Wavelet threshold estimators for data with correlated noise

Iain M. Johnstone
Department of Statistics
Stanford University
Stanford CA 94305
U.S.A.

Bernard W. Silverman
School of Mathematics
Bristol University
Bristol BS8 1TW
U. K.

December 21, 1994

Abstract

Wavelet threshold estimators for data with stationary correlated noise are constructed by the following prescription. First, form the discrete wavelet transform of the data points. Next, apply a *level-dependent* soft threshold to the individual coefficients, allowing the thresholds to depend on the level in the wavelet transform. Finally, transform back to obtain the estimate in the original domain. The threshold used at level j is $s_j\sqrt{2\log n}$, where s_j is the standard deviation of the coefficients at that level, and n is the overall sample size. The minimax properties of the estimators are investigated by considering a general problem in multivariate normal decision theory, concerned with the estimation of the mean vector of a general multivariate normal distribution subject to squared error loss. An ideal risk is obtained by the use of an 'oracle' that provides the optimum diagonal projection estimate. This 'benchmark' risk can be considered in its own right as a measure of the sparseness of the signal relative to the noise process, and in the wavelet context it can be considered as the risk obtained by ideal spatial adaptivity. It is shown that the level-dependent threshold estimator performs well relative to the benchmark risk, and that its minimax behaviour cannot be improved upon in order of magnitude by any other estimator.

Key Words and Phrases: Decision theory, Level-dependent thresholding, Minimax estimation, Nonlinear estimators, Nonparametric regression, Oracle inequality, Wavelet transform.

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 some noise process. In the case where (e_i) is white noise, Donoho and Johnstone (1994) showed that an estimator based on thresholded wavelet expansions has very desirable minimax properties. Their theory concentrates entirely on the white noise case; in this paper we extend their results to deal with e_i that have more general structure.

The structure of this paper is as follows. First of all, we briefly review the relevant aspects of the Donoho–Johnstone paper and define key notation. We then consider the case where e_i are drawn from some stationary process, and define a wavelet threshold estimator for this case. The key practical difference from the white noise case is that the thresholds are chosen separately for each level in the wavelet transform. The mild conditions we impose on the covariance structure of the noise process are then set out and discussed.

The key results of the paper are obtained by considering a general problem in multivariate normal decision theory, concerned with the estimation of an n-vector θ of parameters from a vector observation $Y \sim N(\theta, V)$. In order to provide a benchmark risk, we consider idealised diagonal projection estimators, where individual components of Y are 'kept' or 'killed' according to instructions from an 'oracle'. This benchmark risk can be considered in its own right as a measure of the 'sparsity' of the signal relative to the noise variance. Estimates constructed using a soft thresholding approach, corresponding to that applied to the wavelet coefficients in the wavelet threshold estimator, are shown to have risk at most $O(\log n)$ times the benchmark. By considering a suitable prior for θ, it is then shown that this order of magnitude of behaviour cannot be improved by *any* estimator based on the given data Y.

In the original function estimation context, the results demonstrate that wavelet threshold estimates with level-dependent thresholds have good theoretical properties. Many spatially inhomogeneous functions have economical wavelet expansions as judged by the benchmark risk, and for these the asymptotic performance of the estimator is excellent.

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).

2 Wavelet estimators for white noise

2.1 Basic definitions and notation

To construct the Donoho–Johnstone estimator, let $\mathcal{W}$ be a periodic discrete wavelet transform operator, and let Y be the n-vector of observations $Y_1, \ldots, Y_n$. Write

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

with the remaining element labeled w_{00}. Let $\theta = \mathcal{W}\mathbf{f}$ be the wavelet transform of the signal $\mathbf{f} = \left(f\{(i-1)/n\}\right)_{i=1}^{2^m}$, and $\mathbf{z} = \mathcal{W}\mathbf{e}$ be the wavelet transform of the noise. Suppose that the (e_i) are independent identically distributed $N(0, \sigma^2)$ random variables, for some known σ^2.

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

$$\eta_s(w, \lambda) = \operatorname{sgn}(w)\left(|w| - \lambda\right)_+ . \tag{2}$$

The Donoho–Johnstone estimator is constructed by soft thresholding the wavelet coefficients w_{jk} at threshold $\lambda_n = \sigma\sqrt{2\log n}$, and then transforming back. Thus we define $\hat{\theta}$ by

$$\hat{\theta}_{jk} = \eta_s(w_{jk}, \lambda_n) \tag{3}$$

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

$$\hat{\mathbf{f}} = \mathcal{W}^T \hat{\theta}. \tag{4}$$

In practice the transformations $\mathcal{W}$ and the inverse transform $\mathcal{W}^T$ are carried out by an extremely fast algorithm. For more details see, for example, Mallat (1989) or Nason and Silverman (1994).

In most of our subsequent discussion, we shall 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.

2.2 Minimax properties

Donoho and Johnstone (1994) showed that the estimator $\hat{f}$ as defined in (3) and (4) has attractive spatial adaptivity properties. These were demonstrated in a minimax framework that we now set out.

The theory is worked out in the context of a problem from multivariate normal decision theory. Suppose observations $X_i, i = 1, \ldots, n$ satisfy

$$X_i = \theta_i + z_i$$

where the z_i are independent $N(0, \sigma^2)$ random variables, for some known σ^2, and 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 the empirical wavelet coefficients w_{jk}.

It is of particular interest to consider situations in which the θ vector contains only a few elements substantially different from zero. This is because a wide variety of 'regular' functions have wavelet transforms of this kind, not only those that are smooth in conventional senses (for example 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. With this in mind, we could consider the effect of having an *oracle* that told us which of the θ_i were near zero, and consider the best estimators that could be obtained under such conditions. Of course such estimators cannot be used in practice, but one might hope that they were mimicked well by adaptive estimators.

We shall in fact consider oracles among all *diagonal projection* estimators of the form $\hat{\theta}_i = \delta_i X_i$, $\delta_i = 0$ or 1, so that the oracle has the option of telling us whether to 'keep' or 'kill' coordinate i. We let $R(DP, \theta)$ be the risk obtained for the ideal choice of sequence $\{\delta_i\}$. Under the assumption that the errors z_i are independent $N(0, \sigma^2)$ random variables, the ideal diagonal projection estimator is obtained by setting

$$\delta_i = I[|\theta_i| > \sigma]. \tag{5}$$

If the oracle told us which θ_i were numerically larger than the noise standard deviation σ we would be able to use the values given in (5) and attain the risk

$$R(DP, \theta) = \sum_i (\theta_i^2 \wedge \sigma^2). \tag{6}$$

Of course, we cannot actually use this diagonal projection estimator, but the risk (6) can be used to construct a benchmark against which to judge the behaviour of estimates that can actually be realised in practice. For technical reasons, the benchmark risk will be $\sigma^2 + \sum_i (\theta_i^2 \wedge \sigma^2)$; the additional σ^2 is the mean square error in estimating a single parameter unbiasedly, and for all but the sparsest signals will be small compared with $R(DP, \theta)$.

As in (3), set the threshold λ_n to $\sigma\sqrt{2 \log n}$ and define

$$\hat{\theta}_i = \eta_s(X_i, \lambda_n).$$

Donoho and Johnstone (1994, Theorem 1) show that this estimator satisfies

$$E\|\hat{\theta} - \theta\|^2 \leq (1 + 2 \log n)\{\sigma^2 + \sum_i (\theta_i^2 \wedge \sigma^2)\} \text{ for all } \theta \in I\!R^n. \tag{7}$$

The result (7) shows that, for all possible θ, the estimator $\hat{\theta}$ comes within a factor $(1 + 2 \log n)$ of achieving the benchmark risk $\sigma^2 + R(DP, \theta)$.

Theorem 2 of Donoho and Johnstone (1994) shows that this behaviour cannot essentially be improved. They show that

$$\frac{1}{2 \log n} \inf_{\tilde{\theta}} \sup_{\theta \in I\!R^n} \frac{E\|\tilde{\theta} - \theta\|^2}{\sigma^2 + \sum_{i=1}^n (\theta_i^2 \wedge \sigma^2)} \to 1 \tag{8}$$

as $n \to \infty$, where for each n the infimum is taken over all estimators $\tilde{\theta}$ that depend on $X_1, \ldots, X_n$. The threshold estimator $\hat{\theta}$ thus asymptotically attains the best possible behaviour of any estimator relative to the benchmark risk.

Finally we note that although the benchmark risk has been derived by reference to an ideal diagonal projection estimator, it can equally be considered in its own right as a measure of the sparsity of the wavelet representation of a signal. For example, suppose 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 order $\log n$ as $n \to \infty$ and the benchmark risk therefore of 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 the results (7) and (8) 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 optimal wavelet diagonal projection estimator. The risk compares extremely favourably with the corresponding rate $O(n^{-1/2})$ for nonadaptive linear methods.

Many other ramifications of these results are discussed in Donoho and Johnstone (1994).

3 Wavelet estimators for coloured noise

3.1 Estimation by level-dependent thresholding

Suppose, now, that we have observations $Y_1, \ldots, Y_n$, with $n = 2^m$, satisfying

$$Y_i = f\{(i-1)/n\} + e_i, \qquad i = 1, \ldots, 2^m.$$

Suppose that the errors e_i have a multivariate normal distribution with mean zero and covariance matrix Γ_n. We shall assume that the errors are stationary so that Γ_n has entries $\gamma^{(n)}_{|s-t|}$, say. For simplicity we shall also initially assume periodicity, so that Γ_n is a circulant matrix. Some remarks about relaxing this assumption will be made in Section 6.2 below.

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

$$V_n = \mathcal{W} \Gamma_n \mathcal{W}^T. \tag{9}$$

Because of the stationarity assumptions, the variance of z_{jk} will depend only on the level j. In general the errors at different levels will be correlated to some extent, but within each level the distribution of the errors will be stationary. Write $s_j^2 = \operatorname{var} z_{jk}$ for each j.

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; we shall apply soft thresholding at threshold λ_j to the coefficients at level j, and define $\hat{\theta}$ to be the estimator

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

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 are a considerable number of coefficients at each level and the signal θ_{jk} can be assumed to be sparse, the variance at each level can be estimated from the data, and the threshold can be chosen as a suitable multiple of the variance at that level.

We shall consider, in this paper, thresholding rules of the form

$$\lambda_j = s_j\sqrt{2\log n}, \tag{10}$$

and assume that the variances s_j^2 are known. We write $\hat{\theta}$ for the corresponding estimator of θ, and set

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

The main aim in the remainder of the paper will be to obtain results analogous to (7) and (8) for this more general case.

3.2 Conditions for minimax results

The theoretical results for the level-dependent estimators defined in Section 3.1 above will depend on more general multivariate normal results set out in Sections 4 and 5 below. In this subsection we discuss the conditions that will allow these results to apply to the present context.

In order to provide a framework for our asymptotic results, we consider a sequence of problems wherein the function f remains fixed, but the number of observation points on the interval $[0, 1]$ increases as $n \to \infty$. The noise process will also be allowed to depend on n. It will be assumed throughout that n is a power of 2.

We shall make two assumptions on the covariances Γ_n. Firstly, we assume that the sum of squares of the autocorrelations of $\mathbf{e}$ remains bounded as $n \to \infty$, so that, for some constant c_1,

$$(\gamma_0^{(n)})^{-2}\sum_{j=0}^{n-1}(\gamma_j^{(n)})^2 \leq c_1 < \infty \qquad \text{for all } n. \tag{11}$$

We shall also need a condition on the reciprocal of the Fourier transform of the autocorrelations. For each n, let $\tilde{\gamma}^{(n)}$ be the discrete Fourier transform of the vector $\gamma^{(n)}$, so that the $\tilde{\gamma}^{(n)}$ are the eigenvalues of the covariance matrix Γ_n. Define

$$k_n = n^{-1}\gamma_0^{(n)}\sum_{j=0}^{n-1}(\tilde{\gamma}_j^{(n)})^{-1}. \tag{12}$$

We shall assume that, for some constant c_2,

$$k_n \leq c_2 < \infty \qquad \text{for all } n. \tag{13}$$

By standard time series properties, we will have $k_n \geq 1$, with equality only if the noise is uncorrelated.

If the noise can be considered as being drawn approximately from an infinite sequence with fixed spectral density $g(\omega)$, then both conditions (11) and (13) can be expressed in terms of $g(\omega)$. (The approximation arises because of the periodicity assumptions we are making on the autocovariances.) Condition (11) corresponds approximately to

$$\int_0^{\pi/2} g(\omega)^2 d\omega < \infty$$

while condition (13) is elucidated by the approximation

$$k_n \approx \frac{2}{\pi}\gamma_0 \int_0^{\pi/2} \frac{1}{g(\omega)}\, d\omega.$$

It can be seen that conditions (11) and (13) will then be satisfied for a very wide range of noise processes, though certain forms of long-range dependence are not included. Of course, the conditions also allow for a more general sequence of noise models, where the dependence between the observations does not remain essentially the same as n increases. Wang (1994) discusses some aspects of wavelet methods for processes with long-range dependence.

4 Mimicking oracles by threshold estimators

4.1 A general multivariate problem

In this section we set out a general problem in multivariate normal decision theory with dependent noise, extending the general white noise problem described in Section 2.2 above.

Suppose we have a vector X of observations $X_1, \ldots, X_n$ satisfying

$$X \sim N(\theta, V) \tag{14}$$

where θ is a vector of parameters to be estimated and V is a covariance matrix. Define the risk measure $R(\hat{\theta}, \theta) = E\|\hat{\theta} - \theta\|^2$, where the norm is the usual Euclidean norm. Define $\sigma_i^2 = \sigma_{ii} = \mathrm{var}(X_i)$ and

$$\overline{\sigma^2} = n^{-1} \sum_1^n \sigma_i^2 = n^{-1} \mathrm{tr}\, V. \tag{15}$$

Consider the effect in this problem of having an oracle among diagonal projection estimators, and let $R(DP, \theta)$ be the risk obtained for the ideal choice of sequence

$\{\delta_i\}$. Then

$$\begin{aligned} R(DP,\theta) &= \min_{\{\delta_i\}} \sum_i E(\delta_i X_i - \theta_i)^2 = \sum_i \min\left\{\theta_i^2, E(X_i-\theta_i)^2\right\} \\ &= \sum_i (\theta_i^2 \wedge \sigma_i^2) \qquad (16) \end{aligned}$$

where $\delta_i = I[\theta_i^2 \geq E(X_i-\theta_i)^2] = I[\theta_i^2 \geq \sigma_i^2]$. Just as before, we shall use the quantity $R(DP,\theta)$ to construct a benchmark risk; in the more general case this risk will be $\overline{\sigma^2} + R(DP,\theta)$, which, by extension of the arguments given in Section 2.2, can again be considered as a measure of the sparseness of θ relative to the noise variances $\{\sigma_i^2\}$ in its own right.

4.2 Risk of adaptive estimators

We now consider thresholded adaptive estimators of θ. Defining the soft threshold function η_s as in (2) above, consider the estimator

$$\hat{\theta}_i = \eta(X_i,\, \sigma_i\sqrt{2\log n}).$$

The following theorem shows that, apart from the additive quantity $\overline{\sigma^2}$, which for all but the sparsest signals will be small compared with $\sum(\theta_i^2 \wedge \sigma_i^2)$, the estimator $\hat{\theta}$ comes within a logarithmic factor of mimicking the oracle by achieving the ideal diagonal projection risk.

Theorem 1 *The estimator $\hat{\theta}$ satisfies, for* all $\theta \in \mathbb{R}^n$,

$$R\left(\hat{\theta},\theta\right) \leq (2\log n + 1)\{\overline{\sigma^2} + \sum(\theta_i^2 \wedge \sigma_i^2)\}.$$

Proof. This follows the proof of the corresponding theorem in Donoho and Johnstone (1994). The method given there shows that, for each i, for all $\delta \leq \frac{1}{2}$ and with $t = \sqrt{2\log\delta^{-1}}$ that

$$E\left\{\eta(X_i,\sigma_i t) - \theta_i\right\}^2 \leq (2\log\delta^{-1} + 1)\{\delta\sigma_i^2 + (\theta_i^2 \wedge \sigma_i^2)\}.$$

Setting $\delta = n^{-1}$ and summing over n, we have

$$E\|\hat{\theta} - \theta\|^2 \leq (2\log n + 1)\{n^{-1}\sum\sigma_i^2 + \sum(\theta_i^2 \wedge \sigma_i^2)\}$$

as required. ∎

5 A lower bound on the estimation risk

This section contains the main technical part of the paper. We show, under suitable assumptions, that 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.

5.1 The 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 $\tau_i^2 \leq \sigma_i^2$ for all i, with equality when the X_i are independent. Set

$$\overline{\tau^2} = n^{-1} \sum_1^n \tau_i^2. \tag{17}$$

Our asymptotic 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, τ_i, $\overline{\sigma^2}$ and $\overline{\tau^2}$ will not be made explicit in the notation. The theorem will be proved under the assumption that there exist finite constants C_1 and C_2 such that, for all n,

$$\frac{n^{-1} \sum_{i=1}^n \sigma_i^4}{\left(n^{-1} \sum_{i=1}^n \sigma_i^2\right)^2} \leq C_1 \tag{18}$$

and

$$\frac{\overline{\sigma^2}}{\overline{\tau^2}} \leq C_2. \tag{19}$$

Theorem 2 *Assume the n-vector of observations X has a $N(\theta, V_n)$ distribution and that V_n is such that assumptions (18) and (19) 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*

$$\liminf_{n\to\infty} \frac{1}{2\log n} \frac{\overline{\sigma^2}}{\overline{\tau^2}} \inf_{\tilde{\theta}\in\tilde{\Theta}_n} \sup_{\theta\in\mathbb{R}^n} \frac{E\|\tilde{\theta} - \theta\|^2}{\overline{\sigma^2} + \sum_{i=1}^n (\theta_i^2 \wedge \sigma_i^2)} \geq 1. \tag{20}$$

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 at most a constant multiple of loss is incurred by restricting attention to pointwise threshold estimators with thresholds $\sigma_i\sqrt{2\log n}$.

As in Donoho and Johnstone, the key idea of the proof is to bound the minimax risk in (20) 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 setup.

5.2 Specifying a suitable prior

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

$$p(\theta) = 1 + (\overline{\sigma^2})^{-1} \sum_{i=1}^n (\theta_i^2 \wedge \sigma_i^2). \tag{21}$$

Define the modified loss function

$$\tilde{L}_n(\hat{\theta}, \theta) = (\overline{\sigma^2})^{-1} p(\theta)^{-1} \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2 \tag{22}$$

so that Theorem 1 can be written

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

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, 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 $F[\varepsilon, \mu]$ be the three-point distribution that places mass $\frac{1}{2}\varepsilon$ 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 (22). Write $\hat{\theta}^b$ for the Bayes estimator $E(\theta|X)$ relative to squared error loss, and $\tilde{\theta}^b$ for the Bayes estimator relative to the loss function $\tilde{L}_n$. Let the respective Bayes risks be

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

and

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

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

5.3 Some key lemmas

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

Lemma 1 *Suppose 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*

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

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,

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

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 $\varepsilon > 0$, Donoho and Johnstone (1994) show that the Bayes risk $E\{\mathrm{var}(\theta_i|\xi_i)\}$ for this problem satisfies

$$E\{\mathrm{var}(\theta_i|\xi_i)\} \geq (1-\varepsilon)\mu_n^2 \tau_i^2 (n^{-1}\log n)\Phi(a) \tag{23}$$

for all sufficiently large n, independently of τ_i^2. (They actually show 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 $(n^{-1}\log n)\mu_n^2\Phi(a)$ 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

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

applying (23). 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 (21) above, the Bayes rule for π_n relative to the modified loss $\tilde L_n$ is given by*

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

and satisfies

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

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

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

as required, proving the first part of the lemma.

For every $\theta \in \text{supp}\ \pi_n$, we then have

$$\begin{aligned}
\|\theta\|^2 &= \sum_{i=1}^n \mu_n^2 \tau_i^2 I[\theta_i \neq 0] = \mu_n^2 \sum_{i=1}^n \left(\sigma_i^2 \wedge \tau_i^2\right) I[\theta_i \neq 0] \\
&\leq \mu_n^2 \sum_{i=1}^n \{\sigma_i^2 \wedge (\mu_n^2 \tau_i^2)\} I[\theta_i \neq 0] \quad = \quad \mu_n^2 \sum_{i=1}^n \left(\sigma_i^2 \wedge \theta_i^2\right) \\
&= \mu_n^2\, \overline{\sigma^2}\{p(\theta) - 1\} < \mu_n^2\, \overline{\sigma^2}\, p(\theta), \qquad (24)
\end{aligned}$$

Applying Cauchy–Schwarz and then inequality (24), we obtain

$$\begin{aligned}
\|\tilde{\theta}^b\|^2 &= \|E^X\{\theta p(\theta)^{-1}\}\|^2/\{E^X p(\theta)^{-1}\}^2 \\
&\leq E^X\{\|\theta\|^2 p(\theta)^{-1}\}/E^X p(\theta)^{-1} < \mu_n^2\, \overline{\sigma^2}\, \{E^X p(\theta)^{-1}\}^{-1}.
\end{aligned}$$

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 C_1 be the constant in condition (18) above,*

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

and

$$\text{var}\, p(\theta) < C_1 \log n.$$

Let A_n be the event $\{p(\theta) \leq \log n + (\log n)^{2/3} + 1\}$. Then

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

Proof. We have

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

where the indicator variables are independent Bernoulli $(n^{-1} \log n)$ random variables. Hence

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

and

$$\text{var}\, p(\theta) \leq (\overline{\sigma^2})^{-2} n^{-1} \sum \sigma_i^4 \log n < C_1 \log n$$

applying condition (18) above. The last part of the lemma follows by 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 $\tilde{\theta}^b$ satisfies*

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

as $n \to \infty$.

Proof. Consider the term $E\left(\|\tilde{\theta}^b\|^2 I[\theta \in A_n^c]\right)$. By Lemma 2 and Cauchy-Schwarz we have

$$\begin{aligned} E\left(\|\tilde{\theta}^b\|^2 I[\theta \in A_n^c]\right) &\leq \mu_n^2 \, \overline{\sigma^2} \, E\Big[E\{p(\theta)|X\} I[A_n^c]\Big] \\ &\leq \mu_n^2 \, \overline{\sigma^2} \{E \, p(\theta)^2\}^{1/2} P(A_n^c)^{1/2} \quad = \quad o(\mu_n^2 \, \overline{\sigma^2} \log n). \end{aligned}$$

On the other hand, we have

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

and

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

so that

$$\begin{aligned} E\Big(\|\theta\|^2 I[\theta \in A_n^c]\Big) &\leq (E\|\theta\|^4)^{1/2} P(A_n^c)^{1/2} \\ = \quad O(\mu_n^2 \, \overline{\sigma^2} \log n) P(A_n^c)^{1/2} &= o(\mu_n^2 \, \overline{\sigma^2} \log n), \end{aligned}$$

completing the proof of the lemma. ∎

5.4 Completion of the proof

We are now in a position to complete the proof of Theorem 2. The Bayes risk for π_n with respect to the loss $\tilde{L}_n$ satisfies

$$\begin{aligned} \overline{\sigma^2} \, \tilde{\rho}(\pi_n) &= E_{\pi_n} E_\theta \frac{\|\tilde{\theta}^b - \theta\|^2}{p(\theta)} \\ &\geq \frac{E_{\pi_n} E_\theta (\|\tilde{\theta}^b - \theta\|^2 I[A_n])}{1 + \log n + (\log n)^{2/3}} \\ &= \frac{E_{\pi_n} E_\theta \|\tilde{\theta}^b - \theta\|^2}{1 + \log n + (\log n)^{2/3}} - o(\mu_n^2 \, \overline{\tau^2}\,), \end{aligned} \tag{25}$$

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

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

Substituting into (25) it follows that, as $n \to \infty$,

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

hence by the minimax theorem of decision theory

$$\begin{aligned} \inf_{\tilde{\theta}} \sup_{\theta} \frac{\overline{\sigma^2}}{\tau^2} E_\theta \tilde{L}_n(\tilde{\theta}, \theta) &\geq \frac{\overline{\sigma^2}}{\tau^2}\, \tilde{\rho}_n(\pi_n) \\ &\geq (2 \log n)\Phi(a)\{1 + o(1)\}. \end{aligned} \tag{26}$$

Since (26) is true for all a, it follows that

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

completing the proof. ∎

6 Results for wavelet transforms

6.1 Invariance under orthogonal transformations

Both the conditions for the 'lower bound' Theorem 2, and the conclusion of the theorem itself, depend on properties of the covariance matrix V_n. In the context of wavelet threshold estimators, the theorem will be applied to a vector that has been obtained by wavelet transformation, and the matrix V_n will be an orthogonal transform (9) of the covariance matrix Γ_n of the original observations. Therefore it is of obvious importance to obtain results that depend on properties of V_n that are invariant under orthogonal transformations, such as the discrete wavelet transform and its inverse.

The quantity $\overline{\sigma^2} = n^{-1}\mathrm{tr}\, V_n$ depends only on the eigenvalues of V_n and hence is invariant in this way. 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

$$n^{-1} \sum_{i=1}^{n} \sigma_i^4 = n^{-1} \sum_{i=1}^{n} \sigma_{ii}^2 \leq n^{-1}\mathrm{tr}\, V_n^2,$$

so that condition (18) is implied by the condition

$$(n^{-1}\mathrm{tr}\, V_n^2)/(n^{-1}\mathrm{tr}\, V_n)^2 \leq c_1 < \infty. \tag{27}$$

This condition is invariant under orthogonal transformation of V_n. Therefore if, from (9), $V_n = \mathcal{W}\Gamma_n\mathcal{W}^T$, where Γ_n is a circulant matrix, we have

$$(n^{-1}\mathrm{tr}\, V_n^2)/(n^{-1}\mathrm{tr}\, V_n)^2 = (n^{-1}\mathrm{tr}\, \Gamma_n^2)/(n^{-1}\mathrm{tr}\, \Gamma_n)^2 = \sum_{j=0}^{n-1} (\gamma_j^{(n)})^2/(\gamma_0^{(n)})^2$$

and so condition (27) is precisely equivalent to the condition (11) on the covariance of the original data.

As far as $\overline{\tau^2}$ is concerned, we have

$$\overline{\tau^2} = n^{-1} \sum (\sigma^{ii})^{-1} \geq (n^{-1} \sum \sigma^{ii})^{-1} = (n^{-1} \mathrm{tr}\, V_n^{-1})^{-1}$$

since the arithmetic mean of the quantities $(\sigma^{ii})^{-1}$ is necessarily greater than their harmonic mean. Therefore the ratio

$$\overline{\sigma^2} / \overline{\tau^2} \leq (n^{-1} \mathrm{tr}\, V_n)(n^{-1} \mathrm{tr}\, V_n^{-1}) = \kappa_n, \tag{28}$$

say, and so Theorem 2 will remain true if $\overline{\sigma^2}/\overline{\tau^2}$ is replaced by κ_n, and if condition (19) is replaced by the condition that κ_n remains bounded as $n \to \infty$.

The quantity κ_n is clearly invariant under orthogonal transformations of V_n. If V_n satisfies (9), then we have

$$\kappa_n = (n^{-1} \mathrm{tr}\, \Gamma_n)(n^{-1} \mathrm{tr}\, \Gamma_n^{-1}) = \gamma_0^{(n)} \times n^{-1} \sum_{j=0}^{n-1} (\tilde{\gamma}_j^{(n)})^{-1} = k_n$$

as defined in (12) above. This gives an interpretation of κ_n in terms of the 'two-sided prediction variance' of the original process. Each element on the diagonal of Γ_n is equal to $n^{-1} \mathrm{tr}\, \Gamma_n$ and each element on the diagonal of Γ_n^{-1} is equal to $n^{-1} \mathrm{tr}\, \Gamma_n^{-1}$, since the matrices are circulant matrices. Therefore

$$\kappa_n = \mathrm{var}(Y_i) / \mathrm{var}(Y_i | Y_j, j \neq i),$$

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

Of course in practice the ratio $\overline{\sigma^2}/\overline{\tau^2}$ may be substantially closer to 1 than k_n, because the wavelet transform will often reduce the correlation, and so the factor k_n may be pessimistic. This is particularly likely to be the case because wavelets are 'almost-eigenfunctions' of many operators (see Frazier *et al.*, 1991, Meyer, 1990) and 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. In the extreme case where the sample wavelet coefficients are independent, $\overline{\sigma^2}/\overline{\tau^2} = 1$. A simple example is given by the model for the errors in the original observations

$$e_i = Z_i \sigma^2 \cos\alpha + Z_0 \sigma^2 \sin\alpha \text{ for } i = 1, 2, \ldots, n$$

where $Z_0, Z_1, \ldots, Z_n$ are independent $N(0,1)$ random variables, so that a proportion $\sin^2\alpha$ of the variability is due to a random systematic error. The wavelet coefficients will be independent, with $\mathrm{var}\, w_{00} = (n \sin^2\alpha + \cos^2\alpha)\sigma^2$ and $\mathrm{var}\, w_{jk} = \sigma^2 \cos^2\alpha$ otherwise. Because the coefficients are independent, we will have $\overline{\sigma^2}/\overline{\tau^2} = 1$, but the bound $k_n = 1/\cos^2\alpha$.

In this section, we have discussed the weakening of the main results by the substitution of quantities invariant under orthogonal transformations. While this yields convenient results for many noise processes, some preliminary calculations indicate that certain noise processes will be ruled out, such as those obtained by successively finer sampling of a continuous-time process on a fixed interval. Detailed investigation is a topic for future research.

6.2 Conclusions and remarks

We can now draw conclusions for the properties of wavelet shrinkage for signal extraction in the presence of stationary correlated noise. These extend the corresponding properties for the white noise case. As in Section 3.1, consider the estimator $\hat{\mathbf{f}}$ obtained by level-dependent thresholding of the wavelet transform of the observed data, using the thresholds (10).

Firstly, Theorem 1 implies

$$E\|\hat{\mathbf{f}} - \mathbf{f}\|^2 \leq (1 + 2\log n)\{\gamma_0^{(n)} + \sum_{j,k}(\theta_{jk}^2 \wedge s_j^2)\} \qquad \text{for all } \mathbf{f} \in I\!R^n, \tag{29}$$

so that $\hat{\mathbf{f}}$ is able to imitate a wavelet diagonal projection oracle up to a factor of $O(\log n)$. Thus wavelet shrinkage retains the ability to exploit sparsity of the underlying signal in the wavelet basis.

Further, under conditions (18) and (19), the rate at which $\hat{\mathbf{f}}$ mimics the oracle is the best possible in a minimax sense, since from Theorem 2,

$$\liminf_{n\to\infty} (2\log n)^{-1} k_n \inf_{\hat{\mathbf{f}}} \sup_{\mathbf{f}\in I\!R^n} \frac{E\|\hat{\mathbf{f}} - \mathbf{f}\|^2}{\gamma_0^{(n)} + \sum_{j,k}(\theta_{jk}^2 \wedge s_j^2)} \geq 1. \tag{30}$$

This result only differs from the white noise case by the inclusion of the bounded factor k_n. It is interesting to note that the only dependence in (29) and (30) on the wavelet transform is in the quantity $\sum_{j,k}(\theta_{jk}^2 \wedge s_j^2)$ which is related to the efficiency of compression of the signal f relative to the noise process.

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 pre-conditioning 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 ideal risk bounds in (29) and (30) 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 possess 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 $V = A^{-1}\tilde{V}A^{-T}$): 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 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.

Acknowledgments

This work was supported in part by grants from SERC/EPSRC, National Science Foundation (DMS 92-09130) and National Institutes of Health (CA 59039-18). The work was facilitated by visits by BWS to Stanford and IMJ to Bristol. Helpful comments of Felix Abramovich and Guy Nason are gratefully acknowledged.

References

Brillinger, D. R. (1994). Some asymptotics of wavelet fits in the stationary error case. Technical Report 415, University of California, Berkeley.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harm. Anal.*, to appear.

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

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? (with Discussion). *J. Roy. Statist. Soc.* B, **57**, 000–000.

Frazier, M., Jawerth, B. and Weiss, G. (1991). *Littlewood–Paley Theory and the Study of Function Spaces.* Providence, RI: Americal Mathematical Society.

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

Meyer, Y. (1990). *Ondelettes et Operateurs.* (3 volumes). Paris: Hermann.

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

Wang, Y. (1994). Function estimation via wavelets for data with long-range dependence. Technical report, University of Missouri–Columbia.