the statement of lemma 3, the Bayes rule $\tilde{\theta}^{\mathrm{b}}$ satisfies

$$E_{\pi_n} E_{\theta}(\|\tilde{\theta}^{b} - \theta\|^2 I[A_n^{\mathrm{c}}]) = o(\mu_n^2 \overline{\sigma^2} \log n)$$

as $n \to \infty$.

Proof. Consider the term $E(\|\tilde{\theta}^{\mathrm{b}}\|^2 I[\theta \in A_n^{\mathrm{c}}])$. By lemma 2 and the Cauchy–Schwarz inequality we have

$$E(\|\tilde{\theta}^{\mathrm{b}}\|^2 I[\theta \in A_n^{\mathrm{c}}]) \leqslant \mu_n^2 \overline{\sigma^2}\, E[E\{p(\theta)|X\} I[A_n^{\mathrm{c}}]]$$
$$\leqslant \mu_n^2 \overline{\sigma^2} \{E\,p(\theta)^2\}^{1/2} P(A_n^{\mathrm{c}})^{1/2} = o(\mu_n^2 \overline{\sigma^2} \log n).$$

However, we have

$$E\|\theta\|^2 = \mu_n^2 \sum_i \tau_i^2 n^{-1} \log n \leqslant \mu_n^2 \overline{\sigma^2} \log n$$

and

$$\mathrm{var}\|\theta\|^2 \leqslant \mu_n^4 \sum_i \tau_i^4 n^{-1} \log n \leqslant \mu_n^4 C_1 (\overline{\sigma^2})^2 \log n$$

so that

$$E(\|\theta\|^2 I[\theta \in A_n^{\mathrm{c}}]) \leqslant (E\|\theta\|^4)^{1/2} P(A_n^{\mathrm{c}})^{1/2}$$
$$= O(\mu_n^2 \overline{\sigma^2} \log n) P(A_n^{\mathrm{c}})^{1/2} = o(\mu_n^2 \overline{\sigma^2} \log n),$$

completing the proof of the lemma. □

A.2.3. *Completion of proof*

We can now complete the proof of theorem 2. The Bayes risk for π_n with respect to the loss $\tilde{L}_n$ satisfies

$$\begin{aligned}\overline{\sigma^2}\,\tilde{\rho}(\pi_n) &= E_{\pi_n} E_\theta \frac{\|\tilde{\theta}^{\mathrm{b}} - \theta\|^2}{p(\theta)} \\ &\geqslant \frac{E_{\pi_n} E_\theta(\|\tilde{\theta}^{\mathrm{b}} - \theta\|^2 I[A_n])}{1 + (1+\eta)\log n} \\ &= \frac{E_{\pi_n} E_\theta \|\tilde{\theta}^{\mathrm{b}} - \theta\|^2}{1 + (1+\eta)\log n} - o(\mu_n^2\,\overline{\tau^2}),\end{aligned} \tag{44}$$

applying lemma 4 and the condition $\overline{\tau^2} \geqslant C_2^{-1}\overline{\sigma^2}$. By lemma 1, we have

$$\begin{aligned}\liminf_{n\to\infty}\left(\frac{E_{\pi_n} E_\theta\|\tilde{\theta}^b - \theta\|^2}{\overline{\tau^2}\mu_n^2 \log n}\right) &\geqslant \liminf_{n\to\infty}\left\{\frac{\rho_n(\pi_n)}{\overline{\tau^2}\mu_n^2 \log n}\right\} \\ &\geqslant \Phi(a).\end{aligned}$$

Substituting into equation (44) it follows that, as $n \to \infty$,

$$\frac{\overline{\sigma^2}}{\overline{\tau^2}}\tilde{\rho}(\pi_n) \geqslant (1+\eta)^{-1}\mu_n^2\,\Phi(a)\{1 + o(1)\} = 2(1+\eta)^{-1}\,\Phi(a)\{1 + o(1)\}\log n;$$

hence by the minimax theorem of decision theory

$$\begin{aligned}\inf_{\tilde{\theta}}\sup_{\theta}\left\{\frac{\overline{\sigma^2}}{\overline{\tau^2}} E_\theta\,\tilde{L}_n(\tilde{\theta}, \theta)\right\} &\geqslant \frac{\overline{\sigma^2}}{\overline{\tau^2}}\,\tilde{\rho}_n(\pi_n) \\ &\geqslant (1+\eta)^{-1}(2\log n)\,\Phi(a)\{1 + o(1)\}.\end{aligned} \tag{45}$$

Since inequality (45) is true for all $\eta > 0$ and all a, it follows that

$$\frac{\overline{\sigma^2}}{\overline{\tau^2}}\inf_{\tilde{\theta}}\sup_{\theta}\{E_\theta\,\tilde{L}_n(\tilde{\theta}, \theta)\} \geqslant 2\{1 + o(1)\}\log n,$$

completing the proof. □

A.3. *Proof of Theorem 3*

To facilitate the proof of theorem 3, we use a simplified form of an argument from Donoho *et al.* (1995). Writing x for the n-vector with elements x_i, define the quantity

$$W_p(\delta, C; n) = \sup_{\|x\|_p \leqslant C}\left\{\sum_{i=1}^{n}(\delta^2 \wedge x_i^2)\right\}.$$

It is easy to verify that

$$W_p(\delta, C; n) \leqslant \begin{cases}\min(n\delta^2, C^p\delta^{2-p}) & \text{if } 0 \leqslant p \leqslant 2, \\ \min(n\delta^2, C^2 n^{1-2/p}) & \text{if } 2 \leqslant p \leqslant \infty,\end{cases} \tag{46}$$

and that, whatever the value of p, the minimum in inequalities (46) is attained at $n\delta^2$ if and only if $n^{1/p}\delta \leqslant C$. To prove inequalities (46), make use of the inequality $\delta^2 \wedge x_i^2 \leqslant \delta^2 \wedge \delta^{2-p}x_i^p$ if $p < 2$ and $n^{-1/2}\|x\|_2 \leqslant n^{-1/p}\|x\|_p$ if $p > 2$.

Since $b_{p,q}^{\sigma}(C) \subseteq b_{p,\infty}^{\sigma}(C)$ for all q, it suffices to establish the result of theorem 3 for $b_{p,\infty}^{\sigma}(C)$.

Thus, we may assume that

$$\|\theta_j\|_p \leqslant 2^{-js}C \qquad \text{for all } j > 0. \tag{47}$$

Set $n = 2^J$ and define $\Delta_J(\theta) = \Sigma_{j \geqslant J} \theta_\lambda^2$. Apply theorem 1, with $\sigma_j^2 = 2^{-j(1-\alpha)}\epsilon^{2\alpha}$, to obtain

$$\begin{aligned} E\|\hat{\theta} - \theta\|^2 &= E\sum_{j<J}(\hat{\theta}_\lambda - \theta_\lambda)^2 + \Delta_J(\theta) \leqslant (2 \log n + 1)\sum_{j<J}\sum_{k=1}^{2^j}(\sigma_j^2/n + \theta_\lambda^2 \wedge \sigma_j^2) + \Delta_J(\theta) \\ &= (2 \log n + 1)(S_1 + S_2) + \Delta_J(\theta). \end{aligned}$$

Now $S_1 = 2^{-J}\Sigma_{j<J} 2^j\sigma_j^2 = O(2^{-J}) = O(n^{-1}) = o(n^{-r})$.

Next consider the ideal risk S_2. Because of constraint (47) we have

$$S_2 \leqslant \sum_{j<J} W_p(\sigma_j, 2^{-js}C; 2^j). \tag{48}$$

Because of the remark following inequalities (46), the value of each summand in inequality (48) will be determined by whether $2^{j/p}\sigma_j \leqslant 2^{-js}C$. This condition can be re-expressed as $j \leqslant \zeta$, where $2^\zeta = (2^{\alpha J}C^2)^{1/(\alpha+2\sigma)}$. The summands are therefore $2^j\sigma_j^2 = 2^{-(J-j)\alpha}$, a geometrically increasing sequence, for $j < \zeta$, and, whether or not $p < 2$, a geometrically decreasing sequence for $j > \zeta$. The largest term in each geometric sequence will be bounded above by $2^{-(J-\zeta)\alpha}$, and hence the sum in inequality (48) is bounded above by a constant multiple of $2^{-\alpha J}(2^{\alpha J}C^2)^{\alpha/(\alpha+2\sigma)}$, which is proportional to $2^{-2J\alpha\sigma/(\alpha+2\sigma)} = n^{-r(\sigma,\alpha)}$. Therefore $S_2 = O(n^{-r(\sigma,\alpha)})$, corresponding to the bound claimed in the theorem.

To complete the proof, we use the bound

$$\Delta_J(\theta) \leqslant \sum_{j \geqslant J} W_p(\infty, 2^{-js}C; 2^j).$$

When $p \geqslant 2$, inequalities (46) then show that Δ_J is bounded by a multiple of $2^{-2J\sigma} = n^{-2\sigma} = o(n^{-r})$. When $p < 2$, we note that $W_p(\infty, C; n) = C^2$, so that Δ_J is bounded by a multiple of $2^{-2Js} = n^{-2s} = o(n^{-r})$ since, in this case, $s = \sigma + \frac{1}{2} - 1/p > \frac{1}{2} > r/2$. This completes the proof.

REFERENCES

Adak, S. (1995) Tree-based estimation for time-dependent spectra for nonstationary processes. *Proc. Statist. Comput. Sect. Am. Statist. Ass.*, 75–80.

Averkamp, R. and Houdré, C. (1996) Denoising non-gaussian noise by wavelet thresholding.

Beran, J. (1994) *Statistical Methods for Long Memory Processes*. New York: Chapman and Hall.

Brillinger, D. (1994) Uses of cumulants in wavelet analysis. *Proc. Soc. Phot. Instrum. Engrs Adv. Signal Process.*, **2296**, 2–18.

——— (1996) Some uses of cumulants in wavelet analysis. *Nonparam. Statist.*, **6**, 93–114.

Chow, Y. and Teicher, H. (1978) *Probability Theory: Independence, Interchangeability, Martingales*. New York: Springer.

Cohen, A., Daubechies, I., Jawerth, B. and Vial, P. (1992) Multiresolution analysis, wavelets, and fast algorithms on an interval. *C. R. Acad. Sci.* A, **316**, 417–421.

Coifman, R. and Donoho, D. L. (1995) Translation-invariant denoising. *Lect. Notes Statist.*, **103**, 125–150.

Csörgö, S. and Mielniczuk, J. (1993) Nonparametric regression under long-range dependent errors. *Technical Report 212*. Department of Statistics, University of Michigan, Ann Arbor.

Dahlhaus, R. (1997) Fitting time series models to nonstationary processes. *Ann. Statist.*, to be published.

Donoho, D. L. (1992) Interpolating wavelet transforms. *Technical Report 408*. Department of Statistics, Stanford University, Stanford.

——— (1995) Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harm. Anal.*, **2**, 101–126.

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

——— (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.

——— (1997) Minimax estimation via wavelet shrinkage. *Ann. Statist.*, to be published.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia (with discussion)? *J. R. Statist. Soc.* B, **57**, 301–369.

Eisenberg, R. (1994) Biological signals that need detection: currents through single membrane channels. In *Proc. 16th A. Int. Conf. IEEE Engineering in Medicine and Biology Society* (eds J. Norman and F. Sheppard), pp. 32a–33a.

Flandrin, P. (1994) Time-scale analyses and self-similar stochastic processes. In *Wavelets and Their Applications* (ed. J. S. Byrnes), pp. 121–142. Dordrecht: Kluwer.

Frazier, M., Jawerth, B. and Weiss, G. (1991) Littlewood–Paley theory and the study of function spaces. *Regl Conf. Ser. Math.*, **79**.

Fredkin, D., Rice, J., Colquhoun, D. and Gibb, A. (1995) Persistence: a new statistic for characterizing ion channel activity. *Phil. Trans. R. Soc. Lond.* B, **350**, 353–367.

Gao, H.-Y. (1993) Choice of thresholds for wavelet estimation of the log spectrum. *Technical Report 438*. Department of Statistics, Stanford University, Stanford.

Hall, P. and Hart, J. D. (1990) Nonparametric regression with long-range dependence. *Stoch. Process. Applic.*, **36**, 339–351.

Hart, J. D. (1991) Kernel regression estimation with time series errors. *J. R. Statist. Soc.* B, **53**, 173–187.

Houdré, C. (1992) Wavelets, probability and statistics: some bridges. In *Wavelets: Mathematics and Applications* (eds J. J. Benedetto and M. W. Frazier). Boca Raton: CRC.

Istas, J. (1992) Wavelet coefficients of a gaussian process and applications. *Ann. Inst. H. Poincaré Probab. Statist.*, **28**, 537–556.

Johnstone, I. M. and Silverman, B. W. (1996) Asymptotically adaptive wavelet threshold selection for data with correlated noise. To be published.

Khatri, C. (1967) On certain inequalities for normal distributions and their applications to simultaneous confidence bands. *Ann. Math. Statist.*, **38**, 1853–1867.

Mallat, S. G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.*, **11**, 674–693.

Mallat, S., Papanicolaou, G. C. and Zhang, Z. (1997) Adaptive covariance estimation of locally stationary processes. *Ann. Statist.*, to be published.

Meyer, Y. (1990) *Ondelettes et Opérateurs*, vols I–III. Paris: Hermann.

Moulin, P. (1994) Wavelet thresholding techniques for power spectrum estimation. *IEEE Trans. Signal Process.*, **42**, 3126–3136.

Nason, G. P. and Silverman, B. W. (1994) The discrete wavelet transform in S. *J. Comput. Graph. Statist.*, **3**, 163–191.

Neumann, M. H. (1996) Spectral density estimation via nonlinear wavelet methods for stationary non-gaussian time series. *J. Time Ser. Anal.*, to be published.

Neumann, M. H. and von Sachs, R. (1997) Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.*, **25**, in the press.

von Sachs, R. and Schneider, K. (1996) Wavelet smoothing of evolutionary spectra by non-linear thresholding. *Appl. Comput. Harm. Anal.*, **3**, 268–282.

Sidák, Z. (1968) On multivariate normal probabilities of rectangles: their dependence on correlations. *Ann. Math. Statist.*, **39**, 1425–1434.

Stein, C. (1981) Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**, 1135–1151.

Taqqu, M. S. (1975) Weak convergence to fractional brownian motion and to the rosenblatt process. *Z. Wahrsch. Ver. Geb.*, **31**, 287–302.

Wang, Y. (1996) Function estimation via wavelet shrinkage for long-memory data. *Ann. Statist.*, **24**, 466–484.