

# Real nonparametric regression using complex wavelets

Stuart Barber and Guy P. Nason

*University of Bristol, UK*

[May 9, 2003]

**Summary.** Wavelet shrinkage is an effective nonparametric regression technique, especially when the underlying curve has irregular features such as spikes or discontinuities. The basic idea is simple: take the discrete wavelet transform (DWT) of data consisting of a signal corrupted by noise; shrink or remove the wavelet coefficients to remove the noise; and then invert the DWT to form an estimate of the true underlying curve. Various authors have proposed increasingly sophisticated methods of doing this using real-valued wavelets. Complex-valued wavelets exist, but are rarely used. We propose two new shrinkage techniques which use complex-valued wavelets. Extensive simulations show that our methods almost always give significantly more accurate estimates when compared with methods based on real-valued wavelets. Further, one of our methods is both simpler and dramatically faster than its competitors. In an attempt to understand the excellent performance of this latter method we present a new risk bound on its hard thresholded coefficients.

*Keywords:* *Keywords:* Complex-valued wavelets; Complex normal distribution; Curve estimation; Empirical Bayes; Multiwavelets.

## 1 Introduction

Wavelet shrinkage (Donoho and Johnstone, 1994) has become a popular method for the estimation of a signal corrupted by noise. The simplest version of this problem is typically expressed as the estimation of a signal vector  $\mathbf{g} = (g(t_1), \dots, g(t_n))^T$  given noisy data  $\mathbf{y} = (y_1, \dots, y_n)^T$  where

$$y_i = g(t_i) + e_i \quad i = 1, \dots, n, \quad (1)$$

---

*Address for correspondence:* Department of Mathematics, University Walk, University of Bristol, Bristol, BS8 1TW, England  
Email: G.P.Nason@bristol.ac.uk

and  $t_i = i/n$  and the errors  $e_i$  are assumed independently  $N(0, \sigma^2)$  distributed.

The main advantages of wavelet shrinkage are that it is highly adaptive to irregular signals as well as smooth ones and the computations involved are of order  $n$ . There has been an explosive growth in wavelet shrinkage methodology in the past decade with increasingly sophisticated shrinkage rules being applied. Some of these rules are briefly described in sections 2 and 3 below. Almost all of this work has used the well known extremal phase and least asymmetric families of compactly supported wavelets described by Daubechies (1992). Partly this is due to the availability of software to compute these transforms, but mainly the Daubechies' wavelets are used due to their simplicity and elegance. However, many other wavelets do exist, including the complex-valued Daubechies wavelets (cDws) used by Lawton (1993) and Lina and Mayrand (1995). These cDws have been used to analyse images and complex-valued signals. However, for solving (1), they have attracted relatively little attention.

Section 2 briefly reviews some pertinent facts about wavelets and wavelet shrinkage before discussing the derivation and prior use of cDws. For more detailed introductions to wavelets in statistics, see Vidakovic (1999) or Abramovich *et al.* (2000). Section 3 introduces our two new proposals for the estimation of real signals using complex wavelets and derives a risk bound for one of them. We then compare our methods to state-of-the-art shrinkage rules that use real wavelets in section 4. Some concluding remarks and discussion are presented in section 5.

## 2 Wavelets and wavelet shrinkage

### 2.1 Real- and complex-valued wavelets

A mother wavelet,  $\psi(t)$ ,  $t \in \mathbb{R}$ , is an oscillatory function which decays rapidly in time. All the wavelets we consider in this article are compactly supported — that is they are zero outside some compact interval. Wavelets are formed from the mother wavelet by dyadic dilations and translations; the wavelet at resolution level  $j$  and location  $k$  is given by  $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$  for  $j, k \in \mathbb{Z}$ . Associated with the mother wavelet is a father wavelet,  $\phi(t)$ . For certain choices of  $\phi(t)$ ,  $\psi(t)$ , and for any fixed integer  $j_0$ , the set  $\{\phi_{j_0,k}(t), \psi_{j,k}(t) : j, k \in \mathbb{Z}, j \geq j_0\}$  is an orthonormal basis of  $\mathbb{L}_2(\mathbb{R})$ , the space of square

integrable functions on  $\mathbb{R}$ .

Usually in statistical problems we have finite sets of discrete data; if we have  $n = 2^J$  values of  $g(t)$  equally spaced between 0 and 1, we use wavelets at levels  $j = 0, \dots, J - 1$ . Level 0 contains the mother and father wavelets while increasing values of  $j$  correspond to wavelets which describe finer detail. Henceforth, in a slight abuse of notation, we shall use  $\psi(t)$  and  $\phi(t)$  to refer to discrete wavelets (real- or complex-valued) and the data vector  $\mathbf{g} = (g(t_1), \dots, g(t_n))^T$  can be expressed as

$$g(t_i) = c_{0,0}\phi(t_i) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t_i), \quad i = 1, \dots, n.$$

The vector  $\mathbf{d} = (c_{0,0}, d_{0,0}, \dots, d_{J-1, 2^{J-1}-1})^T$  is referred to as the discrete wavelet transform (DWT) of  $\mathbf{g}$ . Due to the design of the wavelets, the expression of  $\mathbf{g}$  in terms of the wavelet basis is typically very sparse; that is only a few elements of  $\mathbf{d}$  will be large and most will be zero or nearly so.

The DWT may be represented by an  $n \times n$  unitary matrix  $W$ , constructed from  $\psi(t)$ , such that  $\mathbf{d} = W\mathbf{g}$ . (A unitary matrix is one where  $W\bar{W}^T = \bar{W}^TW = I_n$  where  $\bar{\cdot}$  denotes complex conjugation and  $I_n$  is the  $n \times n$  identity matrix.) In practice the pyramidal DWT algorithm of Mallat (1989) is used to compute  $\mathbf{d}$  in  $\mathcal{O}(n)$  operations rather than the slow  $\mathcal{O}(n^2)$  of matrix multiplication.

A detailed description and derivation of complex Daubechies' wavelets is given by Lina and Mayrand (1995). Daubechies' wavelets (both real- and complex-valued) are indexed by the number of vanishing moments,  $N$ . Increasing values of  $N$  correspond to smoother wavelets. For a given value of  $N$  there are  $2^{N-1}$  possible solutions to the equations that define the Daubechies' wavelets, but not all are distinct. For example, when  $N = 3$  there are four solutions. However, only two are distinct: two solutions give the real extremal phase wavelet, the other two are a conjugate pair (hence giving equivalent complex-valued wavelets). This wavelet was also derived by Lawton (1993) by "zero-flipping" and pictures of the wavelets appear in that article. Lawton notes that, apart from the Haar wavelet, the only compactly supported wavelets which are symmetric are cDws with an odd number of vanishing moments. As well as these symmetric solutions, asymmetric complex-valued

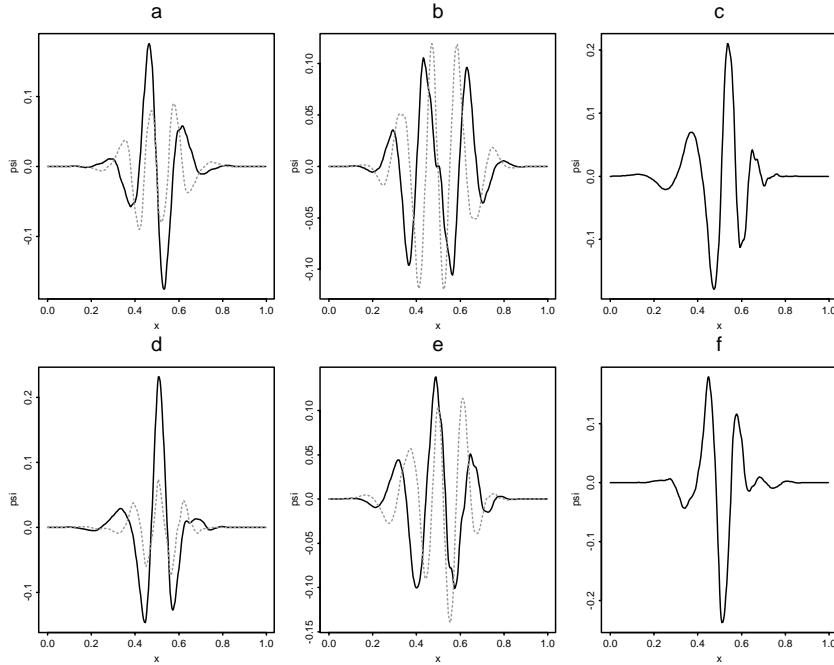


Figure 1: *Daubechies' wavelets with five vanishing moments. In the complex-valued plots the real part is drawn as a solid black line and the imaginary part as a dotted line.*

wavelets occur when there are four or more vanishing moments.

Figure 1 shows the six distinct Daubechies' wavelets with five vanishing moments. Panels (c) and (f) show respectively the familiar extremal phase and least asymmetric real wavelets, while the other panels show the four distinct complex-valued wavelets. Even though these cDws share the same number of vanishing moments, they are visually very different to each other. In particular, the number of oscillations vary, as does the relative magnitude of the real and imaginary parts.

## 2.2 Wavelet shrinkage

Consider data of the form (1) with  $n = 2^J$  for a positive integer  $J$ . Equivalently, we write  $\mathbf{y} = \mathbf{g} + \mathbf{e}$ , where  $\mathbf{y}$ ,  $\mathbf{g}$  and  $\mathbf{e}$  are real column vectors of the observed data, signal and noise respectively. Let the DWT of  $\mathbf{y}$  be  $\mathbf{d}^* = W\mathbf{y} = \mathbf{d} + \boldsymbol{\varepsilon}$ , where  $\mathbf{d} = W\mathbf{g}$  and  $\boldsymbol{\varepsilon} = W\mathbf{e}$  respectively.

Our goal is to recover  $\mathbf{g}$  from the noisy data  $\mathbf{y}$ ; equivalently, we can estimate the true

wavelet coefficients  $\mathbf{d}$  from the empirical coefficients  $\mathbf{d}^*$ . In the case of real wavelets, this problem has been extensively addressed; for a review of wavelet shrinkage, see Abramovich *et al.* (2000). Particularly relevant to this article are the multiwavelet approach of Downie and Silverman (1998) and the empirical Bayes method discussed by Johnstone and Silverman (2002a); these methods are briefly reviewed in sections 3.1 and 3.2 respectively.

With complex-valued wavelets Lina and MacGibbon (1997), Lina (1997) and Lina, Turcotte and Goulard (1999) concentrate on *image* denoising with a Bayesian shrinkage rule. Sardy (2000) considers the estimation of *complex-valued* signals with threshold choices informed by minimax risk minimization (giving a universal threshold of  $\sqrt{2 \log(n \log n)}$ , which is very close to the universal threshold for bi-wavelet shrinkage in Downie and Silverman (1998)). All of this work shrinks only the modulus of the complex coefficients, leaving the phase alone (*phase preserving* shrinkage). The approach we take in Section 3.1 based on ideas from multiwavelet shrinkage uses phase-preserving shrinkage rules.

For image denoising phase preservation is undoubtedly useful, since certain image features are not efficiently captured by error measures such as mean-squared error but will be destroyed by phase alteration. Phase preservation is again possibly good for denoising complex-valued signals. To gain some theoretical understanding of the behaviour of the phase-preserving hard-thresholded complex shrinkage rule we derive an upper bound for its risk in Theorem 1 in section 3.1.3. On the other hand, for real-valued signals, we see no reason why the two-dimensional complex coefficient should have its phase preserved. We cannot do worse if we permit phase alteration as phase preserving rules are a special case of general shrinkage rules. Two of our empirical Bayes estimators in section 3.2 permit phase alteration and perform extremely well in our simulation study in section 4.

Zaroubi and Goelman (2000) describe a method for denoising complex-valued MRI scans by separately thresholding the real and imaginary parts. Our methodology is also capable of native denoising of complex signals and would differ from Zaroubi and Goelman (2000) as we would treat the real and imaginary parts as a pair and not separately. However, we only consider denoising of real-valued signals in this article.

### 2.3 Noise structure of complex wavelet coefficients

The DWT maps the signal domain noise vector  $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 I_n)$  to a noise vector  $\varepsilon = W\mathbf{e}$  in the wavelet domain. If the wavelet used is real-valued then  $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 I_n)$  because  $W$  is orthogonal. This is not the case when the decomposing wavelet is complex-valued. The individual components of  $\varepsilon$ , considered as *complex-valued* random variables, are uncorrelated. However, the real and imaginary parts of  $\varepsilon$  are normal real-valued random variables in their own right and can be strongly correlated as the following proposition demonstrates.

**Proposition 1** *Let  $\varepsilon = W\mathbf{e}$  where  $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 I_n)$  and  $W$  is a unitary matrix. Then*

$$\text{cov}\{\text{Re}(\varepsilon), \text{Im}(\varepsilon)\} = -\sigma^2 \text{Im}(WW^T)/2, \quad (2)$$

$$\text{cov}\{\text{Re}(\varepsilon), \text{Re}(\varepsilon)\} = \sigma^2 \{I_n + \text{Re}(WW^T)\} / 2 \quad (3)$$

$$\text{cov}\{\text{Im}(\varepsilon), \text{Im}(\varepsilon)\} = \sigma^2 \{I_n - \text{Re}(WW^T)\} / 2. \quad (4)$$

*Proof: See appendix A.*

Figure 2 illustrates the covariance matrix  $\text{cov}\{\text{Re}(\varepsilon), \text{Im}(\varepsilon)\}$  for a noise vector of  $n = 128$   $N(0, 1)$  random variables decomposed with the symmetric cDw with  $N = 5$ . The denoising methods that we propose in section 3 exploit the correlation between the real and imaginary components at each time-scale location but, in this article, not between different locations.

For the remainder of this article, we regard a complex-valued random variable as a bivariate real-valued random variable. Any given element  $\varepsilon_{j,k}$  of the vector  $\varepsilon$  has a complex normal distribution equivalent to a bivariate normal with mean vector zero and covariance matrix  $\Sigma_{j,k}$ . This covariance matrix can be formed by selecting appropriate values from the diagonals of the matrices in (2) to (4). In fact, since we are using periodic transforms, it turns out that all the covariance matrices for a given resolution level are equal and henceforth we shall omit the subscript “ $k$ ” on the covariance matrices.

Equations (2) to (4) specify the covariance of the empirical wavelet coefficients assuming that the noise level  $\sigma^2$  is known. This is not the case in practice. The usual approach is to estimate  $\sigma^2$  by the squared median absolute deviation of the wavelet coefficients at the finest resolution level. We use an equivalent estimator, the sum of the squared median absolute deviation of the real and imaginary parts of the finest level coefficients. Regardless

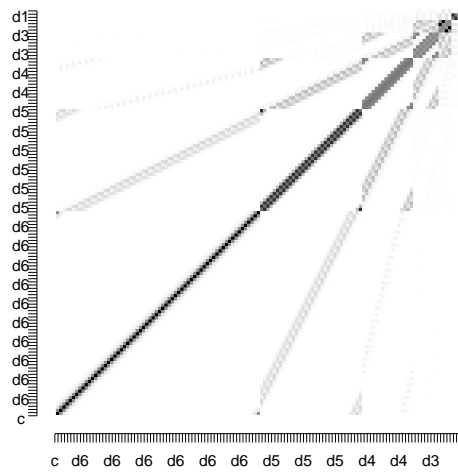


Figure 2: (Absolute) covariance between  $\text{Re}(\varepsilon)$  and  $\text{Im}(\varepsilon)$  for a noise vector of length  $n = 128$  with variance 1 decomposed with the  $cDw$   $N = 5$  shown in Figure 1(e). The axes correspond to wavelet coefficients  $d_j$  at levels  $j = 0, \dots, 6$ . White corresponds to zero covariance and black to the maximum absolute covariance (0.29). Covariances between a coefficient and its neighbour are clearly seen along the diagonal. Weaker covariances between a coefficient and near coefficients on the neighbouring few scales can also be seen as the “wider” diagonals.

of the approximation used, we then use this value and assume that the noise level is well estimated.

### **3 Using complex wavelets to denoise real signals**

#### **3.1 Multiwavelet style shrinkage**

##### **3.1.1 Brief review of multiwavelet shrinkage**

In an attempt to denoise a real signal using complex-valued wavelets, we first consider an approach similar to the multiwavelet scheme used by Downie and Silverman (1998). A multiwavelet transform is similar to the standard wavelet transform except that at each time-scale location there are  $L$  coefficients. The transform derives from an orthonormal, dyadically decimated and translated basis of  $L$  mother wavelets. If  $L = 2$  then the multiwavelet transform has two coefficients at each time-scale location, similar to the complex-wavelet transform. A key difference between two dimensional multiwavelets and complex wavelets is that the two mother wavelets used in a multiwavelet transform are orthogonal, which is not true of the real and imaginary parts of a complex mother wavelet. Various authors have considered denoising real-valued signals using the multiwavelet transform, see Downie and Silverman (1998).

The discrete multiwavelet transform (DMWT) requires each input data value to be an  $L$ -dimensional vector. Unfortunately, the data from our model in (1) is one-dimensional. For multiwavelet shrinkage a prefilter is applied to the real-valued univariate data to obtain  $L$ -dimensional starting vectors. After the usual process of DMWT, shrinkage and inverse DMWT a postfilter (the prefilter inverse) is applied. Downie and Silverman (1998) state that every prefilter, apart from the identity prefilter, gives correlated coefficients, by which they mean that the two components at each time-scale location are correlated. Downie and Silverman (1998) devise a thresholding technique to take account of this correlation. The identity prefilter would result in uncorrelated coefficients but it is strongly not recommended for wavelet shrinkage for other reasons.

### 3.1.2 Complex multiwavelet style (CMWS) shrinkage

In contrast to the multiwavelet case (where the univariate real data points have to be mapped to a multivariate  $L$ -vector input coefficients) the inputs to the complex wavelet transform are univariate. The use of the complex transform for processing real data is natural because  $\mathbb{R}$  is a subset of  $\mathbb{C}$ . Thus, unlike the multiwavelet transform, the complex transform requires no prefilter.

Our CMWS method is based on Downie and Silverman (1998) as follows. Writing the complex-valued empirical wavelet coefficient  $d_{j,k}^*$  as a column vector, recall that  $d_{j,k}^* \sim N_2(d_{j,k}, \Sigma_j)$ . For each  $d_{j,k}^*$ , we compute a ‘‘thresholding statistic’’  $\theta_{j,k} = d_{j,k}^{*T} \Sigma_j^{-1} d_{j,k}^*$ . If this value exceeds the threshold  $\lambda = 2 \log n$ , then the coefficient is retained; otherwise it is set to zero. This defines a hard-thresholding estimation rule  $\widehat{d}_{j,k}^{MH} = d_{j,k}^* I(\theta_{j,k} > \lambda)$ , where  $I(\cdot)$  is the indicator function. Soft thresholding is also possible. In this case, we estimate  $d_{j,k}$  by

$$\widehat{d}_{j,k}^{MS} = \frac{d_{j,k}^*}{|d_{j,k}^*|} \max\{\theta_{j,k} - \lambda, 0\};$$

note the thresholding is done on the  $\theta_{j,k}$  scale. Both these thresholding rules result in an estimated wavelet coefficient with the same phase as the empirical coefficient. Our estimate  $\widehat{\mathbf{g}}$  is then formed by inverting the DWT using the estimated wavelet coefficients. We note that this estimate will not, in general, be purely real-valued; since we know that we are attempting to recover a real signal we discard any imaginary component in our estimate. In practice, we have always found the imaginary parts of  $\widehat{\mathbf{g}}$  to be negligible.

The choice of threshold is based on the fact that  $\theta_{j,k}$  has a non-central  $\chi_2^2$  distribution with non-centrality parameter  $d_{j,k}^{*T} \Sigma_j^{-1} d_{j,k}^*$ . In particular, if  $d_{j,k} = 0$ , then  $\theta_{j,k}$  follows a central  $\chi_2^2$  distribution. Downie and Silverman (1998) use this fact to derive the threshold  $\lambda(L) = 2 \log n + (L - 2) \log \log n$ , which in our bivariate case reduces to  $\lambda = 2 \log n$ .

This problem is related to that of estimating the non-centrality parameter of a single  $\chi_d^2$  distribution. Saxena and Alam (1982) consider this problem and show that soft thresholding with a threshold of  $d$  (considerably lower than the Downie-Silverman threshold) has a lower risk than maximum likelihood estimation. In practice, we have found that a threshold of  $\lambda = 2$  does not do enough thresholding. Johnstone (2000) has also considered threshold estimators in this contest, while Fourdriner *et al.* (2000) consider the Bayesian estimation

of the non-centrality parameter under Stein-type losses.

### 3.1.3 Risk bound for complex multiwavelet style hard thresholding

The simulation study in Section 4 shows that our complex-wavelet denoising outperforms other methods when applied to real signals. The purpose of this section is to try to obtain some theoretical understanding of why this is so. We establish a bound on the risk of hard thresholding for the complex case and compare it to the equivalent result from Johnstone and Silverman (2002b) in the real-valued case.

In this section only we change notation to be consistent with Johnstone and Silverman (2002b). The notation change is  $d^* \rightarrow X$ ,  $d \rightarrow \mu$ ,  $\lambda \rightarrow t$  although both notations are regularly used in the wavelet shrinkage literature. In this section we only consider the theoretical properties of a single wavelet coefficient and hence the  $j, k$  subscripts are omitted.

In the signal plus Gaussian noise case  $X$  is a complex normal random variate with mean  $\mu = \mu_1 + i\mu_2$  and variance matrix  $\Sigma$ . For complex-wavelets, the estimator  $\hat{\mu}(X)$  is a shrinkage rule if  $|\hat{\mu}(X)| \leq |X|$  for all  $X$ .

The hard-thresholding rule  $\hat{\mu}_{\text{HT}}(X, t) = XI(\theta > t)$  where  $\theta = X^T \Sigma^{-1} X$  is the covariance corrected coefficient from Downie and Silverman (1998). We are looking for a bound on the risk of this estimator,

$$r_{\Sigma}(\mu, t) := E|\hat{\mu}_{\text{HT}}(X, t) - \mu|^2.$$

The bound is stated in theorem 1, which is proved in Appendix B.

**Theorem 1** *Let  $a_2 := \Sigma_{11} + \Sigma_{22}$  and let  $\sigma$  and  $\tau$  be respectively the largest and smallest eigenvalues of  $\Sigma$ . Then for  $|\mu| \geq 1$*

$$r_{\Sigma}(\mu, t) \leq (1 + a_2)|\mu|^2.$$

*For  $|\mu| < 1$  and  $|\mu|^2 > t/\tau$  then*

$$r_{\Sigma}(\mu, t) \leq K_1 + K_2 \phi_{0,\sigma}(|\mu|) \exp(tK_3)(|\mu|^2 + tK_4 + K_5) + |\mu|^2 K_6. \quad (5)$$

For  $|\mu| < 1$  and  $|\mu|^2 < t/\tau$  then

$$r_{\Sigma}(\mu, t) \leq K_7 \phi_{0, \sigma^{-1/2}} \left( |\mu| - \sqrt{t/\tau} \right) \left\{ \left( |\mu| - \sqrt{t/\tau} \right)^2 + K_8 \right\} + |\mu|^2 K_6, \quad (6)$$

where  $K_1$  to  $K_8$  are all positive constants.

The equivalent one-dimensional result from Johnstone and Silverman (2002b) is

$$E|\hat{\mu}_{\text{HT}} - \mu|^2 \leq c_2[|\mu|^2 + t\phi(t)].$$

There are clearly similarities between the one- and two-dimensional results although our bound is slightly weaker. In (5) recall that  $t/\tau < |\mu|^2$  and so  $\phi_{0, \sigma}(|\mu|) \leq \phi_{0, \sigma}(\sqrt{t/\tau})$  so control by the equivalent of  $t\phi(t)$  is possible in our version (as in Johnstone and Silverman (2002b)). It is not clear that the bound gives us clear-cut information as to which cDws would exhibit superior overall performance but indications in special cases could be determined by the more detailed bounds in B.6.

### 3.2 Bayesian methods

Bayesian thresholding rules have been shown to be highly effective, such as the ABWS method discussed by Chipman *et al.* (1997), the BayesThresh method by Abramovich *et al.* (1998) and the EBayesThresh method proposed by Johnstone and Silverman (2002a). The majority of these Bayesian rules place an independent prior on each wavelet coefficient. These priors are updated by the empirical wavelet coefficients to give posterior distributions and some location estimator (such as the posterior mean or median) is then applied to “estimate” the “true” coefficients. The more successful Bayesian rules have employed prior distributions which incorporate a point mass at zero to represent the fact that many of the true coefficients are either exactly or nearly zero. Lina and MacGibbon (1997) and Lina *et al.* (1999) have suggested Bayesian methods using complex-valued wavelets in image denoising, but threshold only the magnitude of the empirical coefficients, leaving the phase unchanged.

We now describe our complex empirical Bayes (CEB) procedure. Writing the complex

coefficient  $d_{j,k}$  as a column vector, we consider a prior of the form

$$d_{j,k} \sim p_j N_2(\mathbf{0}, V_j) + (1 - p_j)\delta_0, \quad (7)$$

where  $\delta_0$  is a point mass of probability at  $(0, 0)^T$ . An intuitive interpretation of (7) is that, a priori, all coefficients on level  $j$  independently have “probability of being non-zero”  $p_j$  and those coefficients which are non-zero follow independent bivariate normal distributions with mean zero and covariance matrix  $V_j$ . This is the obvious bivariate extension of the prior used in the BayesThresh method of Abramovich *et al.* (1998).

Given an observed value  $d_{j,k}^*$ , standard results for the multivariate normal distribution show that the posterior distribution of  $d_{j,k}$  is of the same form as (7):

$$d_{j,k} | d_{j,k}^* \sim \tilde{p}_{j,k} N_2(\mu_{j,k}, \tilde{V}_j) + (1 - \tilde{p}_{j,k})\delta_0, \quad (8)$$

where

$$\begin{aligned} \tilde{p}_{j,k} &= \frac{p_j f(d_{j,k}^* | p_j = 1)}{p_j f(d_{j,k}^* | p_j = 1) + (1 - p_j) f(d_{j,k}^* | p_j = 0)}, \\ \tilde{V}_j &= \left( V_j^{-1} + \Sigma_j^{-1} \right)^{-1} \text{ and } \mu_{j,k} = \tilde{V}_j \Sigma_j^{-1} d_{j,k}^*; \end{aligned}$$

see, for example, O’Hagan (1994, p.290).

This leaves two unresolved questions: choice of posterior location measure and specification of the prior parameters. Abramovich *et al.* (1998) use the posterior median which leads to a true thresholding rule and Johnstone and Silverman (2002a) show that in the one-dimensional case the median has better asymptotic properties. However, there are many alternative multivariate medians to choose between (see Small, 1990 for a review) and it is not clear which is more appropriate, nor are they trivial to compute. We have considered defining a form of posterior median by finding the contour where the posterior distribution function equals one half and selecting a point on this contour, but this approach has proved unsatisfactory for several reasons. Firstly, there is an arbitrary choice to be made in selecting a point on the contour; secondly some natural rules for this choice did not always have a solution; and finally the calculations required were highly computationally intensive.

Since Johnstone and Silverman (2002a) also report that in simulation studies the posterior mean is often superior to the median, we consider three estimation rules based on the posterior mean of (8):

$$\begin{aligned}
\text{CEB-Keep or kill} & \quad \widehat{d}_{j,k}^{BH} = d_{j,k}^* I(\tilde{p}_{j,k} > \frac{1}{2}), \\
\text{CEB-Posterior mean} & \quad \widehat{d}_{j,k}^{BM} = \tilde{p}_{j,k} \mu_{j,k}, \\
\text{CEB-MeanKill} & \quad \widehat{d}_{j,k}^{BS} = \tilde{p}_{j,k} \mu_{j,k} I(\tilde{p}_{j,k} > \frac{1}{2}).
\end{aligned}$$

The first estimator is a simple keep-or-kill thresholding procedure based on the posterior mixing parameter, the second is the posterior mean and the third estimator is a hybrid which kills “small” empirical coefficients and estimates the remainder by the posterior mean.

Prior parameter specification is more difficult. Following Clyde and George (2000) and Johnstone and Silverman (2002a) we employ an automatic empirical Bayes approach which uses the maximum likelihood estimates of the parameters  $p_j$  and  $V_j$  for  $j = j_0, \dots, J-1$ . Here,  $j_0$ , is the primary resolution: scales coarser than  $j = j_0$  are left unchanged. For a single coefficient,  $d_{j,k}^*$ , the likelihood is given by

$$L_{j,k} = \frac{p_j}{2\pi\sqrt{|V_j + \Sigma_j|}} \exp\left\{-\frac{1}{2}d_{j,k}^{*T}(V_j + \Sigma_j)^{-1}d_{j,k}^*\right\} + \frac{1-p_j}{2\pi\sqrt{|\Sigma_j|}} \exp\left\{-\frac{1}{2}d_{j,k}^{*T}\Sigma_j^{-1}d_{j,k}^*\right\}. \quad (9)$$

We maximize  $\ell(p_j, V_j | d_{j,k}^*) = \sum_k \log L_{j,k}$  using the E04JYF NAG routine (a version using `nlmnrb` from S-Plus is also available). Alternatives to the empirical Bayes approach often reparameterise the hyperparameters in terms relating to the smoothness of the underlying function and/or the Besov function class parameters, see, e.g. Abramovich *et al.* (1998). The resulting parameterisations can often be difficult to interpret intuitively.

In the one-dimensional case of denoising real data with real wavelets, Johnstone and Silverman (2002a) report that using heavier-tailed priors (such as the Laplace distribution) results in superior mean squared error performance of wavelet thresholding schemes. However, in the bivariate case, the use of a Laplace prior is intractable. If a bivariate random variable  $Z$  follows a Laplace distribution with mean zero and covariance matrix  $C$ , its density is proportional to  $\exp\left\{-\left(z^T C^{-1} z\right)^{1/2}\right\}$ . So if the normal component of the prior (7) is replaced by a bivariate Laplace distribution, convolving the Laplace and normal density

functions results in a posterior distribution with density proportional to

$$\exp \left\{ -\frac{1}{2} (d_{j,k}^* - d_{j,k})^T \Sigma_j^{-1} (d_{j,k}^* - d_{j,k}) - (d_{j,k}^T C^{-1} d_{j,k})^{1/2} \right\}.$$

This density has proved to be analytically intractable. Numerical methods could be used to provide a posterior mean from this distribution, but this would destroy the computational advantages of the wavelet approach. Similar problems occur for other heavy-tailed priors, such as a bivariate Student's- $t$  distribution on one degree of freedom.

## 4 Simulation results

We have used simulation to assess the finite sample performance of our methods. The simulated data sets consisted of the standard test signals blocks, bumps, doppler, heavisine (Donoho and Johnstone, 1994) and ppoly (Nason and Silverman, 1994), corrupted by independent Gaussian noise as in (1). All of the test signals were rescaled so as to have unit variance. The degree of noise is measured by the ratio of the standard deviations of the signal and noise, referred to as the root signal to noise ratio (rsnr).

We also consider two types of basis averaging. One, common in the wavelet literature, uses cycle-spinning (Coifman and Donoho, 1995; Nason and Silverman, 1995). Full cycle-spinning averages the results of wavelet shrinkage applied to all  $n$  cyclically rotated shifts of the data using a fast  $\mathcal{O}(n \log n)$  nondecimated wavelet transform.

A second type of basis averaging, less commonly used, is to analyse the data separately using several different wavelets and to average the results, see Kohn *et al.* (2000). We refer to this as basis averaging over wavelets (BAW). Unlike cycle-spinning, the computational load of this approach remains  $\mathcal{O}(n)$ , although it grows linearly with the number of wavelets used.

Table 1 shows the average mean square error (AMSE) and associated standard errors (both multiplied by  $10^5$  for clarity) obtained when denoising signals using two of our new methods: CMWS hard thresholding and CEB MeanKill using the range of cDws. The other shrinkage options for our new methods were slightly less effective in general, although not substantially so. The results in Table 1 are based on 100 simulated data sets of length

$n = 1024$  with  $\text{rsnr} = 7$ . In each case, no thresholding was done below level 3.

There is little to choose between the multiwavelet style and empirical Bayes estimates, and in each case BAW is superior to using any one cDw when the decimated DWT is used. However, when cycle-spinning is used this situation is reversed, with the best single cDw outperforming the BAW estimates. Usually the best single wavelet is the cDw 3.1 (the Lawton wavelet), regardless of whether the decimated or nondecimated DWT is used. However, for the smoother doppler signal (and to a lesser extent for the heavisine signal), smoother cDws are better.

We have compared all of our methods to a range of methods using real-valued wavelets, using the same simulated data sets as in Table 1. A selection of results is presented in Tables 2 and 3 (the full set of results can be found in Barber and Nason (2003)). Results are shown for CMWS-hard, CEB-MeanKill, the PostBlockMean empirical Bayes block-thresholding procedure from Abramovich *et al.* (2002), the EBayesthresh rule of Johnstone and Silverman (2002a) with a Laplace prior and posterior median threshold, the real-valued multiwavelet hard threshold procedure due to Downie and Silverman (1998) using the Geronimo-Hardin-Massopust biwavelet with repeated signal prefilter (MW-GHM rep), the false discovery rate thresholding (FDR) proposed by Abramovich and Benjamini (1996) with the parameter  $q = 0.05$  and hard thresholding, the cross-validation procedure due to Nason (1996) with soft thresholding (CV-Soft). For all of these competitive procedures we have tried hard to tune their parameters to give as good results as possible (for example, tried both hard and soft thresholding, tried different values of  $q$  in FDR etc.). Real-valued multiwavelet shrinkage, and to a lesser extent cross-validation, was found to be highly sensitive to choice of the primary resolution. All other methods were found to be robust to the choice of primary resolution, see Barber and Nason (2003) for further details. Johnstone and Silverman (2002a) have already demonstrated the superiority of their EBayesThresh rule to simple thresholding with the universal threshold of Donoho and Johnstone (1995), SURE thresholding (Donoho and Johnstone, 1994), and the NeighBlock and NeighCoeff block thresholding rules of Cai and Silverman (2001) and hence we omit these methods in our comparisons.

Tables 2 and 3 show the results when the decimated and nondecimated DWT were used, respectively. Both tables also show the results of BAW; for our methods this is over the six

cDws with three to five vanishing moments, while for the other methods this averaging is over the five real Daubechies' wavelets with the same number of vanishing moments. Our methods always obtain the best results, often by a substantial margin. In fact, with one exception (heavisine in Table 3), the worst of our methods is better than the best of the real-valued wavelet methods and again often substantially better. We have seen similar patterns in extensive simulation results, not shown here but presented in Barber and Nason (2003), for  $n = 256, 512, 1024,$  and  $2048$  and for  $\text{rsnr} 3, 5,$  and  $7$ .

An added advantage of our CMWS procedure is that it is extremely fast compared to those methods that require a numerical optimisation (CEB, PostBlockMean, EBayesThresh, CV). Indeed, in comparison of CPU time execution the PostBlockMean procedure took approximately eight times longer than EBayesThresh which itself took ten times longer than our CMWS method, with all code written in S-Plus and executed on the same machine over multiple simulations.

## 5 Conclusions and discussion

This article has introduced two new methods for denoising real-valued signals using complex-valued wavelets. Extensive simulation studies have shown that our methods work well across a wide range of signal types and generally outperform the existing state-of-the-art by a substantial margin often reducing the AMSE by as much as 25%. As is usual basis averaging over location or wavelet basis is typically beneficial. However, we have discovered that basis averaging over wavelets is counterproductive when cycle-spinning is already being used.

Our CEB method gives slightly better results than our CMWS method but the latter is simpler to use and implement and is also much faster. Indeed, the CMWS method is approximately 10 times faster than EBayesThresh and 80 times faster than PostBlockMean. The slow execution of the empirical Bayesian methods is almost entirely due to their use of numerical optimisation to obtain the hyperparameters. For practical use the CMWS method is particularly simple as it only requires the user to choose the type of wavelet and hard or soft thresholding (in contrast to Bayesian methods which also require a choice of prior distribution and/or (initialisation) parameters).

Wavelet	Blocks		Bumps		Doppler		HeaviSine		Ppoly	
	AMSE	SE	AMSE	SE	AMSE	SE	AMSE	SE	AMSE	SE
<i>Complex multiwavelet style (CMWS) hard thresholding (decimated)</i>										
cDw 3.1	431	5	429	5	314	4	150	3	93	2
cDw 4.1	513	6	443	5	253	4	169	3	105	3
cDw 5.x	537 <sup>4</sup>	6	459 <sup>1</sup>	5	223 <sup>3</sup>	3	145 <sup>1</sup>	3	95 <sup>2</sup>	2
cDw 5	429	5	399	4	180	3	128	2	75	2
cDw 3 to 5	<span style="border: 1px solid black;">390</span>	5	<span style="border: 1px solid black;">366</span>	4	<span style="border: 1px solid black;">179</span>	3	<span style="border: 1px solid black;">120</span>	3	<span style="border: 1px solid black;">70</span>	2
<i>Complex wavelet empirical Bayes (CEB) MeanKill (decimated)</i>										
cDw 3.1	407	5	418	4	285	4	143	3	88	2
cDw 4.1	472	6	425	4	234	3	158	3	102	3
cDw 5.x	485 <sup>4</sup>	5	447 <sup>1</sup>	4	214 <sup>4</sup>	3	143 <sup>1</sup>	3	97 <sup>1</sup>	3
cDw 5	418	5	398	4	<span style="border: 1px solid black;">177</span>	3	133	3	81	2
cDw 3 to 5	<span style="border: 1px solid black;">388</span>	5	<span style="border: 1px solid black;">367</span>	4	178	3	<span style="border: 1px solid black;">128</span>	3	<span style="border: 1px solid black;">75</span>	2
<i>Complex multiwavelet style (CMWS) hard thresholding (nondecimated)</i>										
cDw 3.1	<span style="border: 1px solid black;">322</span>	4	<span style="border: 1px solid black;">321</span>	3	215	3	<span style="border: 1px solid black;">110</span>	2	<span style="border: 1px solid black;">64</span>	2
cDw 3 to 5	359	4	338	3	<span style="border: 1px solid black;">165</span>	3	117	2	66	2
<i>Complex wavelet empirical Bayes (CEB) MeanKill (nondecimated)</i>										
cDw 3.1	<span style="border: 1px solid black;">318</span>	4	<span style="border: 1px solid black;">322</span>	3	213	3	<span style="border: 1px solid black;">124</span>	3	<span style="border: 1px solid black;">65</span>	1
cDw 3 to 5	382	4	360	4	<span style="border: 1px solid black;">161</span>	2	132	3	75	2

Table 1: Comparison of decimated and nondecimated complex wavelet shrinkage methods: CMWS-hard and CEB-MeanKill. Results are average mean square error (AMSE) and standard errors times  $10^5$  for estimation of standard test functions. Smallest AMSE in each category is indicated by a box. The line “cDw 5.x” shows the best results from each of  $x=1, 2, 3,$  and  $4$  indicated by the superscript after the AMSE. Note that cDw 5 and cDw 3 to 5 are BAW.

Method	Blocks		Bumps		Doppler		HeaviSine		Ppoly	
	AMSE	SE	AMSE	SE	AMSE	SE	AMSE	SE	AMSE	SE
<i>Wavelets with three vanishing moments (except MW-GHM which has two)</i>										
CMWS-Hard	431	5	429	5	314	4	150	3	93	2
CEB-MeanKill	<span style="border: 1px solid black;">407</span>	5	<span style="border: 1px solid black;">418</span>	4	<span style="border: 1px solid black;">285</span>	4	<span style="border: 1px solid black;">143</span>	3	<span style="border: 1px solid black;">88</span>	2
PostBlockMean	561	6	571	5	400	6	200	5	161	6
EBayesThresh	580	6	680	6	471	6	171	3	130	3
MW-GHM rep	565 <sup>5</sup>	11	749 <sup>6</sup>	7	515 <sup>5</sup>	7	208 <sup>4</sup>	6	96 <sup>3</sup>	3
FDR	672	7	738	7	568	6	249	5	167	4
CV-Soft	889	7	907	6	668	5	295	4	227	3
<i>Averaged over wavelets with three to five vanishing moments</i>										
CMWS-Hard	390	5	<span style="border: 1px solid black;">366</span>	4	179	3	<span style="border: 1px solid black;">120</span>	3	<span style="border: 1px solid black;">70</span>	2
CEB-MeanKill	<span style="border: 1px solid black;">388</span>	5	367	4	<span style="border: 1px solid black;">178</span>	3	128	3	75	2
EBayesThresh	392	4	431	4	254	3	129	3	86	2
FDR	390	5	426	4	250	3	136	3	83	2
CV-Soft	750	6	737	5	460	4	228	3	157	3

Table 2: Comparison of complex-valued wavelet shrinkage methods with existing methods using real-valued wavelets on standard test functions using decimated transforms. All results have primary resolution equal to 3 except for MW-GHM where, due to its sensitivity, the optimal primary resolution was searched for and is indicated by a superscript. Results are average mean square error (AMSE) and standard errors times  $10^5$  for estimation of standard test functions. Smallest AMSE in each category is indicated by a box.

Method	Blocks		Bumps		Doppler		HeaviSine		Ppoly	
	AMSE	SE	AMSE	SE	AMSE	SE	AMSE	SE	AMSE	SE
<i>Wavelets with three vanishing moments</i>										
CMWS-Hard	322	4	<span style="border: 1px solid black;">321</span>	3	215	3	<span style="border: 1px solid black;">110</span>	2	<span style="border: 1px solid black;">64</span>	2
CEB-MeanKill	<span style="border: 1px solid black;">318</span>	4	322	3	<span style="border: 1px solid black;">213</span>	3	124	3	65	1
EBayesThresh	353	4	433	4	270	3	128	3	70	2
FDR	423	6	472	5	334	4	165	4	101	3
CV-Soft	700	6	720	6	389	4	155	2	131	2
<i>Averaged over wavelets with three to five vanishing moments</i>										
CMWS-Hard	<span style="border: 1px solid black;">359</span>	4	<span style="border: 1px solid black;">338</span>	3	<span style="border: 1px solid black;">165</span>	3	<span style="border: 1px solid black;">117</span>	2	<span style="border: 1px solid black;">66</span>	2
CEB-MeanKill	363	4	344	4	167	3	129	3	71	2
EBayesThresh	372	4	400	4	236	3	124	3	71	2
FDR	415	5	444	4	274	4	148	3	92	2
CV-Soft	742	6	751	6	357	4	158	2	137	2

Table 3: Comparison of complex-valued wavelet shrinkage methods with existing methods using real-valued wavelets on standard test functions using non-decimated transforms. Results are average mean square error (AMSE) and standard errors times  $10^5$  for estimation of standard test functions. Smallest AMSE in each category is indicated by a box.

The thresholding rules proposed in section 3.1 implicitly assume that the thresholding statistics  $\theta_{j,k}$  and  $\theta_{j',k'}$  are independent for  $j \neq j'$  and  $k \neq k'$ . In fact, this is not the case; recall from section 2.3 that even though  $\text{cov}(d_{j,k}^*, d_{j',k'}^*) = 0$  the real and imaginary parts of  $d_{j,k}^*$  and  $d_{j',k'}^*$  are correlated and hence, in general,  $\text{cov}(\theta_{j,k}, \theta_{j',k'}) \neq 0$ . A promising topic for future research would be to exploit the correlation structure between  $\theta_{j,k}$  and  $\theta_{j,k\pm 1}$  to further improve our thresholding schemes. For example, there is no reason why block-thresholding ideas such as those proposed by Cai and Silverman (2001) or Abramovich *et al.* (2002) could not be adapted to the complex-valued wavelet case.

Software to implement our methods is available as an add-on to the WaveThresh package which can be downloaded from [www.stats.bris.ac.uk/~wavethresh](http://www.stats.bris.ac.uk/~wavethresh). This add-on includes all the cDws used in this paper, an implementation of the complex-valued nondecimated wavelet transform and all of our complex-valued wavelet thresholding methods.

## Acknowledgments

We would like to acknowledge helpful discussions with and thank John Kent, Theofanis Sapatinas, Bernard Silverman, and Andy Wood. This research was sponsored by the MoD Corporate Research Programme and forms part of a programme on data fusion lead by Dr. R. H. Glendinning, QinetiQ ltd. GPN was also supported by EPSRC Advanced Research Fellowship AF/001664.

## A Proof of proposition 1

Assume  $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 I_n)$  and  $\varepsilon = W\mathbf{e}$ , where  $W$  is a unitary matrix. We use  $\text{Re}(\varepsilon) = (\varepsilon + \bar{\varepsilon})/2$  and  $\text{Im}(\varepsilon) = (\varepsilon - \bar{\varepsilon})/(2i)$ . Then

$$\begin{aligned} \text{cov}\{\text{Re}(\varepsilon), \text{Im}(\varepsilon)\} &= \frac{-1}{4i} \{\text{cov}(\varepsilon, \varepsilon) - \text{cov}(\varepsilon, \bar{\varepsilon}) + \text{cov}(\bar{\varepsilon}, \varepsilon) - \text{cov}(\bar{\varepsilon}, \bar{\varepsilon})\} \\ &= \frac{-\sigma^2}{4i} (W\bar{W}^T - WW^T + \bar{W}\bar{W}^T - \bar{W}W^T) \\ &= \frac{\sigma^2}{2} \{\text{Im}(I_n) - \text{Im}(WW^T)\} \end{aligned}$$

and noting that  $\text{Im}(I_n) = 0$  gives (2). Similar manipulations give (3) and (4).

## B Proof of Theorem 1

### B.1 Establishment of simple bound

Let  $X$  be a complex normal random variable with mean  $\mu$  and variance matrix  $\Sigma$ . The estimator  $\hat{\mu}(X)$  is a shrinkage rule if  $|\hat{\mu}(X)| \leq |X|$  for all  $X$ . If  $\hat{\mu}$  is hard thresholding then obviously (for some threshold  $t$ )

$$|\hat{\mu} - \mu|^2 = \begin{cases} |\mu|^2 & \text{if } \theta < t \\ |x - \mu|^2 & \text{else} \end{cases}$$

where  $\theta = X^T \Sigma^{-1} X$ . Hence we have the bound

$$|\hat{\mu} - \mu|^2 \leq \max\{|\mu|^2, |x - \mu|^2\} \leq |\mu|^2 + |x - \mu|^2 \quad (10)$$

as in Johnstone and Silverman (2002b) (henceforth J+S).

For a general phase preserving shrinkage rule it is possible to establish the following bound:

$$|\hat{\mu} - \mu|^2 \leq \max\{|2\mu|^2, |x - \mu|^2\} \leq |2\mu|^2 + |x - \mu|^2. \quad (11)$$

So, to bound the risk  $r_\Sigma(\mu, t) := E|\hat{\mu}(X) - \mu|^2$ , we must bound  $|\mu|^2$  and  $|x - \mu|^2$ .

### B.2 Risk bound for $|\mu| \geq 1$

Let us establish a value for  $E|x - \mu|^2$ . Let  $Z = X - \mu := Z_1 + iZ_2$  then  $Z$  is a complex normal distribution with mean 0 and variance matrix  $\Sigma$ .

$$E|X - \mu|^2 = E|Z|^2 = E(Z_1^2 + Z_2^2) = \text{var } Z_1 + \text{var } Z_2 = \Sigma_{11} + \Sigma_{22} := a_2.$$

Hence for  $|\mu| \geq 1$  we have

$$E|\hat{\mu} - \mu|^2 \leq |\mu|^2 + a_2 \leq (1 + a_2)|\mu|^2$$

as required.

### B.3 Risk bound for $|\mu| < 1$

For  $|\mu| < 1$  we take a more direct approach.

$$\begin{aligned}
r_{\Sigma}(\mu, t) &= E |XI(\Theta > t) - \mu|^2 \\
&= \int |xI(\theta > t) - \mu|^2 \phi_{\mu, \Sigma}(x) dx \\
&= \int_{\theta \geq t} |x - \mu|^2 \phi_{\mu, \Sigma}(x) dx + \int_{\theta < t} |\mu|^2 \phi_{\mu, \Sigma}(x) dx \\
&:= I_1 + I_2.
\end{aligned} \tag{12}$$

#### B.3.1 Bound for $I_2$

First we make the transformation  $Y = \Sigma^{-1/2}X \sim N(\Sigma^{-1/2}\mu, I)$ . Let  $\nu = \Sigma^{-1/2}\mu$ . Hence

$$I_2 = |\mu|^2 |\Sigma|^{1/2} \int_{y^T y < t} \phi_{\nu, I}(y) dy. \tag{13}$$

To bound this integral we again change coordinates. To explain this change examine Figure 3. We integrate by increasing  $h$  from  $-t$  to  $t$  and integrating from  $-b_h$  to  $b_h$ . Thus  $I_2$  becomes:

$$I_2 = |\mu|^2 |\Sigma|^{1/2} \int_{-t}^t \int_{-b_h}^{b_h} \phi_{\nu, I}(y) dy_1 dy_2,$$

where  $b_h^2 = t^2 - h^2$ . Hence

$$\begin{aligned}
I_2 &= |\mu|^2 |\Sigma|^{1/2} \int_{-t}^t \phi(y_2 - \nu_2) \int_{-b_{y_2}}^{b_{y_2}} \phi(y_1 - \nu_1) dy_1 dy_2 \\
&= |\mu|^2 |\Sigma|^{1/2} \int_{-t}^t \phi(y_2 - \nu_2) \{\Phi(b_{y_2} - \nu_1) - \Phi(-b_{y_2} - \nu_1)\} dy_2 \\
&\leq |\mu|^2 |\Sigma|^{1/2} \{\Phi(t - \nu_1) - \Phi(-t - \nu_1)\} \int_{-t}^t \phi(y_2 - \nu_2) dy_2.
\end{aligned}$$

The last line because  $t^2 \geq t^2 - y_2^2$ . Hence:

$$I_2 \leq |\mu|^2 |\Sigma|^{1/2} \{\Phi(t - \nu_1) - \Phi(-t - \nu_1)\} \{\Phi(t - \nu_2) - \Phi(-t - \nu_2)\}. \tag{14}$$

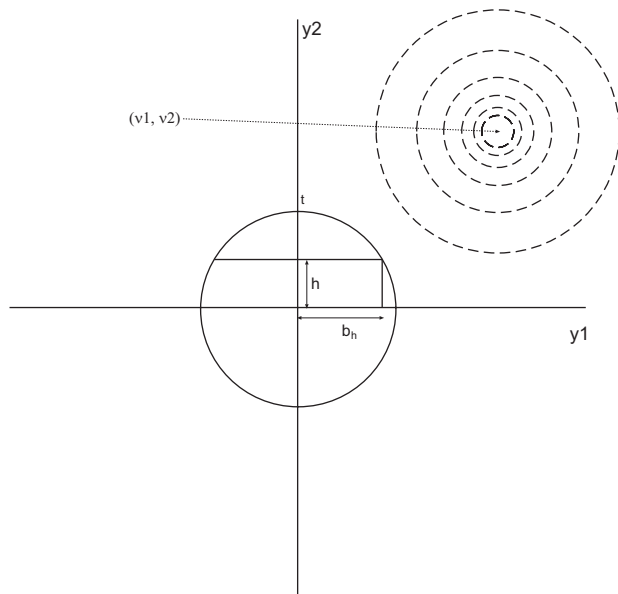


Figure 3: Setup for integral  $I_2$ . Dotted contours correspond to  $\phi_{\nu, I}(y)$  and solid line to the contour  $y^T y = t$

The last line of the above expression should be compared to the one in J+S that reads  $\Phi(t - \mu) - \Phi(-t - \mu)$ . Since  $\Phi : (-\infty, \infty) \rightarrow [0, 1]$  we must have  $I_2 \leq k_1 |\mu|^2$  for all  $t, \mu$  for the positive constant  $k_1 = |\Sigma|^{1/2}$ .

A far simpler route to this bound exists: just note that the integral in equation (13) is always strictly less than one for  $t < \infty$  since  $\phi$  is a density function. However, although this route establishes the bound it does not reveal the finer information given in (14) which is an analogous formula to that in J+S and also gives more information about the risk properties of the Downie-Silverman threshold.

### B.3.2 Bound for $I_1$

Recall the integral  $I_1$  is given by

$$I_1 = \int_{\theta > t} |x - \mu|^2 \phi_{\mu, \Sigma}(x) dx.$$

First, let us make the substitution  $y = x - \mu$ . Thus

$$I_1 = \int_{|(y+\mu)^T \Sigma^{-1}(y+\mu)| > t} |y|^2 \phi_{0, \Sigma}(y) dy. \quad (15)$$

Let  $\sigma$  and  $\tau$  be the largest and smallest eigenvalues of the positive semi-definite matrix  $\Sigma$ . Then, by examining a contour diagram of the densities for example, it can be shown that, for all  $x \in \mathbb{R}^2$ ,

$$\phi_{0, \Sigma}(x) \leq \sigma |\Sigma|^{-1/2} \phi_{0, \sigma I}(x). \quad (16)$$

Further, the ellipse described by the domain condition  $|(y + \mu)^T \Sigma^{-1}(y + \mu)| > t$  in integral (15) contains the disc

$$|(y + \mu)^T \tau I (y + \mu)| > t. \quad (17)$$

Putting together the inequality (16) with small ellipse (17) permits the following bound for  $I_1$

$$I_1 \leq \sigma |\Sigma|^{-1/2} \int_{\tau |(y+\mu)^T (y+\mu)| > t} |y|^2 \phi_{0, \sigma I}(y) dy =: \sigma |\Sigma|^{-1/2} I_3, \text{ say.}$$

We now bound integral  $I_3$ , considering separately the two cases (I):  $|\mu|^2 > t/\tau$  and (II):

$|\mu|^2 < t/\tau$ . Case I is illustrated in Figure 4 where the contour around  $(-\mu_1, -\mu_2)$  does not contain the origin; in Case II it does.

## B.4 Bounding $I_3$ for Case I

### B.4.1 Bounding $I_3$

Figure 4 illustrates the density and the region of integration for integral  $I_3$ . Let  $\alpha_1$  and  $\alpha_2$  be defined as in the caption of Figure 4.

Let the polar representations of  $y = (y_1, y_2)$  and  $\mu = (\mu_1, \mu_2)$  be given by  $y = re^{i\theta}$  and  $\mu = me^{i\phi}$ . We will convert  $I_3$  into polar coordinates:  $y \rightarrow (r, \theta)$ , recall that the Jacobian of this transformation is  $r$ . First, the domain condition becomes:

$$\begin{aligned} |(y + \mu)^T(y + \mu)| > t/\tau &\equiv |y|^2 + 2\mu^T y + |\mu|^2 > t/\tau \\ &\equiv r^2 + 2r\xi(\theta) + m^2 - t/\tau > 0, \end{aligned} \quad (18)$$

where  $\xi(\theta) = \mu_1 \cos(\theta) + \mu_2 \sin(\theta)$ . To compute the forthcoming polar integral we need to know which values of  $(r, \theta)$  satisfy condition (18). Figure 4 shows that the condition is satisfied for all angles  $\theta \in [\alpha_1, \alpha_2]$  for all  $r \in [0, \infty)$ . However, the condition is only satisfied sometimes for  $\theta \in [\alpha_2, \alpha_1]$  for some values of  $r$ .

The  $\alpha_i$  angles are determined as follows. Write the left-hand side of the inequality in (18) as the polynomial  $p(r) = ar^2 + br + c$  where

$$a = 1, b = 2\xi(\theta), c = m^2 - t/\tau. \quad (19)$$

Since  $a > 0$  the polynomial is convex with turning point satisfying  $p'(r) = 2r + b = 0$  i.e. at  $r = -b/2$ . The value of  $p(r)$  at this minimum is  $p(-b/2) = c - b^2/4$ . Also, the polynomial has complex roots if and only if  $b^2 - 4c < 0$  which would imply  $p(-b/2) = c - b^2/4 > 0$ . Summarizing: the polynomial  $p(r)$  is always positive when the roots are complex. What

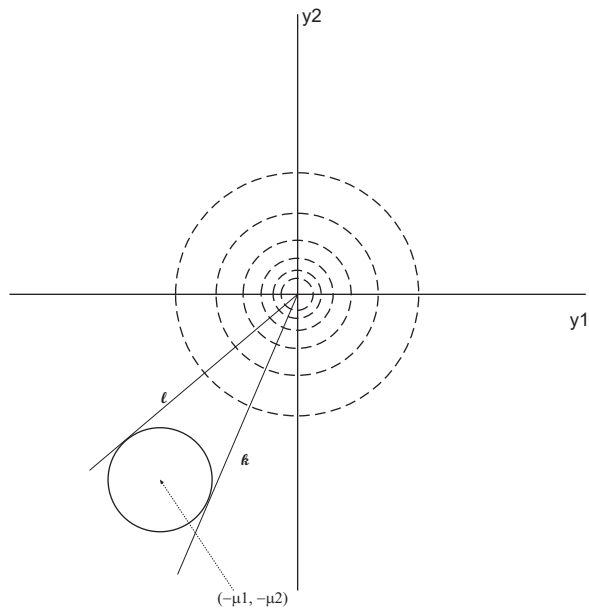


Figure 4: Setup for integral  $I_3$  for case I:  $m^2 > t/\tau$ . Dotted contours correspond to  $\phi_{0,\sigma I}(y)$  and solid line to the contour  $|(y + \mu)^T(y + \mu)| = t/\tau$ . Line  $\ell$  is at angle  $\alpha_2$  and line  $k$  at angle  $\alpha_1$ . Angle  $\phi$  is the angle of the vector  $\mu$ .

are the conditions for the roots to be complex? This occurs when

$$\begin{aligned}
b^2 - 4c < 0 &\equiv 4\xi^2(\theta) < 4(m^2 - t/\tau) \\
&\equiv \{\mu_1 \cos(\theta) + \mu_2 \sin(\theta)\}^2 < m^2 - t/\tau \\
&\equiv \{m \cos(\phi) \cos(\theta) + m \sin(\phi) \sin(\theta)\}^2 < m^2 - t/\tau \\
&\equiv m^2 \cos^2(\theta - \phi) < m^2 - t/\tau \\
&\equiv |\cos(\theta - \phi)| < \sqrt{1 - t/(m^2\tau)}. \tag{20}
\end{aligned}$$

It will be useful later to note that  $b^2 = 4m^2 \cos^2(\theta - \phi)$ . Also note that since we are considering case I the argument of the square root in (20) is always positive. Let  $\alpha$  be the angle such that  $\theta = \phi + \alpha$  causes the inequality in (20) to be an equality. Then  $\alpha_1 = \phi + \alpha$ ,  $\alpha_2 = \phi - \alpha$ . Note that for complex roots as long as  $\theta$  is in the range  $[\alpha_1, \alpha_2]$  condition (18) is always satisfied for all  $r \in [0, \infty)$ . Hence we divide  $I_3$  into two further polar integrals: let  $F(r) = r^3 \exp(-r^2/(2\sigma))$  then:

$$2\pi\sigma I_3 = \int_{\alpha_1}^{\alpha_2} \int_0^\infty F(r) dr d\theta + \int_{\alpha_2}^{\alpha_1} \int_{r \in R} F(r) dr d\theta = I_4 + I_5, \tag{21}$$

where  $R$  is the domain (to be precisely determined below) in the region between  $\alpha_2$  and  $\alpha_1$  where the condition (18) is satisfied.

#### B.4.2 Evaluating $I_4$

The double integral  $I_4$  can be easily calculated using integration by parts after noting that the inner integral does not depend on  $\theta$ :

$$\begin{aligned}
I_4 &= -(\alpha_2 - \alpha_1)\sigma \int_0^\infty r^2 \cdot (-r/\sigma) \exp(-r^2/(2\sigma)) dr \\
&= -(\alpha_2 - \alpha_1)\sigma \left\{ [r^2 \exp(-r^2/(2\sigma))]_0^\infty + 2\sigma \int_0^\infty (-r/\sigma) \exp(-r^2/(2\sigma)) dr \right\} \\
&= -2\sigma^2(\alpha_2 - \alpha_1) [\exp(-r^2/(2\sigma))]_0^\infty \\
&= 2\sigma^2(\alpha_2 - \alpha_1) \\
&= 2\sigma^2(2\pi - 2\alpha). \tag{22}
\end{aligned}$$

### B.4.3 Bounding $I_5$

Now suppose that  $\theta$  is in the region between  $\alpha_2$  and  $\alpha_1$  (the small segment between lines  $\ell$  and  $k$  on Figure 4). Then, clearly, the roots of polynomial  $p(r)$  as defined in the previous section are real: call them  $r_i$  with  $r_1 < r_2$ . From the diagram it can be seen that condition (18) is satisfied when  $0 < r < r_1$  and  $r_2 < r < \infty$ . It is also vital to notice that the  $r_i$  depend on  $\theta$ . By similar integration by parts computations as in the integration for  $I_4$  above it can be shown that

$$\begin{aligned} S(r_1, r_2) &:= \int_0^{r_1} + \int_{r_2}^{\infty} F(r) dr \\ &= \sigma \left\{ 2\sigma + (r_2^2 + 2\sigma)e^{-r_2^2/2\sigma} - (r_1^2 + 2\sigma)e^{-r_1^2/2\sigma} \right\} \end{aligned}$$

Hence

$$I_5 = \int_{\alpha_2}^{\alpha_1} S\{r_1(\theta), r_2(\theta)\} d\theta. \quad (23)$$

The following Lemma enables us to bound  $I_5$ .

#### Lemma 1

$$S(r_1, r_2) \leq 2\sigma^2 + \sigma(2\sigma + b^2/2 - c)T(t, \mu)\sigma, \quad (24)$$

where  $b$  and  $c$  are defined in (19) and where

$$T(t, \mu) = 2 \exp \left\{ (t/\tau - |\mu|^2)/2\sigma \right\} \sinh \left\{ |\mu|(t/\tau)^{1/2}/\sigma \right\}$$

does not depend on  $\theta$ .

**Proof:** See section B.7

The only quantity in (24) depending on  $\theta$  is  $b^2$ . Therefore, in order to bound  $I_5$  we need

to know

$$\begin{aligned}
I_6 &= \int_{\alpha^4}^{\alpha_1} b^2 d\theta \\
&= 4m^2 \int_{\alpha^4}^{\alpha_1} \cos^2(\theta - \phi) d\theta \\
&= 2m^2 \int_{\alpha_2}^{\alpha_1} 1 + 2 \cos\{2(\theta - \phi)\} d\theta \\
&= 2m^2 \{2\alpha + \sin(2\alpha)\}.
\end{aligned} \tag{25}$$

#### B.4.4 Reassembling bounds

We now reassemble the results and bounds from Sections B.3.1 to B.4.3. Formulae (23) to (25) give

$$\begin{aligned}
I_5 &= \int_{\alpha_2}^{\alpha_1} S\{r_1(\theta), r_2(\theta)\} d\theta \\
&\leq 2\sigma^2(\alpha_1 - \alpha_2) + T(t, \mu)\sigma \{(2\sigma - c)(\alpha_1 - \alpha_2) + \frac{1}{2}I_6\} \\
&= 4\alpha\sigma^2 + \sigma T(t, \mu) [2\alpha(2\sigma - c) + m^2 \{2\alpha + \sin(2\alpha)\}]
\end{aligned} \tag{26}$$

Now using the formula for  $I_3$  in (21), formula (22) for  $I_4$  and the bound (26) above we obtain:

$$\begin{aligned}
2\pi\sigma I_3 &\leq 2\sigma^2(2\pi - 2\alpha) + \sigma T(t, \mu) [2\alpha(2\sigma - c) + m^2 \{2\alpha + \sin(2\alpha)\}] \\
2\pi I_3 &\leq 4\sigma\pi + T(t, \mu) [2\alpha(2\sigma - c) + m^2 \{2\alpha + \sin(2\alpha)\}] \\
&\leq 4\sigma\pi + T(t, \mu) (2\sigma\pi + m^2 + t\pi/\tau),
\end{aligned}$$

since  $\alpha \in [0, \pi/2)$ .

Now since  $t/\tau \leq m^2 < 1$  we have

$$\begin{aligned}
T(t, \mu) &\leq 2 \exp\{-m^2/2\sigma\} \exp\{t/(2\sigma\tau)\} \sinh(1/\sigma) \\
&= 2\sqrt{2\pi\sigma}\phi_{0,\sigma}(m) \exp\{t/(2|\Sigma|)\} \sinh(1/\sigma).
\end{aligned}$$

Hence

$$I_3 \leq 2\sigma\pi + \sqrt{2\sigma/\pi}\phi_{0,\sigma}(m) \exp\{t/(2|\Sigma|)\} \sinh(1/\sigma) (m^2 + 2\pi\sigma + t\pi/\tau).$$

### B.5 Bounding $I_3$ for Case II

We now consider case II where  $m^2 < t/\tau$ . For integral  $I_3$  we again change to polar coordinates. If  $(r, \theta)$  are the coordinates of the contour  $|(y + \mu)^T(y + \mu)| = t/\tau$  then clearly  $r$  takes its minimum and maximum value at  $\phi + \pi$  and  $\phi$  respectively. The minimum value of  $r$  at  $\theta = \phi + \pi$  is  $r^* = \sqrt{t/\tau} - m$ . We obtain an upper bound for  $I_3$  by integrating over the area that is greater than the *circular* contour with radius  $r^*$  (which, of course, does not depend on  $\theta$ ). Hence

$$\begin{aligned} I_3 &\leq \int_{y^T y > r^*} |y|^2 \phi_{0,\sigma} I(y) dy \\ &= (2\pi\sigma)^{-1} \int_0^{2\pi} \int_{r^*}^{\infty} r^3 \exp\left(-\frac{r^2}{2\sigma}\right) dr d\theta \\ &= \exp\left(-\frac{r^{*2}}{2\sigma}\right) (r^{*2} + 2\sigma) \\ &\leq (2\pi\sigma)^{1/2} \phi_{0,\sigma^{1/2}}(m - \sqrt{t/\tau}) \left\{ \left(\sqrt{t/\tau} - m\right)^2 + 2\sigma \right\}. \end{aligned}$$

### B.6 Risk bound for $r_\Sigma(\mu, t)$

We now put everything together to obtain the risk bound for  $r_\Sigma(\mu, t)$  for  $|\mu| < 1$ .

For case I:  $m^2 > t/\tau$ :

$$\begin{aligned} r_\Sigma(\mu, t) &= I_1 + I_2 \\ &\leq \sigma|\Sigma|^{-1/2} I_3 + |\Sigma|^{1/2} |\mu|^2 \\ &\leq \left(\frac{2\sigma^3}{\pi|\Sigma|}\right)^{1/2} \phi_{0,\sigma}(|\mu|) \exp\left(\frac{t}{2|\Sigma|}\right) \sinh\left(\frac{1}{\sigma}\right) (|\mu|^2 + 2\sigma\pi + t\pi/\tau) \\ &\quad + \frac{2\sigma^2}{|\Sigma|^{1/2}} + |\Sigma|^{1/2} |\mu|^2 \\ &\sim K_1 + K_2 \phi_{0,\sigma}(|\mu|) \exp(tK_3) (|\mu|^2 + tK_4 + K_5) + |\mu|^2 K_6. \end{aligned}$$

The last line of this formula includes  $\tau$  and  $\sigma$  as part of the constants so the form in terms

of  $t$  and  $\mu$  can be discerned.

For case II:  $m^2 < t/\tau$ :

$$\begin{aligned}
r_\Sigma(\mu, t) &= I_1 + I_2 \\
&\leq \sigma|\Sigma|^{-1/2}I_3 + |\Sigma|^{1/2}|\mu|^2 \\
&\leq K_7\phi_{0,\sigma^{-1/2}}\left(|\mu| - \sqrt{t/\tau}\right) \left\{ \left(|\mu| - \sqrt{t/\tau}\right)^2 + K_8 \right\} + |\mu|^2 K_6.
\end{aligned}$$

## B.7 Proof of Lemma 1

We prove Lemma 1. First, note that if  $r_i$  is a root of  $r^2 + br + c = 0$  then it is immediate that  $r_i^2 = -br_i - c$ . For notational simplicity let  $e_i = \exp(-r_i^2/2\sigma)$  in the following.

$$\begin{aligned}
\sigma^{-1}S(r_1, r_2) - 2\sigma &= (-br_2 - c + 2\sigma)e_2 - (-br_1 - c + 2\sigma)e_1 \\
&= (2\sigma - c)(e_2 - e_1) - (b/2)(-b + \sqrt{b^2 - 4c})e_2 \\
&\quad + (b/2)(-b - \sqrt{b^2 - 4c})e_1 \\
&= (2\sigma + b^2/2 - c)(e_2 - e_1) - b/2\sqrt{b^2 - 4c}(e_2 + e_1) \quad (27)
\end{aligned}$$

The term  $e_2 - e_1$  in (27) can be written:

$$\begin{aligned}
e_2 - e_1 &= \exp\left(\frac{br_2 + c}{2\sigma}\right) - \exp\left(\frac{br_1 + c}{2\sigma}\right) \\
&= \exp(c/2\sigma) \left\{ \exp\left(\frac{br_2}{2\sigma}\right) - \exp\left(\frac{br_1}{2\sigma}\right) \right\} \\
&= \exp(c/2\sigma) \left\{ \exp\left[\frac{b}{4\sigma}(-b + \sqrt{b^2 - 4c})\right] - \exp\left[\frac{b}{4\sigma}(-b - \sqrt{b^2 - 4c})\right] \right\} \\
&= \exp(c/2\sigma - b^2/4\sigma) \left\{ \exp\left(\frac{b\sqrt{b^2 - 4c}}{4\sigma}\right) - \exp\left(-\frac{b\sqrt{b^2 - 4c}}{4\sigma}\right) \right\} \\
&= 2 \exp\{(c - b^2/2)/2\sigma\} \sinh\left(\frac{b\sqrt{b^2 - 4c}}{4\sigma}\right). \quad (28)
\end{aligned}$$

We consider the arguments of exp and sinh in (28) in more detail. The argument of exp is

$$\begin{aligned}
c - b^2/2 &= m^2 - t/\tau - 2m^2 \cos^2(\theta - \phi) \\
&= m^2\{1 - 2\cos^2(\theta - \phi)\} - t/\tau, \quad (29)
\end{aligned}$$

the expansion of  $b^2$  into a  $\cos^2$  term can be found in equation (20). By considering the  $\cos^2$  term in (29) it can be shown that the argument  $c - b^2/2$  takes its maximum value when  $\theta = \phi + \pi/2$ . In this constrained context (for real roots) the furthest  $\theta$  can be from  $\phi$  is  $\alpha$ . Hence the maximum value of the argument is

$$\begin{aligned} m^2\{1 - 2\cos^2(\alpha)\} - t/\tau &= m^2\{1 - 2(1 - t/(m^2\tau))\} - t/\tau \\ &= t/\tau - m^2. \end{aligned}$$

The argument of sinh in (28) is (without the  $4\sigma$ )

$$b\sqrt{b^2 - 4c} = 4m \cos(\theta - \phi) \{m^2 \cos^2(\theta - \phi) - (m^2 - t/\tau)\}^{1/2} \quad (30)$$

As  $\theta$  moves away from  $\phi$  the term  $\cos(\theta - \phi)$  decreases (although it is always positive) and hence the maximum value of the argument must be achieved when  $\theta = \phi$  and takes the value  $4m(t/\tau)^{1/2}$ .

The minimum of  $\cos(\theta - \phi)$  occurs when  $\theta - \phi = \alpha$ , at which point (30) is zero, so  $b\sqrt{b^2 - 4c}$  is non-negative, as is  $e_2 + e_1$ . Therefore the quantity subtracted in (27) has a lower bound of zero and can be discarded. Defining  $T(t, \mu)$  to be (28) with the arguments replaced by the maximum values determined in the previous two paragraphs completes the proof.

## References

- Abramovich, F., Bailey, T. C. and Sapatinas, T. (2000) Wavelet analysis and its statistical applications. *Statistician* **49**, 1–29.
- Abramovich, F. and Benjamini, Y. (1996) Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis* **22**, 351–361.
- Abramovich, F., Besbeas, P. and Sapatinas, T. (2002) Empirical Bayes approach to block wavelet function estimation. *Computational Statistics and Data Analysis* **39**, 435–451.
- Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998) Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc. B* **60**, 725–749.

- Barber, S. and Nason, G. P. (2003) Simulations comparing thresholding methods using real and complex wavelets. *Research Report 03:07*, Statistics Group, University of Bristol.
- Cai, T. T. and Silverman, B. W. (2001) Incorporating information on neighbouring coefficients into wavelet estimation. *Sankhya B* **63**, 127–148.
- Chipman, H., Kolaczyk, E. and McCulloch, R. (1997) Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Ass.* **92**, 1413–1421.
- Clyde, M. and George, E. I. (2000) Flexible empirical Bayes estimation for wavelets. *J. R. Statist. Soc. B* **62**, 681–698.
- Coifman, R. R. and Donoho, D. L. (1995) Translation-invariant de-noising. In *Wavelets and Statistics* (eds A. Antoniadis and G. Oppenheim) pp 125–150. New York: Springer-Verlag.
- Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: SIAM.
- Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.
- Donoho, D. L. and Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.* **90**, 1200–1224.
- Downie, T. R. and Silverman, B. W. (1998) The discrete multiple wavelet transform and thresholding methods. *IEEE Transactions on Signal Processing* **46**, 2558–2561.
- Fourdriner, D., Philippe, A. and Robert, C. P. (2000) Estimation of a non-centrality parameter under Stein-type losses. *Journal of Statistical Planning and Inference* **87**, 43–54.
- Johnstone, I. M. (2000) Chi-square inequalities. In *State of the Art in Probability & Statistics* (eds M. de Gunst, C. Klassen and A. van der Vaart) 399–418. Beachwood: Institute of Mathematical Statistics.
- Johnstone, I. M. and Silverman, B. W. (2002a) Empirical Bayes selection of wavelet thresholds. *Research Report 02:17*, Statistics Group, University of Bristol.
- Johnstone, I. M. and Silverman, B. W. (2002b) Needles and hay in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Research Report 02:15*, Statistics Group, University of Bristol.
- Kohn, R., Marron, J. S. and Yau, P. (2000) Wavelet estimation using Bayesian basis selection and basis averaging. *Statistica Sinica* **10**, 109–128.
- Lawton, W. (1993) Applications of complex valued wavelet transforms to subband decom-

- position. *IEEE Transactions on Signal Processing* **41**, 3566–3568.
- Lina, J.-M. (1997) Image processing with complex Daubechies wavelets. *Journal of Mathematical Imaging and Vision* **7**, 211–223.
- Lina, J.-M. and MacGibbon, B. (1997) Non-linear shrinkage estimators with complex Daubechies wavelets. In *Proceedings of SPIE*, **3169**, pp 67–79.
- Lina, J.-M. and Mayrand, M. (1995) Complex Daubechies wavelets. *Applied and Computational Harmonic Analysis* **2**, 219–229.
- Lina, J.-M., Turcotte, P. and Goulard, B. (1999) Complex dyadic multiresolution analysis. In *Advances in Imaging and Electron Physics*, volume 109. Academic Press.
- Mallat, S. G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.
- Nason, G. P. (1996) Wavelet shrinkage using cross-validation. *J. R. Statist. Soc. B* **58**, 463–479.
- Nason, G. P. and Silverman, B. W. (1994) The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics* **3**, 163–191.
- Nason, G. P. and Silverman, B. W. (1995) The stationary wavelet transform and some statistical applications. In *Wavelets and Statistics* (eds A. Antoniadis and G. Oppenheim) pp 281–229. New York: Springer-Verlag.
- O’Hagan, A. (1994) *Bayesian Inference: Kendall’s Advanced Theory of Statistics, vol. 2B*. London: Edward Arnold
- Sardy, S. (2000) Minimax threshold for denoising complex signals with waveshrink. *IEEE Transactions on Signal Processing* **48**, 1023–1028.
- Saxena, K. M. L. and Alam, K. (1982) Estimation of the non-centrality parameter of a Chi squared distribution. *Ann. Statist.* **10**, 1012–1016.
- Small, C. G. (1990) A survey of multidimensional medians. *International Statistical Review* **58**, 263–277.
- Vidakovic, B. (1999) *Statistical Modeling by Wavelets*. New York: Wiley.
- Zaroubi, S. and Goelman, G. (2000) Complex denoising of MR data via wavelet analysis: application for functional MRI. *Magnetic Resonance Imaging* **18**, 59–68.