Wavelet Decomposition Approaches to Statistical Inverse Problems

F. Abramovich[1] and B. W. Silverman[2]

[1]Department of Statistics & Operations Research,
Raymond & Beverly Sackler Faculty of Exact Sciences,
University of Tel Aviv, Ramat Aviv 69978, Israel.
[2]Department of Mathematics,
University of Bristol, Bristol BS8 1TW, UK.

17 October 1996

Abstract

A wide variety of scientific settings involve *indirect* noisy measurements where one faces a linear inverse problem in the presence of noise. Primary interest is in some function $f(t)$ but the data is accessible only about some transform $(Kf)(t)$, where K is some linear operator, and $(Kf)(t)$ is in addition corrupted by noise. The usual linear methods for such inverse problems, for example those based on singular value decompositions, do not perform satisfactorily when the original function $f(t)$ is spatially inhomogeneous. One alternative that has been suggested is the wavelet–vaguelette decomposition method, based on the expansion of the unknown $f(t)$ in wavelet series.

The vaguelette–wavelet decomposition method proposed in this paper is also based on wavelet expansion. In contrast to wavelet–vaguelette decomposition, the observed data are expanded directly in wavelet series. Using exact risk calculations, the performances of the two wavelet-based methods are compared with one another and with singular value decomposition methods, in the context of the estimation of the derivative of a function observed subject to noise. A result is proved demonstrating that, with a suitable universal threshold somewhat larger than that used for standard denoising problems, both wavelet-based approaches have an ideal spatial adaptivity property.

Some key words: Exact risk analysis; Near-minimax estimation; Singular Value Decomposition; Spatially adaptive estimation; Statistical linear inverse problems; Vaguelettes; Wavelets.

1 Introduction

In a wide variety of statistical problems direct observations on the object of interest are inaccessible and one faces the problem of estimation and inference from indirect experimental data. To be specific, suppose we wish to estimate an unknown function $f(t)$ but we can observe data only about $(Kf)(t)$, where K is some linear operator. Suppose also that the data are observed at discrete points t_i and are corrupted by noise, so that the observed data $y(t)$ are

$$y(t) = (Kf)(t) + \varepsilon(t), \tag{1}$$

where $\varepsilon(t)$ is a white noise process. We use the term *statistical linear inverse problem* for the problem of estimating f from noisy data y in the model (1). For most such inverse problems one cannot recover f simply as $\hat{f} = K^{-1}y$ in practice, either because the inverse operator K^{-1} does not exist at all, or because it is an unbounded operator, which means that small changes in y would cause large changes in $K^{-1}y$. Such inverse problems are called ill-posed, and are usually treated by applying some linear regularization procedure, often based on a singular value decomposition; see Tikhonov & Arsenin (1977) for general theory and O'Sullivan (1986) for a more specifically statistical discussion. Turning to non-linear methods, Donoho (1995) proposed the *wavelet–vaguelette decomposition* which works by expanding the function f in a wavelet series, constructing a corresponding vaguelette series for Kf, and then estimating the coefficients using a suitable thresholding approach.

As an alternative to wavelet–vaguelette decomposition we propose the use of a *vaguelette–wavelet decomposition* where Kf is expanded in a wavelet series. The corresponding wavelet coefficients of Kf are estimated by thresholding the empirical wavelet coefficients of the data. Mapping them back into the vaguelette expansion in the original space yields the vaguelette–wavelet decomposition estimator of f. Some important conceptual aspects of vaguelette–wavelet decomposition are discussed in the conclusions in Section 6.

We shall use as a test problem the estimation of the derivative of a function g on the basis of observations of g itself. This is an important statistical problem: one typical context in which it arises is in the study of growth curves. For example, longitudinal data on the height of a child may be used to make inferences about the child's growth rate. Estimation of derivatives is of clear importance in economics, for example when estimating inflation rates from prices, and in many other fields of application. In order to pose derivative estimation as an inverse problem, let K be the integration operator. The problem is then that of estimating a function f from noisy observations of Kf, so that in the notation of equation (1) the function f corresponds to the derivative g' of primary interest.

In a numerical study, the two wavelet-based methods yield similar results, generally better than those obtained by a singular value decomposition approach. Interestingly, the ideal threshold levels are somewhat larger than those appropriate for the estimation of the directly observed function $g = Kf$. We study the theoretical ground for these phenomena, and prove that both wavelet-based methods for linear inverse problems have an ideal adaptivity property. The theory indicates the amount by which the thresholds for function estimation from direct data should be inflated for inverse problems.

The paper is set out as follows. In Section 2 we review the singular value decomposition and wavelet–vaguelette decomposition approaches, providing the framework within which the vaguelette–wavelet decomposition method is then defined and discussed. In Section 3, exact risk formulae are obtained for the various approaches, and these are used in Section 4 to carry out a comparison on several examples without any need for simulation. In Section

5, the theoretical minimax properties of the estimators are explored, and a result proved showing that, with an appropriate definition of a universal threshold, both wavelet-based methods have an ideal adaptivity property. The general conclusions are set out in Section 6.

2 Different approaches to statistical linear inverse problems

2.1 Singular value decompositions and Fourier series

The underlying idea of singular value decomposition methods is the use of a pseudo-inverse operator $(K^*K)^{-1}K^*$, where K^* is the adjoint operator to the operator K. The unknown f is expanded in a series of eigenfunctions e_j of the self-adjoint operator K^*K as

$$f = \sum_j \gamma_j^{-1} \langle Kf, h_j \rangle e_j = \sum_j c_j e_j, \tag{2}$$

say, where γ_j^2 are the eigenvalues of K^*K and $h_j = Ke_j/\|Ke_j\|$. We use the notation $\langle \cdot, \cdot \rangle$ for the standard inner product in L_2. Ill-posed problems are characterized by the fact that the eigenvalues γ_j^2 tend to zero.

In the presence of noise in the data we can replace Kf in (2) by y and define $\hat{c}_j = \gamma_j^{-1}\langle y, h_j \rangle e_j$. The *truncated singular value decomposition* estimator of f is then defined to be

$$\hat{f}_M^{\mathrm{SVD}} = \sum_{j=1}^{M} \hat{c}_j e_j, \tag{3}$$

for some truncation point M. An intuitive reason for the truncation is to avoid the division by near-zero values γ_j that would be necessary if all the $\hat{c}_j$ had to be calculated. Singular value decomposition has been widely used in practice in various theoretical and applied problems. Moreover, Johnstone & Silverman (1990, 1991) showed that a properly chosen truncated singular value decomposition estimator is asymptotically the best estimator, in a certain minimax sense, over classes of functions that display homogeneous variation.

Applying truncated singular value decomposition we essentially assume that f has a parsimonious representation in terms of the functions e_j, in that the coefficients in (2) decay rapidly to zero, so that only the first few coefficients in (3) convey any information about the underlying signal. However, this basis is defined entirely by the operator K and ignores the specific physical nature of the problem under study. For example, for stationary operators the corresponding eigenfunctions generate a Fourier sine and cosine basis. Fourier series are appropriate for representation of smooth spatially homogeneous functions, but do not provide a parsimonious approximation of inhomogeneous signals which are smooth in some regions while having rapid local changes in others. Thus, for such operators the use of singular value decomposition inherently restricts one within the class of spatially homogeneous functions.

2.2 The wavelet–vaguelette decomposition

In response to these limitations of the singular value decomposition approach, Donoho (1995) proposed the *wavelet–vaguelette decomposition* method, which depends on expanding the function f as a wavelet series. Wavelet series are generated by translations and dilations of a single fixed function ψ, called the mother wavelet: $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$,

$j, k \in \mathbb{Z}$. Examples of mother wavelets can be found in Chui (1992) and Daubechies (1992). Wavelets can be easily calculated and are localized in both the time and frequency domains, and allow parsimonious representation of a wide class of functions. Choosing the mother wavelet with appropriate regularity properties one can generate not only an orthonormal basis in $L_2(\mathbb{R})$ but unconditional bases in a wide range of more specific spaces corresponding to varying degrees and kinds of smoothness; see Section 5 below for more detailed discussion.

In the wavelet–vaguelette decomposition approach, the unknown function f is represented in terms of a wavelet expansion

$$f = \sum_j \sum_k \langle f, \psi_{jk} \rangle \psi_{jk}.$$

Let $\Psi_{jk} = K\psi_{jk}$. The crucial point of wavelet–vaguelette decomposition is that for some operators K there exist constants $\tilde{\beta}_{jk}$ such that the set of scaled functions $v_{jk} = \Psi_{jk}/\tilde{\beta}_{jk}$ form a Riesz basis in L_2 norm, that is there exist two constants $0 < A \leq B < \infty$ such that

$$A \sum_j \sum_k c_{jk}^2 \leq \| \sum c_{jk} v_{jk} \|^2 \leq B \sum_j \sum_k c_{jk}^2 \tag{4}$$

for all square summable sequences $\{c_{jk}\}$. The functions v_{jk} are called *vaguelettes*.

Obviously, only special operators K satisfy (4), but the condition does hold, for example, for *homogeneous* operators, which, for all t_0, satisfy $K[f\{a(t-t_0)\}] = a^{-\alpha}(Kf)\{a(t-t_0)\}$ for some α. The constant α is called the *index* of the operator. Examples of such operators include integration, fractional integration and, in the two-dimensional case which we will not consider in any detail in this paper, the Radon transform. For homogeneous operators $\tilde{\beta}_{jk} = C_\beta 2^{-j\alpha}$, and the corresponding vaguelettes v_{jk} are translations and dilations of a single mother function, but, unlike wavelets, are not mutually orthogonal. The property (4) also holds for various convolution operators; see Donoho (1995). In what follows we mainly consider operators K satisfying (4); however some comments about the relevance of the method we propose for more general operators are made at the end of Section 2.3 below.

Provided the wavelet basis ψ_{jk} is chosen appropriately, any function g in the range of K can be expanded in a vaguelette series as

$$g = \sum_j \sum_k \langle g, u_{jk} \rangle v_{jk},$$

where (u_{jk}) is a dual vaguelette basis satisfying $K^* u_{jk} = \tilde{\beta}_{jk} \psi_{jk}$. The dual bases (u_{jk}) and (v_{jk}) are biorthogonal, i.e. $\langle v_{jk}, u_{lm} \rangle = \delta_{jl}\delta_{km}$. Thus if we observed the signal Kf without noise, we could expand it in a vaguelette series:

$$Kf = \sum_j \sum_k \langle Kf, u_{jk} \rangle v_{jk}$$

and then recover the original function f as

$$f = \sum_j \sum_k \langle Kf, u_{jk} \rangle \tilde{\beta}_{jk}^{-1} \psi_{jk} = \sum_j \sum_k \langle Kf, \tilde{\Psi}_{jk} \rangle \psi_{jk}, \tag{5}$$

where $\tilde{\Psi}_{jk} = u_{jk}/\tilde{\beta}_{jk}$ and, hence, $K^* \tilde{\Psi}_{jk} = \psi_{jk}$. The formula (5) is a key formula for the wavelet–vaguelette decomposition method.

In the case of noisy data, we expand the observed signal y in terms of vaguelettes, with coefficients $\hat{b}_{jk} = \langle y, \tilde{\Psi}_{jk} \rangle$ which satisfy

$$\hat{b}_{jk} = b_{jk} + w_{jk}, \tag{6}$$

where $b_{jk} = \langle Kf, \tilde{\Psi}_{jk} \rangle$ are the noiseless vaguelette coefficients from (5) and $w_{jk} = \langle \varepsilon, \tilde{\Psi}_{jk} \rangle$ are the vaguelette decomposition of a white noise.

From (6), $\hat{b}_{jk} \sim N(b_{jk}, \sigma_0^2 \|\tilde{\Psi}_{jk}\|^2)$ for some σ_0^2. Construct rescaled coefficients $\hat{b}^0_{jk} = \hat{b}_{jk} / \|\tilde{\Psi}_{jk}\|$ which will all have the same variance σ_0^2. Extraction of the important $\hat{b}^0_{jk}$ is then based on the idea that only the 'large' $|\hat{b}^0_{jk}|$ contribute to the real signal, and can be naturally performed by thresholding, applying the soft threshold function

$$\delta_\lambda(x) = \mathrm{sign}(x)(|x| - \lambda)_+$$

or the hard threshold function

$$\delta_\lambda(x) = x \text{ if } |x| > \lambda, \; 0 \text{ otherwise}$$

for some threshold value $\lambda \geq 0$. It is important to note that the linear weighting of eigenvalues that characterizes linear singular value decomposition methods is replaced here by the nonlinear thresholding of vaguelette coefficients and, hence, the corresponding wavelet–vaguelette decomposition estimator is nonlinear. Mapping the thresholded coefficients back into the wavelet expansion in the original space yields the resulting wavelet–vaguelette decomposition estimator $\hat{f}_\lambda^{\mathrm{WVD}}$:

$$\hat{f}_\lambda^{\mathrm{WVD}} = \sum_j \sum_k \|\tilde{\Psi}_{jk}\| \delta_\lambda(\hat{b}^0_{jk}) \psi_{jk}. \tag{7}$$

In particular, for homogeneous operators the $\tilde{\Psi}_{jk}$ will be multiples of translations and dilations of a single mother vaguelette $\tilde{\Psi}_{00}$: $\tilde{\Psi}_{jk} = 2^{j\alpha} 2^{j/2} \tilde{\Psi}_{00}(2^j t - k)$. This makes all computations substantially easier. The norms $\|\Psi_{jk}\|$ are equal within each level j and, therefore, although the variances of the vaguelette coefficients $\hat{b}_{jk}$ on the j-th level increase like $2^{2j\alpha}$ as a direct consequence of the ill-posedness of the inverse problem, within each given level j they all have the same variance. Therefore (7) is exactly equivalent to the level-dependent thresholding of the coefficients $\hat{b}_{jk}$ using thresholds λ_j proportional to $2^{\alpha j}$. Kolaczyk (1996) considers in details the wavelet–vaguelette decomposition algorithms for the Radon transform operator which arises, for example, in positron emission tomography, a problem also considered by Johnstone and Silverman (1990).

The wavelet–vaguelette decomposition estimator, provided with optimally chosen threshold λ, has attractive theoretical properties, especially for spatially inhomogeneous functions f; for details see Section 5 below.

2.3 The vaguelette–wavelet decomposition

A natural alternative to wavelet–vaguelette decomposition is the *vaguelette–wavelet decomposition*: expand the observed data y in terms of wavelets, threshold the resulting wavelet coefficients appropriately, and then map back by K^{-1} to obtain an estimate of f in terms of vaguelette series. Thus, it is Kf, rather than f, that is expanded in a wavelet series.

Suppose we have a wavelet expansion

$$Kf = \sum_j \sum_k d_{jk} \psi_{jk}, \tag{8}$$

where ψ_{jk} are wavelets constructed to ensure that ψ_{jk} is in the range of K for all j and k, and $d_{jk} = \langle Kf, \psi_{jk} \rangle$. We should note that, although for convenience we keep the same notation for wavelets, the ψ_{jk} are now wavelets in the range of K and generally will be different from those in the original domain used in Section 2.2. The same will be true for the vaguelettes introduced below. Assume the existence of constants β_{jk} such that (4) holds for $v_{jk} = K^{-1}\psi_{jk}/\beta_{jk}$. If K is homogeneous of index α then the β_{jk} will be proportional to $2^{\alpha j}$. The function f is then recovered from (8) by expanding in the vaguelette series

$$f = \sum_j \sum_k \langle Kf, \psi_{jk} \rangle \beta_{jk} v_{jk} = \sum_j \sum_k \langle Kf, \psi_{jk} \rangle \Psi_{jk}, \tag{9}$$

where $\Psi_{jk} = K^{-1}\psi_{jk}$.

As in wavelet–vaguelette decomposition the wavelet coefficients of a noisy signal y, $\hat{d}_{jk} = \langle y, \psi_{jk} \rangle$, are contaminated by noise

$$\hat{d}_{jk} = d_{jk} + w_{jk}, \tag{10}$$

where $w_{jk} = \langle \varepsilon, \psi_{jk} \rangle$ are the coefficients of the wavelet decomposition of a white noise, and therefore are themselves a white noise; note that this is not the case for the corresponding vaguelette coefficients $\hat{b}_{jk}$ in (6) used in wavelet–vaguelette decomposition. Therefore the $\hat{d}_{jk}$ need to be denoised, for example by thresholding. The resulting vaguelette–wavelet decomposition estimator $\hat{f}_\lambda^{\mathrm{VWD}}$ will then be

$$\hat{f}_\lambda^{\mathrm{VWD}} = \sum_j \sum_k \delta_\lambda(\langle y, \psi_{jk} \rangle) \Psi_{jk}, \tag{11}$$

where $\delta_\lambda(\cdot)$ is a soft or hard thresholding operator.

We can of course consider the vaguelette–wavelet decomposition approach as a 'plug-in' estimator, in that we find a wavelet-based estimator of Kf and then apply K^{-1} to estimate f itself. The way in which the wavelets have been specified means that the operator K^{-1} can be applied to each wavelet individually. This allows the obvious extension of the vaguelette–wavelet decomposition approach to more general linear operators K; as long as the individual wavelets are in the range of K, the $K^{-1}\psi_{jk}$ can be found either analytically or by stable numerical methods, and therefore an estimate $\widehat{Kf}$ that has been found by a wavelet thresholding approach can be inverted term by term to give an estimate of f. The fact that wavelet thresholding has been used means that $\widehat{Kf}$ will in any case be a linear combination of only a small number of wavelets ψ_{jk}, thus contributing to the numerical stability of the procedure. Furthermore, in cases where the $K^{-1}\psi_{jk}$ have to be found individually by a numerical technique, it is only necessary to find those $K^{-1}\psi_{jk}$ that correspond to nonzero coefficients in $\widehat{Kf}$.

2.4 Discrete wavelet and vaguelette transforms

In practice, given the discrete data, we implement both approaches making use of a discrete wavelet or vaguelette transform to find the corresponding empirical wavelet or vaguelette coefficients. For the discrete case, all inner products are re-defined in the ℓ_2-sense, $\langle f, g\rangle = f_1 g_1 + \ldots + f_n g_n$.

Consider the discrete version of the original model (1)

$$y_i = (Kf)(t_i) + \varepsilon_i,$$

where $n = 2^J$ for some J, $t_i = i/n$ and the ε_i are independent $N(0, \sigma^2)$ random variables. The discrete wavelet transform, as described, for example, by Mallat (1989), derives empirical wavelet coefficients $\hat{d}_{jk}$ given by $\hat{d} = Wy$, where W is a particular orthogonal matrix satisfying $\sqrt{n} W_{jk,i} \approx \psi_{jk}(i/n) = 2^{j/2}\psi(2^j i/n - k)$. Both the transform and its inverse require only $O(n)$ operations. Several standard implementations are available, for example the `WaveThresh` package described by Nason and Silverman (1994). Note that, for any function f, the coefficients produced by a discrete wavelet transform of the sequence $f(t_i)$ are approximately $\sqrt{n}$ times the corresponding continuous wavelet coefficients $\langle f, \psi_{jk}\rangle$.

By analogy we can define the discrete vaguelette transform of the data y as $b = Vy$, where $\sqrt{n} V_{jk,i} = \tilde{\Psi}_{jk}(i/n)$. The matrix V is no longer an orthogonal matrix, and varies for different K, so performing the discrete vaguelette transform and its inverse may be computationally expensive in the general case. Because there will be no zeroes in the vaguelette expansion of y, in general, the wavelet–vaguelette decomposition method may require the whole matrix V, implicitly or explicitly. For homogeneous operators, Kolaczyk (1996) provides efficient algorithms for the discrete vaguelette transform and its inverse, each requiring $O(n \log^2 n)$ operations.

The following diagram summarizes the wavelet–vaguelette decomposition and vaguelette–wavelet decomposition approaches in practice:

$$\begin{array}{ll} \text{WVD:} & y \xrightarrow{\text{DVT}} \hat{b} \xrightarrow{\text{rescale}} \hat{b}^0 \xrightarrow{\text{threshold}} \delta_\lambda(\hat{b}^0) \xrightarrow{\text{IDWT}} \hat{f} \\ \text{VWD:} & y \xrightarrow{\text{DWT}} \hat{d} \xrightarrow{\text{threshold}} \delta_\lambda(\hat{d}) \xrightarrow{\text{IDVT}} \hat{f} \end{array}$$

where the inverse discrete vaguelette transform step may be carried out by performing an expansion in terms of the functions $\Psi_{jk} = K^{-1}\psi_{jk}$, or by performing an inverse discrete wavelet transform and then applying K^{-1} to the result.

3 Derivation of average mean square errors

A natural measure of the global performance of an estimate $\hat{f}$ of an unknown function f is the mean integrated square error $E\int(\hat{f} - f)^2$. In order to compare the performance of the various approaches, we studied the discrete version of this error measure, the average mean square error, defined as $\text{AMSE}(\hat{f}) = n^{-1}E\|\hat{f} - f\|_2^2 = n^{-1}E\sum(\hat{f}_i - f_i)^2$. In this section we derive exact formulae for the mean square error in the individual thresholded vaguelette and wavelet coefficients in (7) and (11) respectively, which are then used to make comparisons in the average mean square error without the need for any simulations.

The discrete vaguelette transform of the observed y in the wavelet–vaguelette decomposition, followed by appropriate rescaling, yields vaguelette coefficients $\hat{b}^0_{jk} \sim N(b^0_{jk}, \sigma^2)$. Straightforward calculations, given as (46) and (47) of Donoho & Johnstone (1994) for the special case $\sigma = 1$, give the following formulae for the mean square error of the individual vaguelette coefficients:

$$\begin{aligned} E\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\}^2 \;=\; & (b^0_{jk})^2 - \{(b^0_{jk})^2 - \sigma^2 - \lambda^2\}[\{\tilde{\Phi}\{(\lambda - b^0_{jk})/\sigma\} + \tilde{\Phi}\{(\lambda + b^0_{jk})/\sigma\}] \\ & - (\lambda + b^0_{jk})\sigma\phi\{(\lambda - b^0_{jk})/\sigma\} - (\lambda - b^0_{jk})\sigma\phi\{(\lambda + b^0_{jk})/\sigma\} \qquad (12) \end{aligned}$$

for soft thresholding and

$$
\begin{aligned}
E\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\}^2 &= (b^0_{jk})^2 - \{(b^0_{jk})^2 - \sigma^2\}[\tilde{\Phi}\{(\lambda - b^0_{jk})/\sigma\} + \tilde{\Phi}\{(\lambda + b^0_{jk})/\sigma\}] \\
&\quad + (\lambda - b^0_{jk})\sigma\phi\{(\lambda - b^0_{jk})/\sigma\} + (\lambda + b^0_{jk})\sigma\phi\{(\lambda + b^0_{jk})/\sigma\} \quad (13)
\end{aligned}
$$

for hard thresholding, where ϕ and Φ are the standard normal probability density and cumulative distribution functions, and $\tilde{\Phi} = 1 - \Phi$.

Since the wavelet basis (ψ_{jk}) is orthonormal, even though the $\hat{b}^0_{jk}$ are not independent, the average mean square error of the wavelet–vaguelette decomposition estimator is just the sum of the mean square errors of the individual coefficients, and therefore

$$
\mathrm{AMSE}(\hat{f}^{\mathrm{WVD}}_\lambda) = n^{-1} \sum_j \sum_k \|\tilde{\Psi}_{jk}\|^2 E\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\}^2, \qquad (14)
$$

where $E\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\}^2$ is given by (12) or (13) as appropriate.

The discrete wavelet transform of y used in vaguelette–wavelet decomposition yields noisy wavelet coefficients $\hat{d}_{jk} \sim N(d_{jk}, \sigma^2)$. The mean square error $E\{\delta_\lambda(\hat{d}_{jk}) - d_{jk}\}^2$ of individual thresholded coefficients is given by essentially the same formulae as (12) and (13), substituting d_{jk} for b^0_{jk} throughout. Note, however, that the wavelet coefficients $\hat{d}_{jk}$ are independent because of the orthogonality of the discrete wavelet transform, while the vaguelette coefficients $\hat{b}^0_{jk}$ are not.

Using the fact that the $\hat{d}_{jk}$ are independent, we have

$$
\begin{aligned}
\mathrm{AMSE}(\hat{f}^{\mathrm{VWD}}_\lambda) &= n^{-1} E\|\sum_{j,k}\{\delta_\lambda(\hat{d}_{jk}) - d_{jk}\}\Psi_{jk}\|^2 \\
&= n^{-1}\sum_{j,k} E\{\delta_\lambda(\hat{d}_{jk}) - d_{jk}\}^2\|\Psi_{jk}\|^2 \\
&\quad + n^{-1} \sum_{(j,k)\neq(i,l)} \{E\delta_\lambda(\hat{d}_{jk}) - d_{jk}\}\{E\delta_\lambda(\hat{d}_{il}) - d_{il}\}\langle\Psi_{jk}, \Psi_{il}\rangle \qquad (15)
\end{aligned}
$$

and by straightforward calculations

$$
\begin{aligned}
E\delta_\lambda(d_{jk}) - d_{jk} &= d_{jk}[\Phi\{(\lambda - d_{jk})/\sigma\} - \Phi\{(-\lambda - d_{jk})/\sigma\}] \\
&\quad + \sigma[\phi\{(\lambda + d_{jk})/\sigma\} - \phi\{(\lambda - d_{jk})/\sigma\}] \\
&\quad + \lambda[\Phi\{(\lambda + d_{jk})/\sigma\} - \Phi\{(\lambda - d_{jk})/\sigma\}]
\end{aligned}
$$

or

$$
\begin{aligned}
E\delta_\lambda(d_{jk}) - d_{jk} &= d_{jk}[\Phi\{(\lambda - d_{jk})/\sigma\} - \Phi\{(-\lambda - d_{jk})/\sigma\}] \\
&\quad + \sigma[\phi\{(\lambda + d_{jk})/\sigma\} - \phi\{\lambda - d_{jk})/\sigma\}]
\end{aligned}
$$

for the soft and hard thresholding rules respectively.

From (14) and (15) we can find, by numerical minimization over λ, ideal optimal thresholds that minimize the average mean square errors for particular wavelet–vaguelette decomposition and vaguelette–wavelet decomposition estimators of a given function f with operator K.

To study the efficiency of the wavelet decomposition methods for solving statistical linear inverse problems we also compare their average mean square errors with that of truncated singular value decomposition as described in Section 2.1. Using the notation defined there, we have $\hat{c}_j \sim N(c_j, \gamma^{-2}\sigma^2)$. By the orthogonality of the e_j,

$$\mathrm{AMSE}(\hat{f}_M^{\mathrm{SVD}}) = n^{-1}\left\{\sum_{j=1}^{M} E(\hat{c}_j - c_j)^2 + \sum_{j>M} c_j^2\right\} = n^{-1}(\sigma^2 \sum_{j=1}^{M} \gamma_j^{-2} + \sum_{j>M} c_j^2). \tag{16}$$

For given K and f we can then use a search to find the value of M that minimizes the average mean square error (16).

4 Comparison between approaches

In this section we use average mean square error to compare between different estimators. In order to provide a concrete basis on which to compare the methods, we consider the estimation of the derivative of a function as a particular example of an ill-posed problem. From (14) and (15) we find ideal optimal wavelet–vaguelette decomposition and vaguelette–wavelet decomposition estimators in terms of average mean square error for various test functions. In addition, we compare them with the optimal truncated singular value decomposition estimator.

4.1 Wavelet–vaguelette and vaguelette–wavelet decomposition approaches for estimating a derivative

Suppose that f is a function of interest, defined on the interval $[0,1]$. Let f be the n-vector of values $f(i/n)$, for $i = 1, \ldots, n$ and Kf be the n-vector defined by

$$(Kf)_i = \sum_{j \le i} f(j/n).$$

Suppose we observe a vector $y = (y_1, ..., y_n)^T$ of independent normally distributed observations with means $(Kf)_i$ and variance σ^2. Define Δ to be the finite difference operator $(\Delta g)_i = g\{(i+1)/n\} - g(i/n)$. In order to avoid complications caused by boundary effects, we shall consider functions for which $f(0) = f(1)$ and $\bar{f} = n^{-1}\sum f_i$ is zero, and use the periodic versions of the discrete wavelet and vaguelette transforms.

First we consider wavelet–vaguelette decomposition. To derive the vaguelette vectors $\tilde{\Psi}_{jk}$, it is easy to verify that $(K^*\tilde{\Psi}_{jk})_i = \psi_{jk,i}$ implies that $\tilde{\Psi}_{jk,i} = -(\Delta\psi_{jk})_i$. From (7) we then have

$$(\hat{f}_\lambda^{\mathrm{WVD}})_i = -\sum_j \sum_k \|\Delta\psi_{jk}\| \delta_\lambda\left(\langle y, \Delta\psi_{jk}\rangle / \|\Delta\psi_{jk}\|\right) \psi_{jk,i}\,.$$

To perform the vaguelette–wavelet decomposition note that $(K^{-1}f)_i = (\Delta f)_{i-1}$ and the vaguelette vectors corresponding to the expansion (9) are $\Psi_{jk,i} = (K^{-1}\psi_{jk})_i = (\Delta\psi_{jk})_{i-1}$. The resulting vaguelette–wavelet decomposition estimator is

$$\hat{f}_{\lambda,i}^{\mathrm{VWD}} = \sum_j \sum_k \delta_\lambda(\langle y, \psi_{jk}\rangle)(\Delta\psi_{jk})_{i-1},$$

where the wavelet coefficients $\langle y, \psi_{jk}\rangle$ are obtained by a discrete wavelet transform of the data.

4.2 Examples

The four test functions we considered are all based on the test functions 'bumps', 'blocks', 'HeaviSine' and 'Doppler' used by Donoho and Johnstone (1995), to which readers should refer for detailed descriptions. In each case, the number of sample points n was set to 512, so the various wavelet transforms have nine levels. The actual test vector f was defined from the vector of values f^0 of the Donoho–Johnstone function by performing the standardization $f = (f^0 - \bar{f^0})/\|f^0 - \bar{f^0}\|$, so that $\bar{f_i} = 0$ and $\sum f_i^2 = 1$ in each case.

The discrete wavelet or vaguelette transform yields $n-1$ wavelet or vaguelette coefficients d_{jk} or b_{jk} for $j = 0, ..., J-1$, $k = 0, ..., 2^j - 1$, the extra coefficient corresponding in the periodic case to the overall sum of the values of the function under consideration. Because of the shift transformation resulting in the condition $\bar{f} = 0$, this coefficient was not considered in the average mean square error; this would correspond to the use of an estimator where the result was transformed to satisfy the condition $\bar{f} = 0$.

For each of the test functions, and for a range of signal-to-noise ratios, we used our exact risk formulae to find the optimal values, in terms of average mean square error, of the thresholds λ for the wavelet–vaguelette decomposition and vaguelette–wavelet decomposition estimators. Soft threshold functions were used throughout. Because we are estimating the derivative of the directly observed function Kf, the signal-to-noise ratio was taken to be the ratio of the root-mean-square of the function values f to the population standard deviation $\sigma\sqrt{2}$ of the differenced noise $\Delta\varepsilon_i$. Note that $b_{jk}^0 = \langle Kf, \tilde{\Psi}_{jk}\rangle = \langle f, \psi_{jk}\rangle$, so the b_{jk}^0 in (14) can be calculated simply by performing a discrete wavelet transform of f.

Since integration increases the order of a function's smoothness by one, the regularity of the mother wavelet in the vaguelette–wavelet decomposition should be larger by one than that in the wavelet–vaguelette decomposition for a proper comparison. The wavelets used were the compactly supported extremal phase wavelets as defined in Section 6.4 of Daubechies (1992), writing D_m for the wavelet with $N = m$ in Daubechies' notation. These wavelets have m vanishing moments. The mother wavelets D_4 and D_8 were used in the wavelet–vaguelette decomposition and D_5 and D_9 in the vaguelette–wavelet decomposition.

The various average mean square errors yielded by the application of the exact risk formulae are given in Table 1. To make a comparison with singular value decomposition, we calculated the minimal average mean square error for a truncated singular value decomposition estimator with optimally chosen cut-off point M. The results are also given in Table 1. Table 2 gives the optimal values of the thresholds in terms of σ, each found by a grid search at grid interval 0.05σ, where σ is the standard deviation of the noise. For comparison, the ideal value of the threshold for the estimation of Kf itself by a wavelet thresholding method is also given in each case, for the wavelet D_5; the results for D_9 were very similar. All computations were done in the statistical package S-Plus using the `WaveThresh` software (Nason, 1993; Nason & Silverman, 1994).

Table 1 near here

Table 2 near here

4.3 Analysis of the results

Table 1 shows that the choice of whether to use wavelet–vaguelette decomposition or vaguelette–wavelet decomposition does not make a strong difference in terms of average mean square error, nor does the choice between the different wavelet functions. The largest discrepancy between the two wavelet-based methods is the improvement afforded by using vaguelette–wavelet decomposition for the HeaviSine function.

Both wavelet-based methods outperform the singular value decomposition method, especially at the larger signal-to-noise ratios. For the 'bumps' function, the singular value decomposition method has a generally relatively poor performance, presumably because of the substantial high-frequency component of this function. The 'blocks' and 'Doppler' functions are more inhomogeneous, but have a reasonable amount of signal at low frequencies. This probably explains why it is only at higher signal-to-noise ratios that the wavelet-based methods begin to show substantial improvements, but also why the relative advantage of a nonlinear method increases more rapidly with the signal-to-noise ratio. The fact that the performance of singular value decomposition is more reasonable in the HeaviSine case is probably not so surprising since this function is better approximated than the other test functions by Fourier expansions to limited numbers of terms.

It is especially informative to study the behaviour of the estimates as the signal-to-noise ratio varies. Decreasing the noise level is somewhat equivalent to increasing the sample size n and in this sense the signal-to-noise ratio can be considered as a surrogate sample size. Thus, we can examine the convergence for different methods by studying their performance as a function of signal-to-noise ratio. Table 1 clearly indicates that the rates of convergence for the wavelet-based methods are much faster than that of the singular value decomposition approach, especially for the 'blocks' and 'Doppler' functions. The theoretical ground for this phenomenon is given in Section 5.

The threshold values shown in Table 2 show that to a first approximation the same thresholds should be used for either wavelet-based approach. The thresholds for vaguelette–wavelet decomposition are if anything very slightly larger. It is interesting that the best threshold to use does depend substantially on the unknown function, indicating that there is still considerable progress to be made in the theoretical understanding of wavelet methods, and that universal thresholding is not likely to be always the best choice in practice.

The variation of the threshold with the signal to noise ratio is more dramatic in the case of the more inhomogeneous 'blocks' and 'Doppler' functions. Because of the overall greater importance of high frequency effects, smaller thresholds were needed for 'bumps' and 'blocks'. Finally, it is noteworthy that much smaller thresholds were appropriate for the estimation of Kf than of f. In terms of the vaguelette–wavelet decomposition method as a 'plug-in' approach, this indicates that the estimator to be plugged in is more strongly smoothed than would be the case if we were estimating the observed function Kf itself.

5 Theoretical results

In this section we consider the theoretical aspects of the wavelet-based estimators more rigorously, establish the asymptotic near-optimality, in the minimax sense, of the vaguelette–wavelet decomposition estimator and compare with analogous results for wavelet–vaguelette decomposition.

5.1 Notation and assumptions

First we introduce some definitions and notation. Consider the original model (1) and suppose that the operator K acts on some space $B^s_{p,q}$ from the Besov scale of functions of a real variable. This includes, in particular, such well-known spaces as Sobolev H^s ($B^s_{2,2}$), Hölder C^s ($B^s_{\infty,\infty}$), spaces of functions of bounded variation ($p = 1, s = 1$), and many other interesting examples. See Meyer (1992) or Donoho & Johnstone (1997) for rigorous definitions and details. The parameter s measures the number of derivatives whose existence is understood in an L_p-sense, while q provides some further flexibility. The fundamental property of wavelets is that given the mother wavelet of regularity τ, the corresponding wavelet basis is an unconditional basis within the whole range of Besov scale with $s < \tau$. This allows a parsimonious wavelet expansion for a wide set of different functions.

Define

$$r = (s + \alpha + 1/2)^{-1} s. \tag{17}$$

Let $s' = s + 1/2 - 1/p$, and assume that the parameters are such that

$$s > s_0 \text{ where } s_0 = 1/p + \max\{0, (2-p)\alpha/p - 1/2\}. \tag{18}$$

We shall assume that the function f is observed at points i/n for $i = 1, \ldots, n$, where $n = 2^J$ for some integer J. The observations will be assumed to have independent $N(f(i/n), \sigma^2)$ distributions, and it will be assumed that σ^2 is known. For simplicity of exposition, the function will be assumed to be periodic on $[0,1]$, and periodic versions of the various function spaces and wavelet transforms will be used. We measure accuracy of estimation of f in terms of the standard L_2 risk function $R(\hat{f}, f) = E \int (\hat{f} - f)^2$.

5.2 Minimax risks for particular Besov balls

Within the structure defined in Section 5.1, let R_n^* be the minimax risk for the estimation of f within a particular Besov ball $B_{p,q}^s(C_0)$ with radius C_0. The results of Donoho (1995) show that R_n^* converges to zero at exactly rate n^{-r} as n tends to infinity, with r defined as in (17). Donoho also shows that this rate of estimation is obtained by a suitable wavelet estimator, necessarily tuned to the particular Besov space under consideration. On the other hand, for $p < 2$ the corresponding rate of convergence $n^{-r'}$ for the minimax *linear* estimator has exponent $r' = (s' + \alpha + 1/2)^{-1} s'$, which is strictly less than r. Thus, neither the truncated singular value decomposition estimator nor any other linear estimator can attain the optimal performance within Besov spaces with $p < 2$, and this explains the inferior performance of singular value decomposition relative to wavelet–vaguelette decomposition for spatially inhomogeneous functions, especially for large values of the signal-to-noise ratio.

We shall call an estimator $\hat{f}$ *near-minimax in* $B_{p,q}^s(C_0)$ if up to a logarithmic factor it achieves the optimal rate of convergence, i.e. if, for all sufficiently large $n = 2^J$, $\sup_{f \in B_{p,q}^s(C_0)} R(\hat{f}, f) \leq C(\log n) R_n^*$ for some constant C. The reason for considering near-minimaxity rather than strict minimaxity is that it is possible, under suitable conditions, to construct estimators that are simultaneously near-minimax over a wide range of Besov scales, including both Sobolev and Hölder classes and various classes of spatially inhomogeneous functions. Such estimators may be considered as being spatially adaptive to the unknown smoothness. The logarithmic factor that the definition of near-minimaxity includes is a price for such spatial adaptivity that cannot be avoided; see also Lepskii (1990) and Goldenshluger & Nemirovski (1997).

5.3 Universal thresholding for linear inverse problems

Assume the mother wavelet ψ, generating the wavelet basis in $B_{p,q}^s$, has regularity $\tau > s$. In both the wavelet–vaguelette decomposition and vaguelette–wavelet decomposition estimators, we shall consider the use of soft threshold estimators with threshold

$$\lambda = \sigma \sqrt{2(1 + 2\alpha) \log n}. \tag{19}$$

We will show that, if this threshold is used, both approaches lead to spatially adaptive near-minimax estimators.

The threshold (19) can be considered as a universal threshold for the estimation of f from observations of Kf. It is noteworthy that it is higher by a factor of $\sqrt{1+2\alpha}$ than the usual 'universal threshold' used for the estimation of a function itself. Universal thresholding is known to be excessively conservative in general, and this is borne out by the fact that the thresholds in Table 2 for the estimation of Kf are substantially less than the value $\sigma\sqrt{2\log 512} = 3.53\sigma$. However, it is interesting that the values for the estimation of f in that table are generally between 1.5 and 2 times the value for Kf, in line with the ratio of $\sqrt{3}$ between the universal thresholds.

We can now state and prove the main theorem of this section.

Theorem 1 *For some fixed τ suppose that the wavelet–vaguelette decomposition estimator is constructed using wavelets of regularity τ and that the vaguelette–wavelet decomposition estimator is constructed using wavelets in the expansion (8) of regularity $\tau+\alpha$. With threshold (19), both the wavelet–vaguelette decomposition and vaguelette–wavelet decomposition estimators of f are then simultaneously near-minimax over all $B^s_{p,q}(C_0)$ with $p,q \geq 1$, $C_0 > 0$ and s satisfying $s_0 < s < \tau$, where s_0 is defined in (18).*

5.4 Proof

5.4.1 Setting up the problem in sequence space

Consider the wavelet–vaguelette decomposition approach first. Assume that the wavelets ψ_{jk} have regularity τ. Write $f = \sum_j \sum_k \theta_{jk}\psi_{jk}$, where $\theta_{jk} = \langle f, \psi_{jk}\rangle$ are continuous wavelet coefficients. We use slightly different notation from previously in order to be able to draw connections between the two approaches, and to avoid any confusion about the $\sqrt{n}$ factor introduced by the discrete wavelet transform. The condition that a set $\mathcal{F}$ has bounded $B^s_{p,q}$ norm is equivalent to a condition on the coefficients θ_{jk} of the form

$$\left.\begin{array}{ll} \sum_{j=0}^{\infty} 2^{js'q}\|\theta_j\|_p^q \leq C^q & q<\infty \\ \sup_{j\geq 0} 2^{js'}\|\theta_j\|_p \leq C & q=\infty \end{array}\right\} \tag{20}$$

where each θ_j is the 2^j-vector with elements θ_{jk}. See, for example, Meyer (1992).

Following the standard wavelet theory approach set out in Johnstone and Silverman (1997), for example, we consider the estimation problem within a sequence space context, and assume that we have observations $X_{jk} \sim N(\theta_{jk}, \sigma_j^2)$ for $j = 0, \ldots, J-1$ and $k = 0, \ldots, 2^j - 1$, where $\sigma_j^2 = \sigma_0^2 2^{2\alpha j}$. By assuming, without loss of generality, that the constant C_β as defined in Section 2.2 is equal to 1, we have $\sigma_0^2 = \sigma^2/n$. Note that the observations X_{jk} are not in general independent. The estimator of f is obtained by setting

$$\hat\theta_{jk} = \begin{cases} \delta_{\lambda\sigma_j}(X_{jk}) & \text{for } j<J \\ 0 & \text{for } j \geq J. \end{cases} \tag{21}$$

By the orthogonality of the wavelet basis, the integrated squared error $\int(\hat f - f)^2$ is equal to the sum of squares of errors in the individual coefficients

$$\|\hat\theta - \theta\|_2^2 = \sum_{j=0}^{\infty}\sum_k (\hat\theta_{jk} - \theta_{jk})^2.$$

Now consider vaguelette–wavelet decomposition. This time we have $f = \sum_j \sum_k \theta_{jk} v_{jk}$, where the v_{jk} are non-orthogonal vaguelettes and the θ_{jk} are now continuous *vaguelette* coefficients. The available observations, and the definition of the estimator $\hat{\theta}$, are exactly as above, only in this case the observations X_{jk} are independent.

Under the assumption that K is a homogeneous operator of index $\alpha > 0$, the image of any Besov space $B^s_{p,q}$ under K is another Besov space $B^{s+\alpha}_{p,q}$. Since we have assumed that the wavelets ψ_{jk} in the expansion (8) are of regularity $\tau > s + \alpha$, the function Kf will be in $B^{s+\alpha}_{p,q}$ if and only if its wavelet coefficients $\langle Kf, \psi_{jk}\rangle = \beta_{jk}^{-1}\theta_{jk}$ satisfy the condition (20) with s' replaced by $s' + \alpha$. Since, without loss of generality, the coefficients β_{jk} are equal to $2^{j\alpha}$, the coefficients θ_{jk} will satisfy exactly the same sequence space condition for f in $B^s_{p,q}$ as if the v_{jk} had been orthogonal wavelets of suitable regularity.

Finally, by the Riesz basis property (4), error as measured in terms of $\|\hat{\theta} - \theta\|_2^2$ will be equivalent in order of magnitude, though not in this case necessarily equal, to the integrated square error of $\hat{f}$.

5.4.2 The maximum risk of the sequence space problem

The discussion of the previous subsection shows that, whether one is considering the wavelet–vaguelette decomposition or the vaguelette–wavelet decomposition case, the proof can be completed by bounding the risk $E\|\hat{\theta} - \theta\|_2^2$ over θ satisfying (20), with the estimator $\hat{\theta}$ being defined as in (21) on the basis of observations $X_{jk} \sim N(\theta_{jk}, \sigma_j^2)$ where $\sigma_j^2 = n^{-1}\sigma^2 2^{2\alpha j}$. Without loss of generality we will assume that $\sigma^2 = 1$.

An argument given in Section 9.3 of Johnstone and Silverman (1997) shows that the tail sum

$$\sum_{j=J}^{\infty} \sum_k (\hat{\theta}_{jk} - \theta_{jk})^2 = \sum_{j=J}^{\infty} \sum_k \theta_{jk}^2 \tag{22}$$

is $O(n^{-2s})$ if $p \geq 2$ and $O(n^{-2s'})$ if $p < 2$. In either case, the definitions and conditions of Theorem 1 ensure that the sum is $o(n^{-r})$ as $n \to \infty$.

The bound given by Donoho and Johnstone (1994) for the mean square error of a single coefficient implies that, for $j < J$,

$$E(\hat{\theta}_{jk} - \theta_{jk})^2 \leq \{1 + 2(1 + 2\alpha)\log n\}\{n^{-(1+2\alpha)}\sigma_j^2 + \min(\theta_{jk}^2, \sigma_j^2)\}.$$

so that, since the sum (22) for $j \geq J$ is $o(n^{-r})$,

$$\begin{aligned} E\|\hat{\theta} - \theta\|^2 &\leq \{1 + 2(1 + 2\alpha)\log n\}\{n^{-(1+2\alpha)} \sum_{j=0}^{J-1} 2^j \sigma_j^2 + \sum_{j=0}^{J-1}\sum_k \min(\theta_{jk}^2, \sigma_j^2)\} + o(n^{-r}) \\ &< \{1 + 2(1 + 2\alpha)\log n\}\{n^{-1}(2^{1+2\alpha} - 1)^{-1} + S_2\} + o(n^{-r}). \end{aligned} \tag{23}$$

where we define

$$S_2 = \sum_{j=0}^{J-1} \sum_k \min(\theta_{jk}^2, \sigma_0^2 2^{2\alpha j}).$$

To obtain a bound for S_2, we extend the argument of the main part of Section 9.3 of Johnstone and Silverman (1997), which shows that for θ in $\mathcal{F}$

$$S_2 \leq \sum_{j=0}^{J-1} M_p(\sigma_0 2^{\alpha j}, C 2^{-js'}; 2^j) \tag{24}$$

where

$$M_p(\delta, c; m) = \begin{cases} \min(m\delta^2, c^p\delta^{2-p}) & \text{if } p \le 2 \\ \min(m\delta^2, c^2 n^{1-2/p}) & \text{if } 2 \le p \le \infty. \end{cases}$$

Define ζ by $2^\zeta = (Cn^{1/2})^{1/(s'+\alpha+1/p)}$. Then for $j \le \zeta$ the summands in (24) are proportional to $2^{(1+2\alpha)j}$, a geometrically increasing sequence. For $j \ge \zeta$ the summands are proportional to $2^{-2s'j}$ if $p \ge 2$ and $2^{-\{ps'-(2-p)\alpha\}j}$ if $p < 2$. In both cases the conditions of the theorem ensure that the sequence is geometrically decreasing. Hence, uniformly for all θ in $\mathcal{F}$, S_2 is bounded by a constant multiple of $2^{(1+2\alpha)\zeta}\sigma_0^2$ which is proportional to n^{-r}, since

$$1 - \frac{\alpha + 1/2}{s' + \alpha + 1/p} = \frac{s}{s + \alpha + 1/2} = r$$

by definition. Substituting back into (23) completes the proof of Theorem 1.

5.5 Comparative remarks

It is interesting that the theoretical basis for the two methods is so similar. Deeper examination of the results reveals some interesting features of the two methods, however. Because of the non-orthogonality of the vaguelette basis, the thresholding of vaguelette coefficients in wavelet–vaguelette decomposition is performed within coloured noise, while in vaguelette–wavelet decomposition the coefficients are uncorrelated. Johnstone and Silverman (1997) provide a lower bound theorem, which shows that the behaviour attained by a pointwise thresholding estimator of the kind we have considered is within a constant of the best possible behaviour. This constant will be subsumed within the implicit bounding constants in the definition of near-minimaxity in the theorem above. One factor that affects the constant is the degree of dependence between the coefficients, and on this basis the Johnstone–Silverman theory suggests that, of the two wavelet-based estimators, the vaguelette–wavelet decomposition should be closer to being uniformly minimax. This would certainly be the case if we compared the two estimators only on the basis of the performance measured in terms of $\|\hat{\theta} - \theta\|_2^2$.

However, in the case of vaguelette–wavelet decomposition, the Parseval identity does not hold, and this ℓ_2 norm is only equivalent, rather than equal, to the integrated square error norm on the functions themselves. Arising from the non-orthogonality of the vaguelette basis may be an extra factor in the bounding constant.

Finally, we note in passing that, by considering the zero function which is a member of all the function classes considered, it can be shown that the bound will not hold for any smaller value of the threshold λ_n. Indeed if $\lambda_n < \sqrt{4\alpha \log n}$, the estimator of f will not even be consistent in mean integrated square.

6 Conclusions

In the exact risk analyses we have presented the irregularity of the test functions has clearly had a deleterious effect on the performance of the truncated singular value decomposition method, and the wavelet-based estimators generally give better results.

For the examples considered, the two wavelet-based estimators are roughly comparable in terms of average mean square error. As shown in Section 5, both the vaguelette–wavelet decomposition and the wavelet–vaguelette decomposition can achieve similar asymptotic near-minimax behaviour. Further research is needed, however, to understand in what situations and problems the wavelet–vaguelette decomposition or the vaguelette–wavelet decomposition is to be preferred.

Our proposed vaguelette–wavelet decomposition approach has two features that are conceptually attractive. Firstly, for independent errors the thresholding is performed within white noise. Thresholding treats every coefficient separately, and while arguments such as those presented by Johnstone and Silverman (1997) show that thresholding correlated coefficients need not damage the order of magnitude of the estimation error, it is nevertheless clearly preferable for the individual thresholded quantities actually to be independent of one another. Secondly, the 'plug-in' characterization of vaguelette–wavelet decomposition, as discussed in Sections 2.3 and 2.4 makes it conceptually straightforward, and opens the possibility of using the approach for a wide range of linear inverse problems, though the computational details may have to be worked out in individual cases. It is certainly interesting that the use of a simple plug-in estimate gives performance as good as the conceptually somewhat more involved wavelet–vaguelette decomposition method.

Acknowledgements

The authors gratefully acknowledge the support of the British Engineering and Physical Sciences Research Council and the United States National Science Foundation. The referees made excellent suggestions which improved the paper substantially.

References

[1] Chui, C.K. (1992). *An Introduction to Wavelets.* San Diego: Academic Press.

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

[3] Donoho, D.L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comput. Harm. Anal.*, **2**, 101–126.

[4] Donoho, D.L. & Johnstone, I.M. (1994). Ideal spatial adaption by wavelet shrinkage. *Biometrika* **81**, 425–455.

[5] Donoho, D.L. & Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Ass.*, **90**, 1200–1224.

[6] Donoho, D.L. & Johnstone, I.M. (1997). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, in press.

[7] Donoho, D.L., Johnstone I.M., Kerkyacharian G. & Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc., Ser.* **B 57**, 301–369.

[8] Goldenshluger, A. & Nemirovski, A. (1997). On spatial adaptive estimation of nonparametric regression. *Math. Methods Statist.*, in press.

[9] Johnstone, I.M. & Silverman, B.W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18**, 251–280.

[10] Johnstone, I.M. & Silverman, B.W. (1991). Discretization effects in statistical inverse problems. *J. Complexity* **7**, 1–34.

[11] Johnstone, I.M. & Silverman, B.W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Statist. Soc.* B, in press.

[12] Kolaczyk, E.D. (1996). A wavelet shrinkage approach to tomographic image reconstruction. *J. Amer. Statist. Ass.* **91**, 1079–1090.

[13] Lepskii O. (1990). On a problem of adaptive estimation in gaussian white noise. *Theory Probab. Applic.* **35**, 454–466.

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

[15] Meyer, Y. (1992). *Wavelets and Operators.* Cambridge: Cambridge University Press.

[16] Nason, G.P. (1993). The `wavethresh` package: wavelet transform and thresholding, available from the StatLib archive.

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

[18] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Stat. Sci.* **1**, 502–527.

[19] Tikhonov, A.N. & Arsenin, V.Y. (1977). *Solutions of Ill-posed Problems.* New York: Wiley.

Table 1: Exact ideal average mean square error for the estimation of various test functions using the singular value decomposition (SVD), wavelet–vaguelette decomposition(WVD), and vaguelette–wavelet decomposition(VWD) approaches, for various levels of the signal-to-noise ratio SNR.

	SNR	SVD	WVD (D_4)	VWD (D_5)	WVD (D_8)	VWD (D_9)
Bumps	10	0.00999	0.00466	0.00478	0.00529	0.00530
	5	0.03946	0.01598	0.01662	0.01855	0.01870
	2	0.12337	0.07369	0.07466	0.08411	0.08576
	1	0.21752	0.19462	0.19165	0.21340	0.20711
Blocks	10	0.00947	0.00368	0.00341	0.00436	0.00394
	5	0.01855	0.01143	0.01118	0.01322	0.01219
	2	0.03475	0.03809	0.03732	0.03943	0.03705
	1	0.05223	0.06754	0.06686	0.06798	0.06441
HeaviSine	10	0.00099	0.00074	0.00063	0.00080	0.00071
	5	0.00146	0.00159	0.00137	0.00156	0.00134
	2	0.00238	0.00311	0.00301	0.00307	0.00262
	1	0.00340	0.00481	0.00465	0.00453	0.00443
Doppler	10	0.00532	0.00248	0.00260	0.00231	0.00216
	5	0.01070	0.00678	0.00738	0.00693	0.00668
	2	0.02224	0.02222	0.02091	0.01913	0.02029
	1	0.03636	0.04629	0.03942	0.04092	0.03975

Table 2: Optimal thresholds, in terms of the standard deviation σ of the noise, for various test functions using the wavelet–vaguelette decomposition(WVD) and vaguelette–wavelet decomposition(VWD) approaches, for various levels of the signal-to-noise ratio SNR. The last column gives the optimal threshold for the estimation of the integral of f rather than f itself, using the D_5 wavelet.

	SNR	WVD (D_4)	VWD (D_5)	WVD (D_8)	VWD (D_9)	Kf
Bumps	10	0.95	1.00	0.85	0.90	0.60
	5	1.05	1.05	0.95	1.00	0.65
	2	1.25	1.35	1.20	1.25	0.75
	1	1.55	1.65	1.55	1.60	0.90
Blocks	10	1.15	1.20	1.00	1.15	0.70
	5	1.30	1.35	1.20	1.30	0.80
	2	1.75	1.80	1.75	1.80	0.90
	1	2.20	2.25	2.20	2.25	1.00
HeaviSine	10	2.00	2.40	2.15	2.10	1.15
	5	2.35	2.85	2.35	2.45	1.25
	2	2.85	3.20	2.90	2.90	1.35
	1	3.15	3.35	3.20	3.15	1.45
Doppler	10	1.35	1.35	1.30	1.45	0.90
	5	1.60	1.55	1.60	1.60	0.95
	2	1.95	2.05	2.05	2.05	1.10
	1	2.25	2.40	2.25	2.35	1.20