Statistica Sinica **9**(1999), 51-83

WAVELET SHRINKAGE FOR CORRELATED DATA AND INVERSE PROBLEMS: ADAPTIVITY RESULTS

Iain M. Johnstone

Stanford University

Abstract: Johnstone and Silverman (1997) described a level-dependent thresholding method for extracting signals from correlated noise. The thresholds were chosen to minimize a data based unbiased risk criterion. Here we show that in certain asymptotic models encompassing short and long range dependence, these methods are simultaneously asymptotically minimax up to constants over a broad range of Besov classes. We indicate the extension of the methods and results to a class of linear inverse problems possessing a wavelet vaguelette decomposition.

Key words and phrases: Adaptation, correlated data, fractional brownian motion, linear inverse problems, long range dependence, mixing conditions, oracle inequalities, rates of convergence, unbiased risk estimate, wavelet vaguelette decomposition, wavelet shrinkage, wavelet thresholding.

1. Introduction

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

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

with $t_i = (i-1)/n$ and e_i drawn from a stationary Gaussian noise process. Johnstone and Silverman (1997) (JS below) discussed a number of wavelet thresholding prescriptions appropriate to estimation of f in the presence of such correlated noise e.

In particular, they described and illustrated a method for estimating thresholds from the data based on an unbiased risk estimate. In addition they introduced a family of asymptotic models encompassing both short and long range dependence and argued that the good asymptotic properties (near adaptive minimaxity) of wavelet threshold estimators are unaffected by the presence of correlations of these types.

One purpose of this paper is to describe the proof of this adaptive minimaxity result (Theorem 5 of JS) for the unbiased risk based thresholding estimates. Even in the i.i.d. error case, the proof given here simplifies and corrects that given in Donoho and Johnstone (1995), for example the technical device of random half samples used there is now avoided.

As an asymptotic model encompassing situations of both short and long range dependence, we adopt the setting used in JS. We provide some details on the decorrelating effect of the wavelet transform – for example long range dependent errors are converted in the wavelet domain into $\rho-$ mixing sequences in our model. We then show how the large deviation inequalities of Bosq (1993) for $\alpha-$ mixing sequences may be exploited to show MSE consistency of the empirical threshold choices.

It turns out that certain linear inverse problems possess a structure (captured in the *wavelet vaguelette decomposition* of Donoho (1995)) that allows many of the methods and ideas to be carried over from the regression with correlated noise setting. We describe this in brief fashion in the concluding section.

This paper is entirely concerned with stationary Gaussian errors. Of course, it would be of considerable interest to extend the results of this paper to stationary non-Gaussian errors and even to non-stationary situations. There is a recent and growing literature based on Gaussian approximations of empirical wavelet coefficients in a variety of situations. In addition to the numerous references cited at the end of Section 8 of JS, we wish to mention Neumann and von Sachs (1995).

1.1. Basic definitions and notation

We first establish some notation and recall the definition of SURE thresholding for observed data. Let $\mathcal{W}$ be a periodic discrete wavelet transform operator (in practice implemented with a fast cascade algorithm), 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 corresponding wavelet transform of the signal $\mathbf{f} = (f(t_i))_{i=1}^n$, and $\mathbf{z} = \mathcal{W}\mathbf{e}$ be the transform of the noise.

To construct the estimator, let η_S be the *soft threshold function*

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

If the noise process e is stationary, then so are the processes $k \to z_{jk}$ in the wavelet domain, and so we denote their standard deviations by σ_j. 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 might be soft or hard thresholding, or some compromise between the two, though in this paper we focus on soft thresholding. We write $\hat{\theta}$ for the corresponding estimator of θ, and set

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

Under this formulation, allowing signal at low levels ($j \leq L$, say) 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

$$\hat{\sigma}_j^2 = \mathrm{MAD}\{w_{jk}, k = 1, \ldots, 2^j\}/.6745, \tag{4}$$

where MAD denotes median absolute deviation from zero and the factor .6745 is chosen for calibration with the Gaussian distribution. Other estimates are of course possible, for example *mean* absolute deviation. We do not dwell on the estimation of the variance; we assume for the rest of this paper that it has been carried out, and treat σ_j^2 as known.

We measure error 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.

As shown in Johnstone and Silverman (1997), the co-ordinatewise nature of thresholding implies that the Stein unbiased risk estimate investigated in the i.i.d. Gaussian error setting by Donoho and Johnstone (1995) remains unbiased, even in the presence of correlation. To be specific, we consider a general multivariate normal model in which $X \sim N_d(\theta, V)$. For this paragraph, the covariance matrix V is unrestricted. Stein's method shows that the mean squared error of an estimator $\hat{\theta} = X + g(X)$ may be written

$$\begin{aligned} E\|X + g(X) - \theta\|^2 &= E\{\mathrm{tr}\ V + \|g(X)\|^2 + 2\mathrm{tr}\ [V\, Dg(X)]\} \\ &= E\{U(t; X)\}, \end{aligned} \tag{5}$$

say, where $Dg(X)$ 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

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

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 (5) is therefore diagonal.

If the covariance matrix V is homoscedastic, $v_{kk} \equiv \sigma^2$. If σ^2 is unknown, it can be estimated by $\hat{\sigma}^2$ as defined in equation (4). We will treat σ as known, and via rescaling ($x = X/\sigma$) we may assume that $\sigma^2 = 1$. The unbiased risk criterion is then obtained by substituting the properties of g:

$$U(t) = d + \sum_k (x_k^2 \wedge t^2) - 2I\{|x_k| \le t\}, \tag{6}$$

which is identical to that used in the i.i.d. case. We therefore propose taking

$$\hat{t}(x) = \operatorname{argmin}_{0 \le t \le \sqrt{2\log d}}\, \hat{U}(t). \tag{7}$$

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

For ‘small’ sparse signals, the noise coordinates can swamp the signal coordinates in their contribution to the SURE criterion. The behaviour of $\hat{t}$ can be erratic, so one alternative is to retreat to the use of higher, ‘fixed’ thresholds $t^F = \sqrt{2\log d}$. For further details, see Donoho and Johnstone (1995). The pretest compares an unbiased estimate of $\|\theta\|^2$, namely $s_d^2 = d^{-1}\sum_1^d x_k^2 - 1$, to a threshold γ_d:

$$\tilde{t}(x) = \begin{cases} \sqrt{2\log d} & s_d^2 \le \gamma_d \\ \hat{t}(x) & s_d^2 > \gamma_d \end{cases}. \tag{8}$$

Thus the unbiased risk choice $\hat{t}$ of (7) is chosen only when the pretest rejects.

Returning to 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 (6). We then set

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

We recall that this estimator, and relatives with different threshold choices are all easily implemented with $O(n \log n)$ algorithms in software (for example 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) and illustrations on various kinds of simulated and physiological data were given in Section 3 of JS.

1.2. Asymptotic model

In Johnstone and Silverman (1997), it was suggested that a useful class of asymptotic caricatures of the finite sample model (1) is given by

$$Y(t) = F(t) + \epsilon^\alpha B_H(t), \qquad t \in [0,1]. \tag{9}$$

Here, our target function for estimation is $f = F'$ and $B_H(t)$ is *fractional Brownian motion*, namely the zero mean Gaussian process on $\mathcal{R}$ with covariance function $r(s,t)$ given by

$$r(s,t) = \frac{V_H}{2}(|s|^{2H} + |t|^{2H} - |t-s|^{2H}), \qquad s,t \in \mathcal{R}, \qquad H \in [1/2, 1).$$

The parameter $\alpha = 2(1-H) \in (0,1]$, and the scale parameter ϵ is thought of as proportional to $n^{-1/2}$.

Let $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$ be a wavelet basis on $\mathcal{R}$ derived from a suitable wavelet ψ of compact support with corresponding scaling function ϕ. Here the index λ runs over a set Λ defined by pairs $(j,k), j \geq L, k = 1, \ldots, 2^j$ for the wavelet functions and $(L-1,k), k = 1, \ldots, 2^L$ for the scaling functions. Form the inner products $y_\lambda = \int \psi_\lambda dY, \theta_\lambda = \int \psi_\lambda f$, and $\gamma_j z_\lambda = \int \psi_\lambda dB_H$ where $\gamma_j^2 = \mathrm{Var}\left\{\int \psi_\lambda dB_H\right\} = \tau^2 2^{-j(1-\alpha)}$. Note that for $j = L-1$, the inner products are taken with ϕ_λ. To avoid annoying but inconsequential end effects, we argue as in Johnstone and Silverman (1997) that a model nearly equivalent to (9) (i.e. involving an approximation of the variance structure valid up to absolute multiplicative constants) may be obtained as

$$y_\lambda = \theta_\lambda + \epsilon^\alpha \gamma_j z_\lambda, \qquad \lambda \in \Lambda. \tag{10}$$

The noise variables z_λ, which all have variance 1, are correlated, but can be shown to have relatively weak dependence, in a sense articulated explicitly in Section 2.

We can think of the initial segments $\{y_{jk} : j < J = \log_2 n, k = 1, \ldots, 2^j\}$ in model (10) with $\epsilon = \tau^{1/\alpha} n^{-1/2}$ as being analogous to the empirical coefficients w_{jk} in (2). While this is not literally correct, of course, one can use this identification to transfer intuition from the asymptotic models to empirical data.

It is, however, simpler to do rates of convergence calculations in the approximating model (10). By some general decision theoretic and wavelet theoretic machinery (Donoho and Johnstone (1999)) we expect that these results can be carried over to the original regression model (1) : some discussion of the issues involved is given in the Appendix.

It will be assumed that the parameters α and τ are known—since the latter is a simple scale parameter, we will set $\tau = 1$ without further loss of generality. (Some discussion of estimation of α, τ appears in Section 7.3 of JS.) We will 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 (10) and, by setting $\alpha = 1$, the short-range dependence approximation.

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 $\|\theta_j\|_p^p = \sum_{k=1}^{2^j} |\theta_{jk}|^p$ and

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

For a fuller discussion of these spaces and the important roles of the indices (σ, p, q) see Frazier, Jawerth and Weiss(1991) and Donoho and Johnstone(1998b). Here we note simply that σ is a smoothness parameter, corresponding to the number of derivatives that the function f possesses in L_p. The case $p = q = \infty$ corresponds to Hölder smoothness, defined by the uniform condition $|D^m f(x) - D^m f(y)| \leq C_0 |x - y|^{\delta}$, where $\sigma = m + \delta$ with $\delta \in (0, 1]$.

To state the main result, we consider the sequence model (10), and soft threshold estimators of the form

$$\begin{aligned} \hat{\theta}_{\lambda}^{*} &= \eta_S(y_{\lambda}, \sigma_j \lambda_j) \\ \lambda_j &= \begin{cases} 0 & j \leq L \\ \tilde{t}(y_j/\sigma_j) & j \geq L, \end{cases} \end{aligned} \tag{11}$$

where $\tilde{t}$ is the pretest threshold given in (8).

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

$$R_{T,\alpha}^{*}(\epsilon; b_{p,q}^{\sigma}(C)) = \inf_{(t_j)} \sup_{\theta \in b_{p,q}^{\sigma}(C)} E\|\hat{\theta}_{(t_j)} - \theta\|^2,$$

where $\hat{\theta}_{(t_j)}$ stands for the estimator $(\eta(y_{j,k}, t_j))_{jk}$ and $\|\theta\|^2 = \sum_{\lambda \in \Lambda} \|\theta_{\lambda}\|^2$ is the $\ell_2(\Lambda)$ sequence norm. From results of Wang (1996), it is known that this minimax threshold risk in model (10) is of the same order in ϵ as the minimax risk over *all* estimators: i.e. there is no great loss of efficiency, even in the presence of long-range dependence, due to co-ordinatewise thresholding, and

$$R_{T,\alpha}(n^{-1/2}, b_{p,q}^{\sigma}(C)) \asymp n^{-r(\sigma,\alpha)}, \qquad r(\sigma, \alpha) = 2\sigma\alpha/(2\sigma + \alpha).$$

Against this background, we have the following result for the estimator $\hat{\theta}^{*}$ of (11) using the unbiased risk based choice of thresholds:

Theorem 1. *Set the pretest threshold in* (8) *at* $\gamma_d = 2^{-\sqrt{\log_2 d}}$ *and let* $\Delta = (1/p - 1/2)_+$. *Suppose that* $1 \leq p, q \leq \infty, 0 < C < \infty$, *and* $\sigma > \max(\alpha\Delta, \Delta - \alpha/2, 2\alpha\Delta - \alpha/2)$. *Then, as* $n = \epsilon^{-1/2} \to \infty$,

$$\sup_{\theta \in b_{p,q}^{\sigma}(C)} E\|\hat{\theta}^{*} - \theta\|^2 \leq R_{T,\alpha}^{*}(n^{-1/2}; b_{p,q}^{\sigma}(C))(1 + o(1)).$$

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 explicitly set optimal thresholds. Note especially that the extra logarithmic term present in Theorem 3 of Johnstone and Silverman (1997) has been removed, due to the lower thresholds chosen by the data-based rule.

In our result we do assume that the dependence parameter α and overall scale τ are known, and so we do not incorporate estimation of scale through estimator (4). In Johnstone and Silverman (1997) it was shown how to incorporate estimation of α in model (10) for a different choice of threshold – we expect that a similar extension here would also be valid, at the cost of further technical complications to the proof. The same comment could also be made regarding estimation of τ.

Outline of paper. Section 2 develops some correlation properties of wavelet noise coefficients in model (10). In Section 3, some asymptotic oracle inequalities are established for the SURE/pretest threshold selector (8). These inequalities are to be used in later sections on the wavelet coefficients derived from a single resolution level - they are stated under a set of dependence assumptions (Model "S") that is general enough to include both the fractional Brownian motion setting of Section 2 and the extensions to linear inverse problems outlined in Section 5. Section 4 gives the principal parts of the proof of Theorem 1, with the details being postponed to the Appendix. In particular, the role of large deviation properties for $\alpha-$ mixing seqences is indicated. Section 5 indicates how these results might be carried over to a class of linear inverse problems where the observed representer coefficients in the WVD model satisfy on each level the dependence assumptions of Model "S". The analog of Theorem 1 is given as Theorem 6. Finally, in addition to proof details, the Appendix contains some remarks on the possible extension of these results to the sampling model (1).

2. Properties of the fBM Sequence Model

Rewriting (10), we consider the model

$$y_{jk} = \theta_{jk} + \epsilon^{\alpha} w_{jk} \tag{12}$$

$$w_{jk} = \int \psi_{jk} dB_H. \tag{13}$$

In this section, we draw conclusions about the reduced dependence structure of the wavelet coefficients w_{jk} which will form the basis for the proof of Theorem 1.

Following for example Barton and Poor (1988), the stochastic integrals defining w_{jk} have mean zero and covariances given by

$$E\int f dB_H \overline{\int g dB_H} = (1/2\pi)\int \hat{f}(\xi)\overline{\hat{g}(\xi)}|\xi|^{-(1-\alpha)}d\xi. \tag{14}$$

In this section, we deduce some properties of the error process $j \to \{w_{jk}\}$ that will be used in the proof of Theorem 2. First, we use scaling properties of $\mu(d\xi) = |\xi|^{-(1-\alpha)}d\xi$ in (14), along with $\widehat{\psi_{jk}}(\xi) = 2^{-j/2}e^{ik2^{-j}\xi}\hat{\psi}(2^{-j}\xi)$, to conclude that

$$r_j(k) \triangleq E w_{jk} w_{j0} = 2^{-j(1-\alpha)} r_0(k). \tag{15}$$

In what follows, chiefly for convenience, we use the *Meyer wavelet* ψ (compare, for example Daubechies (1992), Chapter 4). We recall, in particular, the properties

(i) supp $\hat{\psi} \subset [-8\pi/3, -2\pi/3] \cup [2\pi/3, 8\pi/3]$,

(ii) On $[-2\pi, -\pi] \cup [\pi, 2\pi]$, $|\hat{\psi}| \geq c_0$.

1°. *Decay of autocorrelations* $r_0(k)$. Using (14),

$$r_0(k) = (1/2\pi)\int e^{ik\xi}|\hat{\psi}(\xi)|^2|\xi|^{-(1-\alpha)}d\xi \tag{16}$$

$$= (k^{-\alpha}/2\pi)\int e^{i\omega}|\hat{\psi}(\omega/k)|^2|\omega|^{-(1-\alpha)}d\omega. \tag{17}$$

From property Meyer (i), we can assume $|\hat{\psi}(\omega)| \leq c_{M_0}|\omega|^{M_0} I\{|\omega| \leq 3\pi\}$. This implies

$$\begin{aligned} |r_0(k)| &\leq (c_{M_0}^2/\pi)k^{-\alpha-2M_0}\textstyle\int_0^{3\pi}|\omega|^{2M_0+\alpha-1}d\omega \\ &\leq c_{\alpha M}k^{-\alpha-2M_0}. \end{aligned} \tag{18}$$

Here and throughout, $c_{\alpha M}$ denotes a constant depending on α and M_0, not necessarily the same at each appearance.

2°. *Spectral density of* $k \to w_{0k}$: Decompose $(-\infty, \infty)$ into union of intervals of length 2π centered at $2\pi\mathcal{Z}$ and apply to (16) to get

$$\begin{aligned} r_0(k) &= (1/2\pi)\textstyle\int_{-\pi}^{\pi} e^{ik\xi}\sum_{j\in\mathcal{Z}}|\hat{\psi}(\xi+2\pi j)|^2|\xi+2\pi j|^{-(1-\alpha)}d\xi \\ &= (1/2\pi)\textstyle\int_{-\pi}^{\pi} e^{ik\xi}f_\psi(\xi)d\xi. \end{aligned} \tag{19}$$

Because of the support properties of the Meyer wavelet, we conclude that f_ψ is bounded away from 0 and ∞:

$$0 < f_{\min} \leq f_\psi(\xi) \leq f_{\max} < \infty \qquad \text{for all } |\omega| \leq \pi. \tag{20}$$

From this follows a bound on eigenvalues: For any $(a_j) \in \ell_2(\mathcal{Z})$:

$$f_{\min} \leq \frac{\sum_{jk} a_j r_0(j-k)\bar{a}_k}{\sum_j |a_j|^2} \leq f_{\max}. \tag{21}$$

Indeed using (19), we have

$$\sum_{jk} a_j r(j-k)\bar{a}_k = (1/2\pi)\int_{-\pi}^{\pi} |\sum_j a_j e^{ij\xi}|^2 f_\psi(\xi) d\xi,$$

and using (20) together with orthogonality of exponentials gives (21).

3°. *The sequence $k \to w_{jk}$ is $\rho-$ mixing:* We recall from (e.g. Ibragimov and Rozanov (1978)) that for stationary *Gaussian* sequences $(X_n, n \in \mathcal{Z})$, we may define

$$\rho(n) = \sup_{(a_j),(b_k)} |\text{Corr } (\sum_{j\leq 0} a_j X_j, \sum_{k \geq n} b_k X_k)|. \tag{22}$$

Using correlation bound (18) on $k \to w_{0k}$ and setting $\beta = \alpha + 2M_0$,

$$\begin{aligned}
\text{Cov } (\sum_{j\leq 0} a_j w_{0j}, \sum_{k\geq n} b_k w_{0k}) &\leq c_{\alpha M} \sum_j \sum_k |a_j||j-k|^{-\beta}|b_k| \\
&\leq c_{\alpha M} \sum_{j \leq 0} |a_j| \cdot \|b\| (\sum_{k\geq 0} (n + |j| + k)^{-2\beta})^{1/2} \\
&\leq c_{\alpha M} \|a\| \|b\| [\sum_{j \geq 0} (n+j-1)^{1-2\beta}]^{1/2} \\
&\leq c_{\alpha M} \|a\| \|b\| (n-2)^{-(\beta-1)},
\end{aligned}$$

since $\sum_{k=0}^{\infty} |r_0 + 1 + k|^{-2\beta} \leq r_0^{1-2\beta}/(2\beta-1)$ for $\beta > 1/2$.

From the eigenvalue bound (21) on the spectral density,

$$\sum a_j^2 \leq f_{\min}^{-1} \cdot \text{Var } (\sum_j a_j w_{0j}).$$

This yields a $\rho-$ mixing rate for $k \to w_{0k}$:

$$\rho(n) \leq c_{\alpha M} f_{\min}^{-1} n^{-\alpha - 2M_0 + 1}. \tag{23}$$

Remark. The chief advantage of the Meyer wavelet for this paper is that it possesses infinitely many vanishing moments - this implies, as in step 1° above, that the autocorrelations $r_0(k)$ and hence the $\rho-$mixing rate $\rho(n)$ have decay faster than any polynomial. This makes it trivial to find values of M_0 such that inequalities (69) and (75) below are satisfied. The same conclusions could be achieved with other wavelets (e.g. having compact support in the time domain, such as the various Daubechies families) that have a sufficiently large number M_0 of vanishing moments. We have not carried out this analysis in detail, since it is not clear that the large deviation inequalities which lead to the conditions (69) and (75) are in fact optimal for this application.

3. Asymptotic Oracle Inequalities for Hybrid-SURE

The goal of this section is to formulate and prove Theorem 2, showing that the hybrid-SURE prescription is asymptotically consistent, in the sense that it essentially achieves the best mean squared error possible among threshold estimators. The result is 'local' in the sense that it applies for each signal μ, and not just to worst case behavior over a set of signals.

We assume that the data $(x_i)_{i=1}^d$ is a subset (increasing as d increases) of a stationary Gaussian sequence satisfying a set of conditions that we denote "Model S":

Model S. $x_i = \mu_i + z_i \qquad i = 1, \ldots, d$ where $(z_i)_{i \in \mathcal{Z}}$ is a mean zero, unit variance stationary Gaussian sequence such that

(i) $r_z(k) = \displaystyle\int_{-\infty}^{\infty} e^{ik\xi} |\widehat{\psi}(\xi)|^2 g(\xi)\, d\xi$; where ψ is the Meyer wavelet,

(ii) $g(\xi)$ is continuous and positive for $|\xi| \in [\frac{\pi}{2}, 3\pi]$,

(iii) there exist constants $c_1 > 0, c_2 \in \mathcal{R}$ such that $|g(\xi)| \le c_1 |\xi|^{c_2}$ as $|\xi| \to 0$.

Although model S appears to impose rather special structure on the noise process (z_i), it is precisely the situation that applies to wavelet coefficients $\{w_{jk}\}$ from model (12) when the level j is held fixed. Compare (15) and (16). It also is designed to apply to the noise processes arising in the linear inverse problem settings described in Section 5.

Let us note some consequences of Model S which will be used in the proof of Theorem 2. These properties reflect the fact that Model S forces the sequence (z_i) to be "close to i.i.d.", and were established in the previous section for the case of the noise processes $k \to w_{jk}$ in model (12).

(i) The spectral density $f_z(\xi)$ satisfies

$$0 < f_{\min} \le f_z(\xi) \le f_{\max} < \infty \qquad \text{for} \quad |\xi| \le \pi;$$

(ii) correlation decay: Given $M_0 \in \mathbb{N}$, $\exists\, c_3 = c_3(M_0, g)$ s.t.

$$|r_z(k)| \le c_3 k^{-c_2 - 1 - 2M_0};$$

(iii) ρ-mixing:

$$\rho_z(n) \le (c_3 / f_{\min}) n^{-c_2 - 2M_0}. \tag{24}$$

To state the theorem, we define the *ideal threshold risk:*

$$R(\mu) = \inf_t \, d^{-1} \sum_{i=1}^{d} r(t, \mu_i),$$

where $r(t,\mu_i) = E[\eta_S(x_i,t) - \mu_i]^2$ (compare (3)) and the *fixed threshold risk:*

$$R_F(\mu) = d^{-1} \sum_{1}^{d} r(t_d^F, \mu_i).$$

As in Section 1.1, we define our hybrid-SURE thresholding estimator in Model S by setting $x = (x_i)_{i=1}^d$, $s_d^2 = d^{-1}\sum_1^d (x_i^2 - 1)$, and

$$\hat{\mu}^*(x)_i = \begin{cases} \eta(x_i, t_d^F) \text{ if } s_d^2 \le \gamma_d \\ \eta(x_i, \hat{t}) \quad \text{if } s_d^2 > \gamma_d \end{cases}, \tag{25}$$

where

$$\hat{t} = \mathrm{argmin}_{0 \le t \le t_d^F} SURE(t, (x_i)), \qquad t_d^F = \sqrt{2 \log d}.$$

Thus, when $\tau^2 = \tau^2(\mu) = d^{-1}\sum_1^d \mu_i^2$ is small, the hybrid estimator will usually default to the fixed threshold choice. When $\mu \equiv 0$, this is guaranteed by the following large deviations inequality for s_d^2, established in the Appendix. Suppose that $\{z_i, i = 1, \ldots, d\}$ is drawn from a Gaussian stationary sequence of mean zero, variance one and bounded spectral density $f(\omega)$ on $[0,\pi]$: $f_\infty = \sup\{f(\omega) : 0 \le \omega \le \pi\}$.Then

$$P\{|s_d^2| > t\} \le 2\exp\{-\frac{dt}{8f_\infty} \cdot \min(t/f_\infty, 1)\}. \tag{26}$$

Theorem 2. *Assume stochastic model S, and that*

$$1 \ge \gamma_d >> \sqrt{d^{-1}\log d}.$$

(a) *For each $\eta \in (0, 1/2)$, uniformly in $\mu \in \mathbb{R}^d$*

$$d^{-1}E \parallel \hat{\mu}^* - \mu \parallel^2 \le R(\mu) + R_F(\mu) I\{\tau^2 \le 3\gamma_d\} + O(d^{\eta - \frac{1}{2}}), \qquad d \to \infty. \tag{27}$$

(b) *There exists a positive constant c_* such that uniformly in μ for which $\tau^2(\mu) \le \frac{1}{3}\gamma_d$,*

$$d^{-1}E \parallel \hat{\mu}^* - \mu \parallel^2 \le R_F(\mu) + O(\log d \, e^{-c_* d \gamma_d^2}), \qquad d \to \infty. \tag{28}$$

Notation. We begin the outline of the proof of Theorem 2 with some definitions. Let F_d denote the empirical c.d.f. of $\{\mu_1, \ldots, \mu_d\}$. The empirical loss of thresholding estimator $(\eta_S(x_i, t))_{i=1}^d$ is

$$\hat{L}(t, F_d) = d^{-1} \sum_1^d [\eta(x_i, t) - \mu_i]^2,$$

and its corresponding risk

$$r(t, F_d) = E\hat{L}(t, F_d) = d^{-1} \sum_1^d r(t, \mu_i).$$

The unbiased estimate of the risk of $(\eta_S(x_i, t))$ is given by

$$U_d(t) = d^{-1} \sum_1^d 1 - 2\{x_i^2 \leq t^2\} + x_i^2 \wedge t^2,$$

and it satisfies $EU_d(t) = r(t, F_d)$.

We use a *pretest decomposition* based on the event $A_d = \{s_d^2 > \gamma_d\}$, which tests for the presence of significant signal. Thus,

$$R_d(\mu) = d^{-1} E\| \hat{\mu}^* - \mu \|^2 = R_{1d}(\mu) + R_{2d}(\mu) \tag{29}$$

and to establish Theorem 2 we show that

$$\begin{aligned} R_{1d}(\mu) &\overset{\Delta}{=} d^{-1} E\{\| \hat{\mu}^* - \mu \|^2, A_d\} \leq R(\mu) + c d^{\eta - \frac{1}{2}}, \\ R_{2d}(\mu) &\overset{\Delta}{=} d^{-1} E\{\| \hat{\mu}^* - \mu \|^2, A_d^c\} \leq R_F(\mu) I\{\tau^2 \leq 3\gamma_d\} + o(d^{-1/2}). \end{aligned}$$

Strategy for "signal" term R_{1d}. On A_d, we use SURE threshold $\hat{t}$, so that

$$R_{1d}(\mu) \leq d^{-1} \sum \left[\eta(x_i, \hat{t}\,) - \mu_i\right]^2 = E\hat{L}(\hat{t}, F_d).$$

We first address an issue that was overlooked in Donoho and Johnstone (1995). The minimization yielding $\hat{t}$ is performed only over the interval $[0, t_d^F]$, so that a priori one might expect only that $R_{1d}(\mu)$ approximates

$$\tilde{R}(\mu) = \inf_{0 \leq t \leq t_d^F} d^{-1} \sum r(t, \mu_i) \geq R(\mu).$$

In fact there is little to be gained by searching over larger thresholds. In the Appendix we prove

Lemma 3. *If $d \geq 8$ and $\delta > 0$, then for all $\mu \in \mathcal{R}^d$,*

$$\tilde{R}(\mu) - R(\mu) \leq 2d^{-1} e^{\sqrt{2 \log d}} = O(d^{\delta - 1}).$$

Consequently, in the proof below, we may replace $R(\mu)$ by $\tilde{R}(\mu)$ with no harm to the uniform $O(d^{\eta-1/2})$ term.

$\hat{L}(t, F_d)$ and $\hat{U}(t)$ are both unbiased for $EL(t, F_d)$, so we have the bound and the basic decomposition

$$R_{1d}(\mu) - \tilde{R}(\mu) \leq E\Delta_d,$$

$$\Delta_d = \hat{L}(\hat{t}, F_d) - r(\hat{t}, F_d) + r(\hat{t}, F_d) - U_d(\hat{t}\,) + U_d(\hat{t}\,) - \inf_{t \leq t_d^F} r(t, F_d).$$

We bound Δ_d via maximal deviations of empirical process type:

$$|\Delta_d| \leq \| \hat{L}(\cdot, F_d) - r(\cdot, F_d) \|_\infty + 2 \, \| U_d(\cdot) - r(\cdot, F_d) \|_\infty,$$

where $\|g\|_\infty = \sup\{|g(t)| \ , \ 0 \leq t \leq t_d^F\}$.

For R_{1d}, the task is thus to show a uniform bound on the expected maximal risk deviation

$$\sup_{\mu \in \mathbb{R}^d} E_\mu \| \hat{L}(\cdot, F_d) - r(\cdot, F_d) \|_\infty \leq c d^{\eta-1/2},$$

and a similar bound for the unbiased risk deviation. For this purpose, let

$$W_d(t) = \hat{L}(t, F_d) - r(t, F_d),$$
$$Z_d(t) = U_d(t) - r(t, F_d).$$

For expectations of $V = \|W_d\|_\infty$ or $\|Z_d\|_\infty$, we use the simple bound

$$EV \leq c + (EV^2)^{1/2} P(V > c)^{1/2}, \tag{30}$$

where $c = c_d$ will be of the desired order $O(d^{\eta-1/2})$.

To estimate sup norms, we use a simple discretization: let $t_j = j\delta$, and $\mathcal{J}$ denote the set of indices j for which $t_j \in [0, t_d^F]$. Clearly,

$$\|f\|_\infty \leq \sup_{\mathcal{J}} |f(t_j)| + \sup_{\mathcal{J}} \Delta f(t_j, \delta), \tag{31}$$

where

$$\Delta f(t_j, \delta) = \sup\{|f(t) - f(t_j)| : t_j \leq t \leq t_j + \delta\}.$$

Analysis of W_d We now describe how we reduce the analysis of W_d and Z_d to the application of appropriate exponential inequalities for dependent sequences. Write

$$\begin{aligned} W_d(t) \ &= \ d^{-1} \sum_{i=1}^{d} [\eta(x_i, t) - \mu_i]^2 - r(t, \mu_i) \\ &\overset{def}{=} d^{-1} \sum_{i=1}^{d} Y_i(t), \end{aligned} \tag{32}$$

from which it is seen that $t \to W_d(t)$ is continuous and piecewise differentiable. From the formula for $r(t,\mu)$ in Donoho and Johnstone (1994) and direct calculation,

$$0 \le \partial r(t,\mu)/\partial t \le 2t, \qquad \text{and} \tag{33}$$

$$(\partial/\partial t)\ [\eta(x_i,t)-\mu_i]^2 = -2\text{sgn } x_i[z_i - (\text{sgn } x_i)t], \tag{34}$$

from which

$$\|W_d'\|_\infty \le 2 \text{ ave } |z_i| + 4t_d. \tag{35}$$

Let

$$A_d = \cap_{j\in J}\{|W_d(t_j)| \le d^{\eta-1/2}\},$$
$$B_d = \{\|W_d'\|_\infty \le 6t_d\}.$$

Then, so long as δ satisfies $6t_d\delta \le d^{\eta-1/2}$, we have, from (31)

$$A_d \cap B_d \Rightarrow \|W_d\|_\infty \le 2d^{\eta-1/2}. \tag{36}$$

In view of (30), (36) and the corresponding bound (76) for $\|Z_d\|_\infty$ described in the Appendix, the chief remaining task in the bound for the signal term R_{1d} is to obtain good tail bounds for the probabilities of A_d^c, B_d^c, C_d^c and D_d^c. When the errors z_i are i.i.d., this was accomplished in Donoho and Johnstone (1995) using the well-known Hoeffding exponential inequalities for large deviations. When the errors z_i are dependent, but strongly mixing, Bosq (1993) has provided some explicit large deviations inequalities of Bernstein type which we restate here for the convenience of the reader.

Bosq's inequalities. Let $(X_i, i \in \mathcal{Z})$ be a zero-mean stochastic process, and let $\mathcal{F}_{-\infty}^i$ and $\mathcal{F}_{i+p}^\infty$ be the sigma-fields generated respectively by $\{X_s, s \le i\}$, and $\{X_s, s \ge i+p\}$. The $\alpha-$ strong mixing coefficients are defined by

$$\alpha(p) = \sup_i \sup_{A,B}\{|P(A\cap B) - P(A)P(B)|,\ A \in \mathcal{F}_{-\infty}^i,\ \mathcal{B} \in \mathcal{F}_{i+p}^\infty\}.$$

The first Bosq inequality is oriented towards bounded X_i and is given in the Appendix. The second inequality imposes the 'Cramer conditions':

Proposition 4. *Assume that there exist constants $m \le M$ such that*

(a) $0 < m \le EX_i^2 \le M, \quad i \in \mathcal{Z}$ (37)

(b) $E|X_i|^\gamma \le M^{\gamma-2}\gamma! EX_i^2; \quad \gamma \ge 3, \quad i \in \mathcal{Z}.$

If $S_n = X_1 + \cdots + X_n$, and $1 \le p_n \le \frac{n}{2}$, then for every $\epsilon > 0$ and $\gamma \ge 2$, we have

$$P(|S_n| > n\epsilon) \le (2p_n + 1 + M^{-1})\exp(-\frac{1}{10M}\frac{\epsilon^2}{5+\epsilon}\frac{n}{p_n}) \tag{38}$$

$$+d_\gamma(1+\epsilon^{-1})^{\beta_\gamma} n\alpha(p_n)^{2\beta_\gamma}, \tag{39}$$

where $\beta_\gamma = \gamma/(2\gamma+1)$ *and*

$$d_\gamma = 11[M^{(\gamma-1)/\gamma}(\gamma!)^{1/\gamma}\frac{5}{2}(1+\frac{4}{5\sqrt{m}})]^{\beta_\gamma}.$$

We illustrate the use of Proposition 4 to bound $P(A_d)$. We show that the Cramér conditions apply to $X_i = Y_i(t) + \epsilon_i$, where $Y_i(t)$ is defined in (32) and $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$ are introduced simply to ensure that $EX_i^2 \geq E\epsilon_i^2 = 1$, so that $m = 1$ in (37). Since ϵ_i have a distribution symmetric about zero, one can verify that, with $S_d = \sum_1^d X_i$,

$$P\{|W_d(t)| > c\} \leq 2P\{d^{-1}|S_d| > c\}. \tag{40}$$

Analysis of soft thresholding shows that

$$[\eta(x_i,t) - \mu_i]^2 \leq 2(z_i^2 + t^2), \qquad \text{and} \tag{41}$$

$$r(t,\mu) \leq 1 + t^2. \tag{42}$$

Using the bounds $E|U+V|^\gamma \leq 2^{\gamma-1}\{E|U|^\gamma + E|V|^\gamma\}$ and $EZ^{2k} \leq 2^k k!$ for $Z \sim N(0,1)$, and after some analysis, one verifies that (37) holds with the conservative choice $M = M_d = 16^3(1+t_d^2)^3$.

Bound (24) provides control on the $\rho-$ mixing rate $\rho_Z(p)$ for the underlying stationary Gaussian noise sequence $\{Z_i\}$. We call upon some standard results relating mixing coefficients to derive bounds on $\alpha_X(p)$, the $\alpha-$ mixing rate needed to apply the Bosq bounds. Indeed

$$4\alpha_X(p) \leq \rho_X(p) \tag{43}$$

$$\leq \rho_Y(p) \tag{44}$$

$$\leq \rho_Z(p). \tag{45}$$

It should be noted here that

$$\rho_X(p) \overset{\Delta}{=} \sup_i \sup\{|\text{Corr }(U,V)| \ : \ U \in L_2(\mathcal{F}_{-\infty}^i), V \in L_2(\mathcal{F}_{i+p}^\infty)\},$$

which reduces to the earlier expression (22) only in the stationary Gaussian case. Inequality (43) is standard (Bradley (1986), p. 166), while (44) follows, e.g. from the Csaki-Fischer theorem (Bradley (1986), p. 173), and (45) follows because $Y_i(t)$ is a function of z_i. We conclude, for the Meyer wavelet, that for any M_0,

$$\alpha_X(p) \leq c_{\alpha M} f_{\min}^{-1} p^{-\alpha - 2M_0}. \tag{46}$$

To apply Proposition 4, set $c = d^{\eta - 1/2}$ in (40), and $p_d = d^\eta$. We may write bound (46) in the form $\alpha(p) \leq ap^{-b}$ where a and b depend on (α, M_0). Using

c_1 to denote a constant depending on γ alone, we obtain from (38), after some bounding,

$$P\{|W_d(t)| \geq d^{\eta-1/2}\} \leq \epsilon_{1d}, \tag{47}$$

$$\epsilon_{1d} = 2d^{\eta} \exp(-d^{\eta}/60M_d) + c_1 M_d^{1/2} d^{1-(2b\eta+\eta-1/2)\beta_\gamma}. \tag{48}$$

This bound, with appropriate choice of M allows control of $P(A_d)$. The completion of the bound for $E\|W_d\|_\infty$, the corresponding bound for $E\|Z_d\|_\infty$, and the treatment of the 'noise' term R_{2d}, and the proof of part (b) of Theorem 2 are deferred to the Appendix.

A modified oracle inequality for fixed thresholds. Before concluding this section we state a slightly improved version of the oracle inequality of Donoho and Johnstone (1994) and Johnstone and Silverman (1997). From Lemma 1 of Donoho and Johnstone (1994), the risk of univariate soft thresholding satisfies the following two inequalities for all $t > 0, \mu \in \mathcal{R}$:

$$r(t,\mu) \leq r(t,0) + \mu^2,$$
$$r(t,\mu) \leq t^2 + 1.$$

Combining these yields

$$r(t,\mu) \leq r(t,0) + (t^2+1)(\mu^2 \wedge 1).$$

This yields immediately:

Proposition 5. *If X_i has marginal distribution $N(\mu_i, 1), i = 1, \ldots d$, and $\hat{\mu}_t$ denotes co-ordinatewise soft thresholding at t, then*

$$E\|\hat{\mu}_t - \mu\|^2 \leq dr(t,0) + (t^2+1)\sum_1^d \mu_i^2 \wedge 1. \tag{49}$$

We recall from (A2.6) of Donoho and Johnstone (1994) that for $t \geq 3/2$,

$$r(t,0) \leq 8t^{-3}\phi(t).$$

In particular, if $t_d = \sqrt{2\log d}$, then $r(\sqrt{2\log d}, 0) \leq 2d^{-1}(\log d)^{-3/2}$ and

$$E\|\hat{\mu}_{t_d} - \mu\|^2 \leq 2(\log d)^{-3/2} + (2\log d + 1)\sum_1^d \mu_i^2 \wedge 1.$$

Finally, if now y_i has marginal distribution $N(\theta_i, \sigma^2), i = 1, \ldots, d$, if $\hat{\theta}^F$ denotes soft thresholding at $t_d\sigma$, and if $d = 2^j$, then

$$E\|\hat{\theta}^F - \theta\|^2 \leq c(j^{-3/2}\sigma^2 + j\sum \theta_i^2 \wedge \sigma^2). \tag{50}$$

The significance of (50) vis-a-vis previous versions of the oracle inequality is that the first term $j^{-3/2}\sigma^2$ is now a summable sequence in j. We note also that the bound (A.24) in Donoho and Johnstone (1995) is incorrect – consider $\Xi \equiv 0$ – but that systematic use of (50) below corrects the error.

4. Proof of SURE Minimaxity

In this section we describe the main steps in the proof of Theorem 1. We assume model (12) and note that it implies

$$\text{Var } y_{jk} = \sigma_{jk}^2 = \epsilon^{2\alpha} 2^{-j(1-\alpha)}. \tag{51}$$

We rescale to use variance one oracle inequalities by level:

$$\begin{array}{ll} x_j = (y_{jk}/\sigma_j) & \\ \mu_j = (\theta_{jk}/\sigma_j) & \end{array} \qquad \hat{\theta}^*_{jk}(y) = \begin{cases} y_j & j < L \\ \sigma_j \hat{\mu}^*(x_j) & j \ge L, \end{cases}$$

where L is fixed. Then

$$\begin{aligned} E \parallel \hat{\theta}^* - \theta \parallel^2 &= \sum_{jk} E(\hat{\theta}^*_{jk} - \theta_{jk})^2 \\ &= \sum_{j<L} 2^j \sigma_j^2 + \Big(\sum_{L \le j \le j_0} + \sum_{j > j_0} \Big) \; \sigma_j^2 E \|\hat{\mu}^*(x_j) - \mu_j\|^2 \\ &\le O(\epsilon^{2\alpha}) + S_{1\epsilon} + S_{2\epsilon}. \end{aligned}$$

Our approach is to decompose $E\|\hat{\theta}^* - \theta\|^2$ into low, mid and high levels.

- Low levels ($j \le L$): trivial, since L is fixed.
- Mid levels ($L \le j \le j_0$): use global oracle inequality (27).
- High levels($j > j_0$) : use 'small signal' oracle inequality (28).

Hence, using the global oracle inequality (27) for mid levels:

$$\begin{aligned} S_{1\epsilon} \le \textstyle\sum_{j \le j_0} 2^j \sigma_j^2 \{\tilde{R}(\mu_j) + R_F(\mu_j) I\{\tau_j^2 \le 3\gamma_j\} + c 2^{j\eta - j/2}\} \\ = S_{11\epsilon} + S_{12\epsilon} + S_{13\epsilon}, \end{aligned} \tag{52}$$

and using the small-signal inequality (28) and the bound $d \log d \cdot e^{-c_* d \gamma_d^2} = O((\log d)^{-3/2})$ for high levels:

$$\begin{aligned} S_{2\epsilon} &\le \sum_{j > j_0} 2^j \sigma_j^2 R_F(\mu_j) + c\sigma_j^2 j^{-3/2} \\ &= S_{21\epsilon} + S_{22\epsilon}. \end{aligned}$$

Focus first on the main term, $S_{11\epsilon}$:

$$S_{11\epsilon} \le \sum_{j \le j_0} \inf_{t_j} E \| \hat{\theta}_{j,(t_j)} - \theta_j \|^2.$$

Hence, by maximising over $\Theta = b^{\sigma}_{p,q}(C)$,

$$\begin{aligned}\sup_{\Theta} S_{11\epsilon} &\leq \inf_{(t_j)} \sup_{\Theta} E\| \hat{\theta}_{(t_j)} - \theta \|^2 \\ &= R^*_T(\epsilon, \Theta) \asymp \epsilon^{2r}.\end{aligned}$$

The proof is completed by showing that we may choose j_0 so that all other terms are $o(\epsilon^{2r})$. We start with $S_{12\epsilon}$: let $c^2_{j\epsilon} = 3\gamma_j 2^j \sigma^2_j$ with $\gamma_j = 2^{-\sqrt{j}}$. Then if $\hat{\theta}^F_j$ denotes soft thresholding at $\sigma_j \sqrt{2 \log 2^j}$,

$$\begin{aligned}S_{12\epsilon} &\leq \sum_{j \leq j_0} E\|\hat{\theta}^F_j - \theta_j\|^2 I\{\|\theta_j\|^2 \leq c^2_{j\epsilon}\} \\ &\leq c \sum_{j \leq j_0} \{j^{-3/2}\sigma^2_j + j \sum_k \sigma^2_j \wedge \theta^2_{jk}\} I\{\|\theta_j\|^2 \leq c^2_{j\epsilon}\},\end{aligned}$$

where we have used the modified oracle inequality (50). To further bound $S_{12\epsilon}$, we borrow a definition from Theorem 3 of Johnstone and Silverman (1997):

$$\begin{aligned}W_p(\delta, C; n) &= \sup_{\|x\|_p \leq C} \sum_1^n \delta^2 \wedge x^2_k \\ &\leq \begin{cases} \min(n\delta^2, C^p \delta^{2-p}) & 0 \leq p \leq 2 \\ \min(n\delta^2, C^2 n^{1-2/p}) & 2 \leq p \leq \infty \end{cases},\end{aligned}$$

where, in both cases, the minimum is obtained at $n\delta^2$ if and only if $\delta \leq Cn^{-1/p}$. Defining also the Besov 'rings'

$$\Theta^{(j)} = \Theta \cap \{\theta : \theta_{j'k} = 0, \forall j' \neq j, \forall k\},$$

one checks easily that $\Theta^{(j)}$ is essentially isomorphic with the ℓ_p ball $\{\theta_j : \|\theta_j\|_p \leq C2^{-sj}\}$. Hence

$$\sup_{\Theta} S_{12\epsilon} \leq c \sum_{j \leq j_0} j^{-3/2}\sigma^2_j + c \sum_{j \leq j_0} j \min\{W_p(\sigma_j, C2^{-sj}; 2^j), W_2(\sigma_j, c_{j\epsilon}; 2^j)\}.$$

Write $W_{pj}(\epsilon), W_{2j}(\epsilon)$ as abbreviations for the preceding terms. We first analyse $W_{pj}(\epsilon)$. In the case $p \leq 2$,

$$W_p(\sigma_j, C2^{-sj}; 2^j) \leq \begin{cases} 2^j \sigma^2_j & \text{if } \sigma_j \leq C2^{-sj-j/p} \\ C^p 2^{-sjp} \sigma^{2-p}_j & \text{if } \sigma_j > C2^{-sj-j/p}. \end{cases} \tag{53}$$

The function $j \to 2^j \sigma^2_j = \epsilon^{2\alpha} 2^{j\alpha}$ grows exponentially in j while $j \to C^p 2^{-sjp} \sigma^{2-p}_j$ $= (C\epsilon^{-\alpha})^p \epsilon^{2\alpha} 2^{-jp[\sigma + \alpha(1/2 - 1/p)]}$ decreases exponentially in j. The functions cross at the switching value $j_* \in \mathcal{R}$ given by the solution to the equation $\sigma_j = C2^{-(s+1/p)}j$, namely

$$2^{j_*(\sigma+\alpha/2)} = C\epsilon^{-\alpha}. \tag{54}$$

The value at the crossing point j_* (which yields the maximum of $j \to W_{pj}(\epsilon)$) is then

$$2^{j_*}\sigma_{j_*}^2 = \epsilon^{2\alpha}2^{j_*\alpha} = C^{r/\sigma}\epsilon^{2r}$$

(recalling that $r = 2\sigma\alpha/(2\sigma+\alpha)$).

For the case $p > 2$, the situation is the same as in (53) except that the second term is now $C^2 2^{-2sj}2^{j(1-2/p)} = C^2 2^{-2j\sigma}$. Note, however, that the switching value j_* is still given by (54). In summary, we conclude that $j \to W_{pj}(\epsilon)$ decays geometrically from a maximum j_*:

$$W_{pj}(\epsilon) \le C^{r/\sigma}\epsilon^{2r} \cdot 2^{-\eta_1|j-j_*|}, \qquad \eta_1 > 0, \tag{55}$$

so long as

$$\sigma > \alpha(1/p - 1/2)_+. \tag{56}$$

For small j, however, the small signal constraint contained in $W_{2j}(\epsilon)$ leads to the better bound. Indeed, since $W_2(\delta, C; n) = n\delta^2 \wedge C^2$, it follows that

$$W_{2j}(\epsilon) = 2^j\sigma_j^2 \wedge c_{j\epsilon}^2 = c_{j\epsilon}^2 \le 32^{-\sqrt{j}}2^{j\alpha}\epsilon^{2\alpha}.$$

Recalling that $r = 2\sigma\alpha/(2\sigma+\alpha), \alpha - r = \alpha^2/(2\sigma+\alpha)$ and (54), we obtain

$$\epsilon^{-2r}W_{2j}(\epsilon) \le 3C^{2\alpha/(2\sigma+\alpha)}2^{-\sqrt{j}+(j-j_*)\alpha}. \tag{57}$$

Combining (55) and (57), and letting $c_i = c_i(\alpha, \sigma, C)$, we get

$$\begin{aligned}\epsilon^{-2r}\sup_{\Theta} S_{12\epsilon} &\le o(1) + c\sum_{j\le j_1} j\epsilon^{-2r}W_{2j}(\epsilon) + c\sum_{j_1}^{j_0} j\epsilon^{-2r}W_{pj}(\epsilon)\\ &\le o(1) + c_1\sum_{j<j_1} j2^{-\sqrt{j}+(j-j_*)\alpha} + c_2\sum_{j\ge j_1} j2^{-\eta_1(j-j_*)} = o(1)\end{aligned}$$

if we choose, for example, $j_1 = j_* + (\log_2 j_*)^2$.

Turning now to $S_{13\epsilon}$ we have

$$\epsilon^{-2r}S_{13\epsilon} = c\epsilon^{2(\alpha-r)}\sum_{j\le j_0} 2^{j(\alpha+\eta-1/2)}.$$

Since $(\alpha - r) = \alpha^2/(2\sigma+\alpha)$, this term is automatically $o(1)$ if $\alpha + \eta - 1/2 < 0$. On the other hand, if $\alpha + \eta - 1/2 > 0$, then

$$\epsilon^{-2r}S_{13\epsilon} \le c\epsilon^{2(\alpha-r)}2^{j_0(\alpha+\eta-1/2)}.$$

Writing $j_0 = bj_*$ and recalling (54), we find that the exponent of ϵ is positive when

$$b < \alpha/(\alpha + \eta - 1/2). \tag{58}$$

The analysis of terms in $S_{2\epsilon}$ is straightforward (and deferred to the Appendix) except for the constraint on j_0 imposed by the 'small-signal' requirement of Theorem 2, namely that the inequality $\tau^2 \le (1/3)\gamma_j$ be valid for all $\mu_j = \theta_j/\epsilon$ and $\theta_j \in \Theta^{(j)}$ for $j \ge j_0$. Since $\Theta^{(j)}$ is essentially equivalent to the ℓ_p ball $\{\theta_j : \|\theta_j\|_p \le C2^{-sj}\}$, the requirement is that

$$\sup\{\|\theta_j\|_2^2 \ : \ \|\theta_j\|_p \le C2^{-sj}\} \le (1/3)\gamma_j 2^j \sigma_j^2$$

for all $j \ge j_0$. Noting that $\sup\{\|\theta\|_{2,n}^2 \ : \ \|\theta\|_{p,n} \le r\} = n^{(1-2/p)_+} r^2$, and recalling that $2^j\sigma_j^2 = 2^{j\alpha}\epsilon^{2\alpha}$, the condition is that for all $j \ge j_0$,

$$C^2 2^{-2sj+(1-2/p)_+ j} \le (1/3)\gamma_j 2^{\alpha j}\epsilon^{2\alpha}. \tag{59}$$

Let $\bar\sigma = s - (1/2 - 1/p)_+ = \sigma - \Delta$, since $\bar\sigma$ equals σ if $p \ge 2$ and s if $p < 2$. Since $\gamma_j = 2^{-\sqrt{j}}$, condition (59) becomes

$$3C^2 2^{\sqrt{j}-(\alpha+2\bar\sigma)j} \le \epsilon^{2\alpha}.$$

If $\alpha + 2\bar\sigma > 0$, the function $j \to j(\alpha+2\bar\sigma) - \sqrt{j}$ is increasing for $j \ge j_+(\alpha,\bar\sigma) = [2(\alpha+2\bar\sigma)]^{-2}$. Since $j_0 = j_0(\epsilon) \uparrow$ as $\epsilon \to 0$, we get for sufficiently small ϵ that (59) will hold for all $j \ge j_0$ so long as

$$\alpha + 2\bar\sigma > 0 \tag{60}$$

$$3C^2 2^{\sqrt{j_0}-j_0(\alpha+2\bar\sigma)} \le \epsilon^{2\alpha}. \tag{61}$$

Writing $j_0 = bj_*$ where 2^{j_*} was defined at (54), it follows by comparing exponents of ϵ that (61) is met for all small ϵ so long as

$$b > (\alpha + 2\sigma)/(\alpha + 2\sigma - 2\Delta). \tag{62}$$

In summary, all terms other than $S_{11\epsilon}$ are $o(\epsilon^{2r})$ so long as (56) and (60) hold and $j_0 = bj_*$ can be chosen so that b satisfies both (58) and (62). Some algebra shows that these latter conditions amount to

$$\sigma > \alpha\Delta, \sigma > \Delta - \alpha/2, \quad \text{and} \qquad \sigma > 2\alpha\Delta - \alpha/2,$$

and this completes the proof of Theorem 1.

Remark. Without the pretest, the error term $2^{j\eta - j/2}$ resulting from the global oracle inequality (compare (52)) would have to be summed over all resolution levels, instead of simply up to level j_0. From the treatment of error term $S_{13\epsilon}$ above, it is apparent that if $\alpha \geq 1/2$ (for example), this sum diverges and so conclusions can be drawn using only the global oracle bound.

5. Extensions to a Class of Linear Inverse Problems

The purpose of this section is to sketch how the preceding results for threshold selection in correlated noise might be carried over to a class of linear inverse problems. The discussion will be informal and mostly by example. We imagine data observed indirectly in a model

$$y = Kf + z, \tag{63}$$

where K is a bounded linear operator of L_2, and $Cov(z) = I$. Specific examples that we have in mind include

1. *Integration*

$$Kf(u) = \int_{-\infty}^{u} f(t)dt.$$

2. *Fractional Integration*

$$Kf(u) = \int_{-\infty}^{\infty} \frac{f(t)\Omega(t-u)}{|t-u|^{1-\alpha}} dt, \qquad 0 < \alpha < 1.$$

(Here Ω is a homogeneous function of degree 0, and for example, $\alpha = 1/2$ corresponds to the Abel Transform)

3. *Certain Convolutions:*

$$\begin{aligned} Kf(u) &= \int_{-\infty}^{\infty} k(u-t)f(t)dt, \qquad \text{where} \\ \hat{k}(\omega) &\sim |\omega|^{-\alpha} \qquad \text{as } |\omega| \to \infty. \end{aligned}$$

Examples include

$$\begin{aligned} k(x) = e^{-|x|}I\{x < 0\} \qquad & (\Rightarrow \alpha = 1) \\ \text{or } \tfrac{1}{2}e^{-|x|} \qquad & (\Rightarrow \alpha = 2). \end{aligned}$$

The heuristic connection between correlated noise and linear inverse problems can be expressed almost trivially. Consider a correlated regression model $y = f + e$ where e has covariance operator Σ, and suppose that Σ has invertible non-negative square root L, so that $\Sigma = LL^*$ and $e \overset{\mathcal{D}}{=} Lz$. Formally writing $L^{-1}y = L^{-1}f + z$, we may then identify K in (63) with $\Sigma^{-1/2}$.

To exploit this connection, we use the notion of a *wavelet-vaguelette decomposition* (WVD) of K due to Donoho (1995). This is a modification of the singular value decomposition which aims to simultaneously almost diagonalize K *and* achieve sparse representations of functions f likely to be of interest.

We review some elements of the WVD here. Suppose that the function f we wish to recover has wavelet representation $f = \sum \langle f, \psi_{jk} \rangle \psi_{jk}$. However, the observed data is (a noisy version of) Kf, so we suppose in addition that it is possible to construct *representers* γ_{jk} such that

$$[Kf, \gamma_{jk}] = \langle f, \psi_{jk} \rangle.$$

Then for observed data $y_{jk} = [y, \gamma_{jk}]$, we have

$$Ey_{jk} = [Kf, \gamma_{jk}] = \langle f, \psi_{jk} \rangle,$$

which motivates use of a thresholding based reconstruction rule

$$\hat{f} = \sum_{jk} \eta(y_{jk}, \hat{t}_{jk}) \, \psi_{jk}.$$

The proposal here, of course, is to use a version of the unbiased risk estimate (SURE) to estimate $\hat{t}_j$ from the data y.

Two conditions will be necessary for asymptotic results on the validity of this thresholding proposal. First, the WVD structure itself requires fine scale homogeneity:

$$\|\gamma_{jk}\|_2 \sim 2^{j\gamma}, \qquad j \to \infty \tag{64}$$

uniformly in k. Secondly, we impose the conclusion

$$\text{if } f = 0, \qquad k \to y_{jk} \text{ is a stationary sequence.} \tag{65}$$

The proposed estimator essentially applies existing software to the data $(\{y_{jk}\})$:

1. Data:
$$w_{jk} = [Y, \gamma_{jk}].$$

2. Robust scale estimates: Fix L and
$$\text{for } j \geq L, \quad \hat{s}_j = MAD(y_{jk}, k = 1, \ldots, 2^j)/.6745$$

3. Hybrid-SURE thresholding:
$$\hat{w}_j^* = \begin{cases} \hat{s}_j \hat{\mu}^*(w_j/\hat{s}_j) & L \leq j < J \\ w_j & j < L. \end{cases}$$

4. Reconstruction:
$$\hat{f} = \sum_{\lambda} \hat{w_l}^* \psi_\lambda.$$

By Parseval's inequality

$$E\|\hat{f} - f\|^2 = \sum_j E|\hat{w}_j - w_j|^2,$$

where $\hat{w}_j = \{\hat{w}_{jk}, k = 1, \ldots, 2^j\}$.

To formulate an asymptotic result, consider a Gaussian white noise model for our linear inverse problem:

$$Y(du) = Kf(u)du + \epsilon W(du), \qquad u \in [0,1], \tag{66}$$

and convert it to sequence space form by integration against the collection of representers γ_λ:

$$[\gamma_\lambda, Y] = [\gamma_\lambda, Kf] + \epsilon[\gamma_\lambda, W], \qquad \lambda \in \Lambda,$$

so that, using an obvious notation,

$$y_\lambda = \theta_\lambda + \epsilon w_\lambda, \qquad \lambda \in \Lambda. \tag{67}$$

We proceed by analogy with the fractional Brownian motion model (10), but with a different noise normalization. Indeed, the observed noise components have covariance structure

$$Ew_\lambda \overline{w_{\lambda'}} = \int \gamma_\lambda \overline{\gamma_{\lambda'}} = \int \widehat{\gamma_\lambda} \overline{\widehat{\gamma_{\lambda'}}},$$

and in view of (64) and (65), we have at level j, $\sigma_j^2 = \mathrm{Var}\,(w_\lambda) \sim 2^{2j\gamma}\epsilon^2$. As an example, in the case of fractional integration, where $\widehat{\gamma_\lambda}(\omega) = \frac{|\omega|^\alpha}{\widehat{\Omega}(\omega)} \cdot \widehat{\psi_\lambda}(\omega)$, we find

$$Ew_\lambda \overline{w_{\lambda'}} = \int \widehat{\psi_\lambda}\, \overline{\widehat{\psi_{\lambda'}}}\, \frac{|\omega|^{2\alpha}}{|\widehat{\Omega}(\omega)|^2}\, d\omega.$$

The covariance structure of $k \to \{w_{jk}\}$ is thus analogous to that of the wavelet coefficients of fractional Brownian motion (compare (12)). In particular, in the examples listed earlier, Model S will apply, so long as one starts with a Meyer wavelet or other wavelet with sufficiently many vanishing moments (for example, for fractional integration, $g(\xi) \propto |\xi|^{2\alpha}/|\hat{\Omega}(\xi)|^2|$). Note however that the noise level here is parametrized by ϵ (and not ϵ^α), and that the levelwise variances $\sigma_j^2 \sim 2^{2j\gamma}\epsilon^2$ will, for the typical case in which $\gamma > 0$, grow with j (in contrast to the fractional Brownian motion case in (51)). In the construction of the hybrid estimator (25), it is necessary to use higher thresholds in the small signal case when $s_d^2 < \gamma_d$: here we replace $\sqrt{2\log d}$ by

$$t_d^F = \sqrt{2\beta \log d}, \tag{68}$$

where $\beta = 1 + 2\gamma_+$, and $\gamma_+ = \max(\gamma, 0)$. This phenomenon was originally noted by Abramovich and Silverman (1998). With this modification, Theorem 2 remains true at each resolution level j.

It is now possible to mimick the proof of Theorem 1 to obtain an asymptotic adaptive minimaxity result for the hybrid SURE estimator, simultaneously over a broad scale of Besov constraints.

Theorem 6. *Suppose that operator K in model* (66) *possesses a WVD satisfying* (64)*. Suppose that the sequence data* (67) *satisfies* (65)*. Let $\hat\theta^*$ be the SURE pretest estimator specified in* (11) *with $\sigma_j^2 = \epsilon^2 2^{2\gamma j}$ and $t_d^F = \sqrt{2\beta \log d}$ in* (25) *for $\beta = 2\gamma + 1$. Set $\alpha = 2\gamma + 1$ and $\Delta = (1/p - 1/2)_+$. Then for $1 \le p, q \le \infty, 0 < C < \infty$, and $\sigma > \max\{\alpha\Delta, \Delta - \alpha/2, 2\alpha\Delta - \alpha/2\}$,*

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

and, as shown by Donoho (1995)*,*

$$R_{T,\gamma;b^\sigma_{p,q}(C)}(n^{-1/2}) \asymp n^{-r(\gamma,\sigma)}, \qquad r(\gamma,\sigma) = 2\sigma/(2\sigma + 1 + 2\gamma).$$

Notice that the range of validity of this result is constrained to functions of greater smoothness as γ increases (in the sparse case, $\Delta > 0$.) This phenomenon is discussed further in Mallat (1998) who also shows how appropriate wavelet packet bases can be used to address the problem.

Acknowledgements

Thanks to Brenda MacGibbon for comments on the manuscript, and the hospitality of the Australian National University when the work was revised. Two referees also provided detailed comments that improved the revision. This work was supported in part by NSF grant DMS 9505151, NIH 1 R01 CA72028-02 and the Guggenheim Foundation.

Appendix

Proof of (26). Using the spectral representation of the process $\{z_i\}$, we can find i.i.d. $N(0,1)$ variates ζ_i and eigenvalues $\lambda_{i,d} \le f_\infty$ so that $\sum_1^d z_i^2 = \sum_1^d \lambda_{i,d}\zeta_i^2$. Setting $\alpha_i = d^{-1}\lambda_{i,d}$, we have

$$s_d^2 = \sum_1^d \alpha_i(\zeta_i^2 - 1).$$

Using the elementary identity

$$P\{|\sum \alpha_i(\zeta_i^2 - 1)| > t\} \le 2e^{2u^2 \sum \alpha_i^2 - |u|t}$$

for $|u| \leq 1/4(\max|\alpha_i|)$, and optimizing over u, we obtain (26).

Proof of Lemma 3

Graphs of the risk $r_S(t,\mu)$ of soft thresholding as a function of threshold t are essentially constant (at μ^2) for large t: the proof is essentially a (long-winded) formal verification.

1°. Risk as a function of threshold. Differentiation of the formula for the risk of soft thresholding (e.g. Donoho and Johnstone (1994), Formula (A2.1)) gives

$$r_t(t,\mu) = (\partial/\partial t)r(t,\mu) = 2t[\tilde{\Phi}(t-\mu) + \Phi(-t-\mu)] - 2[\phi(t-\mu) + \phi(t+\mu)].$$

2°. Monotonicity. For $\mu \geq 1$ and $t \geq 2$, $r(t,\mu)$ is an increasing function of t. Indeed, one checks that $\partial r_t/\partial\mu = \mu[\phi(t-\mu) + \phi(t+\mu)] \geq 0$, so that it will suffice to verify that $h(t) = (1/2)r_t(t,1) \geq 0$ for all $t \geq 2$. Further calculus shows that

$$h'(t) = \tilde{\Phi}(t-1) - \phi(t-1) + \tilde{\Phi}(t+1) - \phi(t+1) \leq 0 \qquad \text{for } t \geq 2.$$

Since $h(2) > 0$ and $h(t) \to 0$ as $t \to \infty$, the claim follows.

3°. We now verify that if $t \geq t_1$ and $\mu \geq 0$, then

$$r(t_1,\mu) - r(t,\mu) \leq 2\tilde{\Phi}(t_1-\mu) + 2t\phi(t). \tag{69}$$

The left side equals the expectation of

$$[\eta_S(x,t_1) - \mu]^2 - [\eta_S(x,t) - \mu]^2 \leq \begin{cases} (x-t_1)^2 & x > t_1 \\ 0 & |x| \leq t_1 \\ (x+t_1-\mu)^2 & -t < x < -t_1 \\ 2(x-\mu)(t_1-t) & x < -t, \end{cases}$$

as follows by checking cases. Changing variables to $z = x-\mu$ and integrating by parts gives

$$\int_{t_1}^{\infty}(x-t_1)^2\phi(x-\mu)dx = \tilde{\Phi}(t_1-\mu) - (t_1-\mu)^2\int_{t_1-\mu}^{\infty} z^{-2}\phi(z)dz \leq \tilde{\Phi}(t_1-\mu). \tag{70}$$

On the range $-t < x < -t_1$, setting $u = -(x-\mu)$ yields

$$\int_{-t}^{-t_1}(x+t_1-\mu)^2\phi(x-\mu)dx = \int_{t_1+\mu}^{t+\mu}(u-t_1)^2\phi(u)du \leq \tilde{\Phi}(t_1), \tag{71}$$

exploiting $\mu > 0$ and (70). Finally, on the range $x < -t$,

$$2(t_1-t)\int_{-\infty}^{-t}(x-\mu)\phi(x-\mu)dx = 2(t-t_1)\phi(t+\mu) \leq 2t\phi(t). \tag{72}$$

Combining the last three displays and noting that $\tilde{\Phi}(t_1) \le \tilde{\Phi}(t_1 - \mu)$ for $\mu \ge 0$ yields (69).

4°. We proceed to the proof of Lemma 3. Suppose that $r(t, F_d)$ attains its minimum on $[0, \infty]$ at t_0. Then

$$\begin{aligned} R(\mu) = \inf_t r(t, F_d) &= r(t_1, F_d) + r(t_0, F_d) - r(t_1, F_d) \\ &\ge \inf_{t \le t_1} r(t_1, F_d) + d^{-1} \sum [r(t_0, \mu_i) - r(t_1, \mu_i)] I\{t_0 > t_1\}. \end{aligned}$$

Step 2° implies that when $\mu_i \ge 1$, $r(t_0, \mu_i) \ge r(t_1, \mu_i)$. Hence, on rearranging the previous display and then substituting (69) and setting $t_1 = \sqrt{2 \log d}$, we get

$$\begin{aligned} \tilde{R}(\mu) - R(\mu) &\le d^{-1} \sum [r(t_1, \mu_i) - r(t_0, \mu_i)] I\{\mu_i \le 1, t_0 > t_1\} \\ &\le 2[\phi(t_1 - 1) + t_1 \phi(t_1)] \le 2\phi(t_1)(e^{t_1} + t_1) \\ &\le 4\phi(0) e^{-t_1^2/2 + t_1} \le 2d^{-1} e^{\sqrt{2 \log d}}. \end{aligned}$$

Proof of Theorem 2

Bound for $P(B_d)$. It follows from (35) that $\{\text{ave}_i |z_i| \le t_d\}$ implies B_d. On the other hand $(\text{ave}|z_i|)^2 \le s_d^2 + 1$, so

$$\begin{aligned} P(B_d^c) &\le P\{s_d^2 > t_d^2 - 1\} \\ &\le e^{-d(t_d^2 - 1)/8 f_\infty} = \epsilon_{2d}, \end{aligned}$$

where we have used the tail bound (26) for s_d^2. Clearly $\epsilon_{2d} << \epsilon_{1d}$, so we will ignore this term below.

Completion of bound for $E\|W_d\|$. Bound (30) calls for a crude bound for $E\|W_d\|_\infty^2$. Using

$$\begin{aligned} \|W_d\|_\infty &\le W_d(0) + t_d \|W_d^{'}\|_\infty \\ &\le W_d(0) + 4t_d^2 + 2t_d \text{ave } |z_i|, \end{aligned}$$

together with the observation that $W_d(0) \overset{\mathcal{D}}{=} d^{-1}(\chi_{(d)}^2 - d)$ and $(\text{ave } |z_i|)^2 \le s_d^2 + 1$, leads, after some calculation, to

$$E\|W_d\|_\infty^2 \le (8t_d^2)^2.$$

Substituting this and (47) into (30) (with $c = 2d^{\eta - 1/2}$) and letting $N_d = t_d/\delta \asymp 6t_d^2 d^{1/2 - \eta}$ denote the number of discretization points, we have

$$E\|W_d\|_\infty \le 2d^{\eta - 1/2} + 8t_d^2 \sqrt{N_d \epsilon_{1d}},$$

uniformly in $\mu \in \mathcal{R}$. The dominant term in $\sqrt{N_d \epsilon_{1d}}$ is

$$\sqrt{N_d \epsilon_{1d}} \approx M^{1/4} t_d d^{(1/2)[3/2-\eta-(2b\eta+\eta-1/2)\beta_\gamma]},$$

so some algebra shows that $E\|W_d\|_\infty = O(d^{\eta-1/2})$ so long as

$$b = \alpha + 2M_0 > [5/2 - 3\eta + (1/2-\eta)\beta_\gamma]/(2\eta\beta_\gamma). \tag{73}$$

Analysis of Z_d. By definition,

$$\begin{aligned} Z_d(t) &= d^{-1} \sum_{i=1}^{d} 1 - 2I\{x_i^2 \le t^2\} + x_i^2 \wedge t^2 - r(t,\mu_i) \\ &\overset{def}{=} d^{-1} \sum_i Y_i(t). \end{aligned} \tag{74}$$

The functions $t \to Y_i(t)$ are discontinuous, so we need to give a stochastic estimate of ΔZ_d in (31). If $t < t'$, we have, using (33), and $(t')^2 - t^2 \le 2t_d(t'-t)$,

$$|Y(t') - Y(t)| \le 2I\{t < |x| \le t'\} + 4t_d(t'-t). \tag{75}$$

Let $N_d(t,t') = \#\{i \ : \ t < |x_i| \le t'\}$. From (75),

$$\Delta Z_d(t_j, \delta) \le 2d^{-1} N_d(t_j, t_j + \delta) + 4\delta t_d.$$

Clearly, $Ed^{-1}N_d(t_j, t_j+\delta) \le 2\phi(0)\delta$ where $\phi(x)$ is the standard Gaussian density. Let

$$\begin{aligned} \partial N_d(t_j,\delta) &= d^{-1}\{N_d(t_j, t_j+\delta) - EN_d(t_{j},_{j}+\delta)\}, \qquad \text{and} \\ C_d &= \cap_{j\in J}\{|Z_d(t_i)| \le d^{\eta-1/2}\} \\ D_d &= \cap_{j\in J}\{\partial N_d(t_j,\delta)| \le (1/3)d^{\eta-1/2}\}. \end{aligned}$$

It follows from the above analysis and (31) that if we take $\delta = d^{-1/2}/t_d$, then for sufficiently large d,

$$C_d \cap D_d \Rightarrow \|Z_d\|_\infty \le 2d^{\eta-1/2}. \tag{76}$$

Bound for $E\|Z\|_\infty$. We first state the inequality from Bosq (1993) appropriate for bounded random variables.

Proposition 7. *Assume that there exist constants $m \le M$ such that*

$$0 < mp \le esssup|X_{i+1} + \cdots + X_{i+p}| \le Mp, \qquad i \in \mathcal{Z}, p \ge 1. \tag{77}$$

If $S_n = X_1 + \cdots + X_n$, and $1 \le p_n \le n/2$, then for every $\epsilon > 0$,

$$P(|S_n| > n\epsilon) \le 8 \ \exp(-\frac{\epsilon^2}{25M^2}\frac{n}{p_n}) + 18\frac{M}{\sqrt{m\epsilon}}\frac{n}{p_n}\alpha(p_n). \tag{78}$$

We follow the approach for $E\|W\|_\infty$, using now (78) to estimate $P(C_d^c)$ and $P(D_d^c)$. Consider first C_d^c, and note that (77) hold for $X_i = Y_i(t)$ (defined in (74)) with $m = 1$ and $M = 2 + t_d^2$. As before, we have the α-mixing bound $\alpha_X(p) \leq ap^{-b}$ with $b = \alpha + 2M_0$. We apply Proposition 7 with parameters $\epsilon = d^{\eta-1/2}, p_d = d^\eta$, and find, after simplification, that

$$P\{|Z_d(t)| \geq d^{\eta-1/2}\} \leq \epsilon_{3d},$$
$$\epsilon_{3d} = c_3 \exp\{-d^\eta/25(2+t_d^2)^2\} + c_4 d^{5/4-(b+3/2)\eta}.$$

Turning to D_d^c, we note that (77) holds for $X_i = I\{t_j < x_i < t_j + \delta\} - P\{t_j < x_i < t_j + \delta\}$ with $m = 1/4$ and $M = 1$. Again using Proposition 7, now with $\epsilon = d^{\eta-1/2}/3$ and $p_d = d^\eta$, we eventually obtain

$$P\{|\partial N_d(t_j, \delta)| \geq (1/3)d^{\eta-1/2}\} \leq \epsilon_{4d},$$
$$\epsilon_{4d} = c_5 \exp\{-c_6 d^\eta\} + c_7 d^{5/4-(b+3/2)\eta}.$$

The dominant term in both ϵ_{3d} and ϵ_{4d} is $O(d^{5/4-(b+3/2)\eta})$. Noting that $\delta^{-1}t_d = t_d^2 d^{1/2+\epsilon}$ and assembling pieces, we find that

$$P\{\|Z\|_\infty > 2d^{\eta-1/2}\} \leq c_8 t_d^2 d^{7/4+\epsilon-(b+3/2)\eta}.$$

Since $\|Z_d\|_\infty \leq 2 + t_d^2$, (30) yields

$$E\|Z_d\|_\infty \leq 2d^{\eta-1/2} + (2+t_d^2) \cdot P(\|Z_d\|_\infty > 2d^{\eta-1/2})^{1/2}.$$

Combining these last two displays, some algebra shows that $E\|Z_d\|_\infty = O(d^{\eta-1/2})$ so long as

$$b = \alpha + 2M_0 > (3/4 + \eta/2 + \epsilon)/\eta. \tag{79}$$

Combining the conclusions reached from (73) and (79), we obtain for M_0 chosen sufficiently large, the 'signal term' bound

$$R_{1d} - \tilde{R}(\mu) \leq cd^{\eta-1/2}.$$

Bound for 'noise' term R_{2d}. On the event A_d^c, fixed thresholding is used:

$$R_{2d} = d^{-1}E[\sum_i (\eta(x_i, t_d^F) - \mu_i)^2, A_d^c] \leq R_F(\mu). \tag{80}$$

It remains to show that $\tau^2 \geq 3\gamma_d$ forces $P(A_d^c)$ to be small enough that $R_{2d} = o(d^{-1/2})$. On event A_d^c

$$d^{-1}\sum_i \eta(x_i, t_d^F)^2 \leq d^{-1}\sum_i x_i^2 \leq 1 + \gamma_d,$$

and so

$$R_{2d} \le 2(1+\gamma_d+\tau^2)P(A_d^c). \tag{81}$$

When $\tau^2 \ge 3\gamma_d$, we may write

$$\begin{aligned} A_d^c &= \{d^{-1}\textstyle\sum(z_i^2-1)+d^{-1}\sum 2\mu_i z_i \le -(\tau^2-\gamma_d)\} && (82)\\ &\subset \{s_d^2 \le -\tau^2/3\} \cup \{V_d \triangleq d^{-1}\textstyle\sum 2\mu_i z_i \le -\tau^2/3\} && (83)\\ &\triangleq B_{\tau d} \cup C_{\tau d}. && (84)\end{aligned}$$

For event $B_{\tau d}$, use the large deviation inequality (26) to write

$$\begin{aligned}(1+\tau^2)P(B_{\tau d}) &\le 2(1+\tau^2)\exp\{-\frac{d\tau^2}{72f_\infty^2}\min(\tau^2,3f_\infty)\}\\ &= o(d^{-1/2}) && (85)\end{aligned}$$

since $d\gamma_d^2 >> \log d$.

For event $C_{\tau d}, V_{1d} = \sum \mu_i z_i$ is Gaussian with mean 0 and variance bounded by $f_\infty \sum \mu_i^2 = f_\infty \tau^2 d$. Consequently,

$$\begin{aligned}(1+\tau^2)P(C_{\tau d}) &\le (1+\tau^2)P\{(2/d)\tau\sqrt{f_\infty d}\ N(0,1) > \tau^2/3\} && (86)\\ &= o(d^{-1/2}) && (87)\end{aligned}$$

for $\tau^2 \ge 3\gamma_d$. Combining (85) and (86) with (81) shows that $R_{2d} = o(d^{-1/2})$ uniformly when $\tau^2 \ge 3\gamma_d$.

Proof of Theorem 2(b). We use the decomposition (29) as before, as well as the bound $R_{2d} \le R_F(\mu)$ from (80). By Cauchy-Schwarz, we have

$$dR_{1d} \le P(A_d)^{1/2}\textstyle\sum(E[\eta(x_i,\hat t_S)-\mu_i]^4)^{1/2}. \tag{88}$$

Using (41),

$$E[\eta(x_i,\hat t_S)-\mu_i]^4 \le 4E[z_i^2+t_d^2]^2 \le 16t_d^4. \tag{89}$$

We use a decomposition of A_d similar to (82), along with the bound $\tau^2 \le \gamma_d/3$, to write

$$A_d \subset \{s_d^2 > \gamma_d/3\} \cup \{V_d > \gamma_d/3\} \triangleq B_d \cup C_d.$$

Arguing as before, we obtain

$$\max\{P(B_d),P(C_d)\} \le 2e^{-c_8 d\gamma_d^2}$$

from which we conclude, in conjunction with (88) and (89), that $dR_{1d} = O(d\log d\cdot e^{-c_8 d\gamma_d^2}) = O(e^{-c_9 d\gamma_d^2})$.

Completion of proof of Theorem 1

We now bound the terms in $S_{2\epsilon}$. First

$$\begin{aligned} S_{22\epsilon} &\le c \sum_{j>j_0} \sigma_j^2 j^{-3/2} \\ &= c\epsilon^{2\alpha} \sum_{j>j_0} j^{-3/2} 2^{-j(1-\alpha)} \\ &\le c' \epsilon^{2\alpha}. \end{aligned}$$

Thus

$$\epsilon^{-2r} S_{22\epsilon} \le c' \epsilon^{2(\alpha-r)} = o(1)$$

since $\alpha - r = \alpha^2/(2\sigma+\alpha) > 0$.

Using the oracle inequality (50) and $p-$ maximal ideal risk bound $W_{pj}(\epsilon)$ defined above (53),

$$\begin{aligned} S_{21\epsilon} &\le c \sum_{j>j_0} E\|\hat{\theta}_j^F - \theta_j\|^2 \\ &\le c \sum_{j>j_0} j^{-3/2}\sigma_j^2 + j \sum_k \sigma_j^2 \wedge \theta_{jk}^2) \\ &\le c \sum_{j>j_0} j^{-3/2}\sigma_j^2 + c \sum_{j>j_0} j W_{pj}(\epsilon) \\ &= S'_{21\epsilon} + S''_{21\epsilon}. \end{aligned}$$

The first term is bounded exactly as for $S_{22\epsilon}$, while (55) shows that

$$\begin{aligned} \epsilon^{-2r} S''_{21\epsilon} &\le c \sum_{j>j_0} j C^{r/\sigma} 2^{-\eta_1(j-j_*)} \\ &\le c j_0 2^{-\eta_1(j_0-j_*)} = c j_0 2^{-\eta_1(b-1)j_*} = o(1), \end{aligned}$$

since (62) forces $b > 1$.

Remarks on extension to the sampling model

The first version of this paper projected that further research would show that results given here for the white noise model would carry over to the sampled data model (1). Since the referees have asked for details of this unattempted project, we describe here some *conjectures* as to how the argument might proceed. These conjectures are based on the results of Donoho and Johnstone (1999) (hereafter DJ-I) and a modified approach using orthogonal wavelets (such as Coiflets) described in Donoho and Johnstone (1998a) (DJ-II).

Suppose then that we have observations from model (1), with stationary Gaussian errors e_i with correlation function $r_k \sim Ak^{-\alpha}$ as $k \to \infty$ and $0 < \alpha \le 1$.

Write $\mathbf{y} = (y_i)$, $\mathbf{f} = (f(t_i))$ and $\mathbf{e} = (e_i)$. Take a (boundary corrected) discrete wavelet transform $\tilde{y} = W\mathbf{y}$ using filter coefficients corresponding to the Meyer wavelet (after appropriate truncation because of non-compact support), or using a filter of compact support with a sufficient number of vanishing moments for *both* scaling function and wavelet - the Coiflet family is a key example. Apply the estimator $\hat{\theta}^*$ of (12) to $\tilde{y}$. An estimator $\hat{f}^*(t_i)$ is then obtained by applying the inverse discrete wavelet transform.

Let ϕ and ψ be the scaling function and wavelet corresponding to the discrete transform. Given wavelet coefficients θ_I, let $f = f[\theta] = \sum \theta_I \psi_I$ be the associated function on $[0, 1]$ (with the same convention re scaling coefficients as in the introduction) and given f, let $\theta[f]$ denote the corresponding wavelet coefficients. Define the function space $\mathcal{F}^{\sigma}_{p,q}(C)$ as $\{f[\theta] : \theta \in b^{\sigma}_{p,q}(C)\}$, the definition being justified by the characterization of norms of Besov function spaces in terms of wavelet coefficients (cf. Meyer (1990)).

We conjecture that the result of Theorem 1 extends to the sampling model (1) and that

$$\sup_{f \in \mathcal{F}^{\sigma}_{p,q}(C)} R(\hat{f}^*, f) \leq R^*_{T,\alpha}(n^{-1/2}; b^{\sigma}_{p,q}(C))(1 + o(1)), \tag{90}$$

where $R(\hat{f}^*, f)$ denotes either $\|f[\hat{\theta}^*] - f\|_{L_2[0,1]}$ or $n^{-1}\sum[\hat{f}^*(t_i) - f(t_i)]^2$. We confine further discussion to the former, but the latter might be treated, for example using remarks in DJ-II. We emphasise that these estimators are obtained by treating sampled data (1) in a manner that is a) implementable in computer code, and b) directly analogous to the estimator for which Theorem 1 is proved.

There are two issues in extending the results of DJ-I,II to the current situation: firstly, modification of the results for a given $b^{\sigma}_{p,q}(C)$ from the Brownian noise to the fractional Brownian noise setting, and secondly the incorporation of adaptation over the parameters (σ, p, q, C) of Θ.

We describe how the approach of DJ-I,II is used to deal with the first issue. We employ the Parseval inequality and a decomposition

$$\|\hat{f}^*(\tilde{y}) - f\| = \|\hat{\theta}^*(\tilde{y}) - \theta\| \leq \|\hat{\theta}^*(\tilde{y}) - \tilde{\theta}\| + \|\tilde{\theta} - \theta\|. \tag{91}$$

Here $\tilde{\theta} = W\mathbf{f} = \tilde{\theta}(\theta)$, where the latter form emphasises the dependence on the wavelet coefficients $\theta[f]$. Temporarily, suppose that $\tilde{y}$ and $\tilde{\theta}$ are considered only for levels $j \leq j_0 = \gamma \log_2 n$, where $\gamma = \gamma(\alpha, p, \sigma) < 1$ is chosen so that the full difficulty of estimation over $b^{\sigma}_{p,q}(C)$ occurs at levels up to j_0. Then analogs of two key lemmas of DJ-I can be employed:

$$\sup_{\theta \in b^{\sigma}_{p,q}(C)} \|\tilde{\theta} - \theta\|_2^2 = o(n^{-r}), \qquad r = 2\sigma\alpha/(2\sigma + \alpha), \tag{92}$$

$$\sup_{\theta \in b^{\sigma}_{p,q}(C)} \|\tilde{\theta}(\theta)\|_{b^{\sigma}_{p,q}(C)} \leq (1 + \Delta_n)C, \qquad \Delta_n \to 0. \tag{93}$$

These lemmas crucially require the vanishing moments assumption on the scaling function.

Now (92) is used to show that the second term in (91) is negligible with respect to $n^{-r.}$ For the first term, one uses (93) to write, setting $C_n = (1+\Delta_n)C$,

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

Here y denotes data from model (10), and the factor $1+\Delta_{2n}$ is a bound to allow for the fact that the covariance matrix $Cov(\tilde{y}) = Cov(W\mathbf{e})$ is asymptotically of the form of the covariance of the noise in (10), at least for levels $j \leq \gamma \log_2 n$.

Finally, to handle the adaptation across (σ, p, q, C), it seems possible that one could proceed as in the proof of Theorem 1, using the value $j_0(n)$ constructed there to apply the bounds derived from (92) and (93).

References

Abramovich, F. and Silverman, B. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**, 115-130.

Barton, R. J. and Poor, H. V. (1988). Signal detection in fractional gaussian noise. *IEEE Trans. Inform. Theory* **34**, 943-959.

Bosq, D. (1993). Bernstein-type large deviations inequalities for partial sums of strong mixing processes. *Statistics* **24**, 59-70.

Bradley, R. C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics: A Survey of Recent Results* (Edited by E. Eberlein and M. Taqqu), 165-192. Birkhauser, Boston, Massachusetts.

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Number 61 *in* CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia.

Donoho, D. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Applied Computational and Harmonic Analysis* **2**, 101-126.

Donoho, D. and Johnstone, I. (1998a). Lecture notes on wavelets and function estimation. Unpublished manuscript.

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

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90**, 1200-1224.

Donoho, D. L. and Johnstone, I. M. (1999). Asymptotic minimaxity of wavelet estimators with sampled data, Statist. Sinica **9**, 1-32.

Donoho, D. L. and Johnstone, I. M. (1998b). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**, 879-921.

Frazier, M., Jawerth, B. and Weiss, G. (1991). *Littlewood-Paley Theory and the Study of Function Spaces*, NSF-CBMS Regional Conf. Ser in Mathematics **79**, American Mathematical Society, Providence, Rhode Island.

Ibragimov, I. A. and Rozanov, I. A. (1978). *Gaussian Random Processes.* Springer Verlag, New York.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc. Ser. B* **59**, 319-351.

Mallat, S. (1998). *A Wavelet Tour of Signal Processing.* Academic Press, New York.

Meyer, Y. (1990). *Ondelettes et Opérateurs, I: Ondelettes, II: Opérateurs de Calderón-Zygmund, III:* (with R. Coifman)*, Opérateurs multilinéaires*, Hermann, Paris. English translation of first volume is published by Cambridge University Press.

Neumann, M. H. and von Sachs, R. (1995). Wavelet thresholding: Beyond the Gaussian i.i.d. situation. In *Wavelets and Statistics* Volum 103. (Edited by A. Antoniadis). Springer Verlag Lecture Notes.

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

Department of Statistics, Stanford University, Stanford CA 94305, U.S.A.
E-mail: imj@stat.stanford.edu

(Received November 1997; accepted September 1998)