Biometrika (1998), **85**, 1, *pp.* 115–129
Printed in Great Britain

Wavelet decomposition approaches to statistical inverse problems

BY F. ABRAMOVICH

Department of Statistics & Operations Research, Raymond & Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

felix@math.tau.ac.il

AND B. W. SILVERMAN

Department of Mathematics, University of Bristol, Bristol BS8 1TW, U.K.

b.w.silverman@bristol.ac.uk

SUMMARY

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 data are accessible only about some linear transform corrupted by noise. The usual linear methods for such inverse problems do not perform satisfactorily when $f(t)$ is spatially inhomogeneous. One existing nonlinear alternative is the wavelet–vaguelette decomposition method, based on the expansion of the unknown $f(t)$ in wavelet series. In the vaguelette–wavelet decomposition method proposed here, the observed data are expanded directly in wavelet series. The performances of various methods are compared through exact risk calculations, 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 the 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 problem; Vaguelette; Wavelet.

1. INTRODUCTION

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 $y(t_i)$ are

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

where $\varepsilon(t)$ is a Gaussian 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). Many such problems fall into the category of ill-posed problems, where, even in the absence of noise, one cannot recover f numerically from Kf simply by inverting the transform K. Ill-posed problems are usually treated by applying some linear regularisation 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 nonlinear 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, 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 § 6.

We shall use as a test problem the estimation of the derivative of a function g. This fits into the framework of (1) by setting K to be the integration operator and $g = Kf$. It is an important statistical problem, in contexts such as growth curves, to make inferences about growth rates, in economics, for example when estimating inflation rates from prices, and elsewhere.

In our 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 grounds 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 § 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 § 3, exact risk formulae are obtained for the various approaches, and these are used in § 4 to carry out a comparison on several examples without any need for simulation. In § 5 and the Appendix, the theoretical minimax properties of the estimators are explored. Some concluding remarks are made in § 6.

2. 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 $\langle .,. \rangle$ for the standard inner product in L_2. Ill-posed problems are characterised 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

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

for some truncation point M. 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.

In singular value decomposition, the basis is defined entirely by the operator K and ignores the specific physical nature of the problem under study. For example, the Fourier basis that arises for stationary operators does not provide a parsimonious approximation of 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; see Donoho (1995) for further discussion.

2·2. *The wavelet–vaguelette decomposition*

In response to these limitations, 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, are localised 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 also unconditional bases in a wide range of more specific spaces corresponding to varying degrees and kinds of smoothness; see § 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 $0 < A \leqslant B < \infty$ such that

$$A \sum_j \sum_k c_{jk}^2 \leqslant \left\| \sum c_{jk} v_{jk} \right\|^2 \leqslant B \sum_j \sum_k c_{jk}^2 \tag{3}$$

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

Obviously, only special operators K satisfy (3). For example, the condition holds for homogeneous operators, which, for all t_0, satisfy

$$K[f\{a(t-t_0)\}] = a^{-\alpha}(Kf)\{a(t-t_0)\}$$

for some constant α, called the index of the operator. Examples of homogeneous operators include integration, fractional integration and, in the two-dimensional case with an appropriate coordinate system, 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 are not mutually orthogonal. The property (3) also holds for various convolution operators; see Donoho (1995). We mainly restrict attention to operators K satisfying (3).

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, that is $\langle v_{jk}, u_{lm}\rangle = \delta_{jl}\delta_{km}$. Thus, if we observed the signal Kf without noise, we could expand it in a vaguelette series as

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

and then recover the original function f as

$$f = \sum_j \sum_k \langle Kf, u_{jk}\rangle \tilde{\beta}_{jk}^{-1}\psi_{jk} = \sum_j \sum_k \langle Kf, \tilde{\Psi}_{jk}\rangle \psi_{jk}, \tag{4}$$

where $\tilde{\Psi}_{jk} = u_{jk}/\tilde{\beta}_{jk}$ and, hence, $K^*\tilde{\Psi}_{jk} = \psi_{jk}$.

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

$$\hat{b}_{jk} = b_{jk} + w_{jk}, \tag{5}$$

where $b_{jk} = \langle Kf, \tilde{\Psi}_{jk}\rangle$ are the noiseless vaguelette coefficients from (4) and $w_{jk} = \langle \varepsilon, \tilde{\Psi}_{jk}\rangle$ are the vaguelette decomposition of a white noise. Since the vaguelettes $\tilde{\Psi}_{jk}$ are not orthonormal, the coefficients w_{jk} will not be independent or have equal variances, in general, but the rescaled coefficients defined by $\hat{b}^0_{jk} = \hat{b}_{jk}/\|\tilde{\Psi}_{jk}\|$ 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

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

or the hard threshold function

$$\delta_\lambda(x) = \begin{cases} x & \text{if } |x| > \lambda, \\ 0 & \text{otherwise,} \end{cases}$$

for some threshold value $\lambda \geqslant 0$. Mapping the thresholded coefficients back into the wavelet expansion in the original space yields the resulting wavelet–vaguelette decomposition estimator $\hat{f}^{\mathrm{WVD}}_\lambda$:

$$\hat{f}^{\mathrm{WVD}}_\lambda = \sum_j \sum_k \|\tilde{\Psi}_{jk}\| \delta_\lambda(\hat{b}^0_{jk})\psi_{jk}. \tag{6}$$

The computations will be easier for homogeneous operators, because the $\tilde{\Psi}_{jk}$ will be multiples of translations and dilations of a single mother vaguelette $\tilde{\Psi}_{00}$ with $\tilde{\Psi}_{jk} = 2^{j\alpha}2^{j/2}\tilde{\Psi}_{00}(2^j t - k)$. The norms $\|\Psi_{jk}\|$ are equal within each level j and, therefore, although the variances of the vaguelette coefficients $\hat{b}_{jk}$ on the jth 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 (6) 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 detail the wavelet–vaguelette decomposition algorithms for the Radon transform operator which arises, for example, in positron emission tomography, a problem also considered by Johnstone & 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 § 5 below.

2·3. *The vaguelette–wavelet decomposition*

A natural alternative to wavelet–vaguelette decomposition is the vaguelette–wavelet decomposition, which expands Kf rather than f in a wavelet series: threshold the wavelet coefficients of the observed data y to obtain an estimate of the wavelet expansion of Kf and then map back by K^{-1} to obtain an estimate of f in terms of vaguelette series.

Suppose we have a wavelet expansion

$$Kf = \sum_j \sum_k d_{jk} \psi_{jk}, \tag{7}$$

where ψ_{jk} are wavelets constructed to ensure that ψ_{jk} is in the range of K for all j and k, and $d_{jk} = \langle Kf, \psi_{jk} \rangle$. 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 § 2·2. The same will be true for the vaguelettes introduced below. Assume the existence of constants β_{jk} such that (3) 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 (7) by expanding in the vaguelette series

$$f = \sum_j \sum_k \langle Kf, \psi_{jk} \rangle \beta_{jk} v_{jk} = \sum_j \sum_k \langle Kf, \psi_{jk} \rangle \Psi_{jk}, \tag{8}$$

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, so that $\hat{d}_{jk} = d_{jk} + w_{jk}$, 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 (5) used in wavelet–vaguelette decomposition. Therefore the $\hat{d}_{jk}$ need to be denoised, for example by thresholding. The resulting vaguelette–wavelet decomposition estimator $\hat{f}_\lambda^{\text{VWD}}$ will then be

$$\hat{f}_\lambda^{\text{VWD}} = \sum_j \sum_k \delta_\lambda(\langle y, \psi_{jk} \rangle) \Psi_{jk}, \tag{9}$$

where $\delta_\lambda(.)$ 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 of 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 the estimate of 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.

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 redefined in the l_2-sense, so that $\langle f, g\rangle = f_1 g_1 + \ldots + f_n g_n$.

Suppose the original data are observed at points $t_i = i/n$, where $n = 2^J$ for some J. Let y be the vector of observations $y(t_i)$ in the model (1). The discrete wavelet transform, as described, for example, by Mallat (1989), yields empirical wavelet coefficients $\hat{d}_{jk}$ for $j = 0, \ldots, J-1$ and $k = 0, \ldots, 2^j - 1$, given by $\hat{d} = Wy$. Here W is an orthogonal matrix satisfying $\sqrt{n} W_{jk,i} \simeq \psi_{jk}(i/n) = 2^{j/2}\psi(2^j i/n - k)$. Both the transform and its inverse require only $O(n)$ operations. Several standard implementations are available, for example the WaveThresh package described by Nason & 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. Since there will be no zero value 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 summarises the wavelet–vaguelette decomposition and vaguelette–wavelet decomposition approaches in practice, using the abbreviations DWT and DVT for discrete wavelet and vaguelette transforms respectively, and IDWT and IDVT for their inverses:

$$\text{wavelet–vaguelette decomposition:} \quad y \xrightarrow{\text{DVT}} \hat{b} \xrightarrow{\text{rescale}} \hat{b}^0 \xrightarrow{\text{threshold}} \delta_\lambda(\hat{b}^0) \xrightarrow{\text{IDWT}} \hat{f},$$

$$\text{vaguelette–wavelet decomposition:} \quad y \xrightarrow{\text{DWT}} \hat{d} \xrightarrow{\text{threshold}} \delta_\lambda(\hat{d}) \xrightarrow{\text{IDVT}} \hat{f}.$$

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 SQUARED ERRORS

A natural measure of the global performance of an estimator $\hat{f}$ of an unknown function f is the mean integrated squared 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 squared 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 and use exact formulae for the mean squared error in the individual thresholded vaguelette and wavelet coefficients in (6) and (9) respectively.

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 (A2·1) and (A2·2) of Donoho & Johnstone (1994) for the special case $\sigma = 1$, give the following formulae for the mean squared error of the individual vaguelette coefficients:

$$E[\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\}^2] = (b^0_{jk})^2 - \{(b^0_{jk})^2 - \sigma^2 - \lambda^2\}[\tilde{\Phi}\{(\lambda - b^0_{jk})/\sigma\} + \tilde{\Phi}\{(\lambda + b^0_{jk})/\sigma\}] \\ - (\lambda + b^0_{jk})\sigma\phi\{(\lambda - b^0_{jk})/\sigma\} - (\lambda - b^0_{jk})\sigma\phi\{(\lambda + b^0_{jk})/\sigma\} \quad (10)$$

for soft thresholding, and

$$E[\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\}^2] = (b^0_{jk})^2 - \{(b^0_{jk})^2 - \sigma^2\}[\tilde{\Phi}\{(\lambda - b^0_{jk})/\sigma\} + \tilde{\Phi}\{(\lambda + b^0_{jk})/\sigma\}] \\ + (\lambda - b^0_{jk})\sigma\phi\{(\lambda - b^0_{jk})/\sigma\} + (\lambda + b^0_{jk})\sigma\phi\{(\lambda + b^0_{jk})/\sigma\} \quad (11)$$

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 squared error of the wavelet–vaguelette decomposition estimator is given by

$$\text{AMSE}(\hat{f}^{\text{WVD}}_\lambda) = n^{-1} \sum_j \sum_k \|\tilde{\Psi}_{jk}\|^2 E[\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\}^2], \quad (12)$$

where $E[\{\delta_\lambda(\hat{b}^0_{jk}) - b^0_{jk}\}^2]$ is given by (10) or (11) 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 squared error $E[\{\delta_\lambda(\hat{d}_{jk}) - d_{jk}\}^2]$ of individual thresholded coefficients is given by essentially the same formulae as (10) and (11), 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

$$\begin{aligned} \text{AMSE}(\hat{f}^{\text{VWD}}_\lambda) &= n^{-1} E\left[\left\| \sum_{j,k} \{\delta_\lambda(\hat{d}_{jk}) - d_{jk}\}\Psi_{jk} \right\|^2\right] \\ &= n^{-1} \sum_{j,k} E[\{\delta_\lambda(\hat{d}_{jk}) - d_{jk}\}^2] \|\Psi_{jk}\|^2 \\ &\quad + n^{-1} \sum_{(j,k) \neq (i,l)} [E\{\delta_\lambda(\hat{d}_{jk})\} - d_{jk}][E\{\delta_\lambda(\hat{d}_{il})\} - d_{il}]\langle \Psi_{jk}, \Psi_{il}\rangle \end{aligned} \quad (13)$$

and by straightforward calculations

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

for the soft and hard thresholding rules respectively.

From (12) and (13) we can find, by numerical minimisation over λ, ideal optimum thresholds that minimise the average mean squared 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 methods we also compare their average mean squared errors with that of truncated singular value decomposition as described in § 2·1.

Since the e_j are orthogonal and $\hat{c}_j \sim N(c_j, \gamma^{-2}\sigma^2)$,

$$\text{AMSE}(\hat{f}_M^{\text{SVD}}) = n^{-1}\left\{\sum_{j=1}^{M} E(\hat{c}_j - c_j)^2 + \sum_{j>M} c_j^2\right\} = n^{-1}\left(\sigma^2 \sum_{j=1}^{M} \gamma_j^{-2} + \sum_{j>M} c_j^2\right). \quad (14)$$

For given K and f we can then use a search to find the value of M that minimises (14).

4. Comparison of methods

4·1. *General approach*

To compare the methods, we consider the estimation of the derivative of a function. From (12) and (13) we find ideal optimal wavelet–vaguelette decomposition and vaguelette–wavelet decomposition estimators in terms of average mean squared error for various test functions. In addition, we compare them with the optimal truncated singular value decomposition estimator.

4·2. *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

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

Suppose we observe a vector $y = (y_1, \ldots, y_n)^{\text{T}}$ 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 (6) we then have

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

To perform the vaguelette–wavelet decomposition note that $(K^{-1}f)_i = (\Delta f)_{i-1}$ and the vaguelette vectors corresponding to the expansion (8) are $\Psi_{jk,i} = (K^{-1}\psi_{jk})_i = (\Delta\psi_{jk})_{i-1}$. The resulting vaguelette–wavelet decomposition estimator is

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

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

4·3. *Examples*

Our four test functions are based on those used by Donoho & Johnstone (1995). In each case, n was set to 512, so the various wavelet transforms have nine levels. The actual test vector f was defined by standardising the vector of values f^0 of the Donoho–Johnstone function to $f = (f^0 - \bar{f}^0)/\|f^0 - \bar{f}^0\|$, 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 $j = 0, \ldots, J-1$ and $k = 0, \ldots, 2^j - 1$, together with a scaling function coefficient that corresponds in the periodic case to the overall mean of the values of the function under consideration. Since the shift transformation results in the condition $\bar{f}=0$, this coefficient was not considered in the average mean squared error; this corresponds to an estimator 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 squared error, of the thresholds λ for the wavelet–vaguelette decomposition and vaguelette–wavelet decomposition estimators. Soft threshold functions were used throughout. Since we are estimating the derivative of the directly observed function Kf, the root 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 $\Delta\varepsilon(t_i)$. Note that $b^0_{jk}=\langle Kf, \tilde{\Psi}_{jk}\rangle = \langle f, \psi_{jk}\rangle$, so the b^0_{jk} in (12) 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 § 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 squared errors based on the exact risk formulae are given in Table 1, which also contains the minimal average mean squared error for a truncated singular value decomposition estimator with optimally chosen cut-off point M. Table 2 gives the optimal values of the thresholds in terms of σ, each found by a grid search at grid interval $0{\cdot}05\sigma$, where σ is the standard deviation of the noise, as well as the ideal value of the threshold for the estimation of Kf itself by a wavelet thresholding method, 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 & Silverman, 1994), available from the StatLib archive.

Table 1. *Exact ideal average mean squared 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 root signal-to-noise ratio* (RSNR) *and wavelet D_m*

	Ideal average mean squared error					Ideal average mean squared error				
RSNR	SVD	WVD (D_4)	VWD (D_5)	WVD (D_8)	VWD (D_9)	SVD	WVD (D_4)	VWD (D_5)	WVD (D_8)	VWD (D_9)
			Bumps					Blocks		
10	0·00999	0·00466	0·00478	0·00529	0·00530	0·00947	0·00368	0·00341	0·00436	0·00394
5	0·03946	0·01598	0·01662	0·01855	0·01870	0·01855	0·01143	0·01118	0·01322	0·01219
2	0·12337	0·07369	0·07466	0·08411	0·08576	0·03475	0·03809	0·03732	0·03943	0·03705
1	0·21752	0·19462	0·19165	0·21340	0·20711	0·05223	0·06754	0·06686	0·06798	0·06441
			HeaviSine					Doppler		
10	0·00099	0·00074	0·00063	0·00080	0·00071	0·00532	0·00248	0·00260	0·00231	0·00216
5	0·00146	0·00159	0·00137	0·00156	0·00134	0·01070	0·00678	0·00738	0·00693	0·00668
2	0·00238	0·00311	0·00301	0·00307	0·00262	0·02224	0·02222	0·02091	0·01913	0·02029
1	0·00340	0·00481	0·00465	0·00453	0·00443	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 root signal-to-noise ratio* (RSNR). *The last column gives the optimal threshold for the estimation of the integral of f rather than f itself, using the D_5 wavelet*

	Optimal thresholds					Optimal thresholds				
RSNR	WVD (D_4)	VWD (D_5)	WVD (D_8)	VWD (D_9)	Kf	WVD (D_4)	VWD (D_5)	WVD (D_8)	VWD (D_9)	Kf
			Bumps					Blocks		
10	0·95	1·00	0·85	0·90	0·60	1·15	1·20	1·00	1·15	0·70
5	1·05	1·05	0·95	1·00	0·65	1·30	1·35	1·20	1·30	0·80
2	1·25	1·35	1·20	1·25	0·75	1·75	1·80	1·75	1·80	0·90
1	1·55	1·65	1·55	1·60	0·90	2·20	2·25	2·20	2·25	1·00
			HeaviSine					Doppler		
10	2·00	2·40	2·15	2·10	1·15	1·35	1·35	1·30	1·45	0·90
5	2·35	2·85	2·35	2·45	1·25	1·60	1·55	1·60	1·60	0·95
2	2·85	3·20	2·90	2·90	1·35	1·95	2·05	2·05	2·05	1·10
1	3·15	3·35	3·20	3·15	1·45	2·25	2·40	2·25	2·35	1·20

4·4. *Analysis of the results*

Table 1 exhibits no strong difference in performance between the two wavelet methods, nor between the different wavelet functions considered. Both wavelet-based methods outperform the singular value decomposition method, especially at the larger signal-to-noise ratios. The Blocks and Doppler functions have a reasonable amount of signal at low frequencies, which may explain why it is only at higher signal-to-noise ratios that the wavelet-based methods begin to show substantial improvements, and also why the relative advantage of a nonlinear method increases more rapidly with the signal-to-noise ratio. The relatively favourable performance of singular value decomposition for the HeaviSine function may be accounted for by its reasonable approximation by a limited length Fourier expansion.

Since the signal-to-noise ratio may be considered as a surrogate for sample size, we can examine the convergence for different methods by studying their performance as a function of signal-to-noise ratio. Table 1 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 grounds for this phenomenon are given in § 5.

Table 2 indicates that to a first approximation the same thresholds should be used for either wavelet-based approach. The best threshold depends substantially on the unknown function, so 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, note 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 of Kf to be plugged in is more strongly smoothed than the best estimate of Kf itself.

5. Theoretical results

5·1. *Preamble, notation and assumptions*

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.

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 Sobolev spaces H^s ($B^s_{2,2}$), Hölder spaces C^s ($B^s_{\infty,\infty}$), spaces of functions of bounded variation ($p = 1, s = 1$) and many others; see Meyer (1992) or Donoho & Johnstone (1998) 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

$$r = (s + \alpha + \tfrac{1}{2})^{-1} s. \qquad (15)$$

Let $s' = s + \frac{1}{2} - 1/p$, and assume that the parameters are such that $s > s_0$, where

$$s_0 = p^{-1} + \max(0, 2\alpha p^{-1} - \alpha - \tfrac{1}{2}). \qquad (16)$$

We shall assume that the function Kf is observed at regular intervals with independent $N(0, \sigma^2)$ noise as described in § 4, with σ^2 known and $n = 2^J$ for some integer J, and the same periodicity assumptions. We measure estimation accuracy by the standard 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 § 5·1, let R_n^* be the minimax risk for the estimation of f within a particular Besov ball $B^s_{p,q}(C_0)$ with radius C_0. The result of Donoho (1995) shows that R_n^* converges to zero at exactly rate n^{-r} as n tends to infinity, with r defined as in (15), and that this rate of estimation is obtained by a suitable wavelet estimator, necessarily tuned to the particular Besov space. On the other hand, for $p < 2$ the corresponding rate of convergence $n^{-\rho}$ for the minimax linear estimator has exponent $\rho = (s' + \alpha + \frac{1}{2})^{-1} s'$, 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 $B^s_{p,q}(C_0)$ if up to a logarithmic factor it achieves the optimal rate of convergence, i.e. if, for all sufficiently large n,

$$\sup_{f \in B^s_{p,q}(C_0)} R(\hat{f}, f) \leqslant C(\log n) R_n^*$$

for some constant C. Under suitable conditions, it is possible to construct estimators that are simultaneously near-minimax over a wide range of Besov scales. 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 has regularity $\tau > s$. We shall consider the use of soft threshold estimators with threshold

$$\lambda_n = \sigma\sqrt{\{2(1+2\alpha)\log n\}}. \tag{17}$$

With this threshold, both the wavelet approaches lead to spatially adaptive near-minimax estimators.

The threshold (17) can be considered as a universal threshold for the estimation of f from observations of Kf. 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: note that the thresholds in Table 2 for the estimation of Kf are substantially less than the value $\sigma\sqrt{(2\log 512)} = 3{\cdot}53\sigma$. However, the best thresholds for the estimation of f are generally between 1·5 and 2 times their value for Kf, in line with the ratio of $\sqrt{3}$ between the universal thresholds.

We can now state the main theorem. The proof is given in the Appendix.

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* (7) *of regularity $\tau+\alpha$. With threshold* (17), *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 $p, q \geqslant 1$, $C_0 > 0$ and s satisfying $s_0 < s < \tau$, where s_0 is defined in* (16).

By considering the zero function, which is a member of all the function classes considered, it can be shown that the Theorem 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. CONCLUDING REMARKS

Though the theoretical basis for the two methods is apparently very similar, deeper examination of the results reveals some interesting differences. The thresholding of vaguelette coefficients in wavelet–vaguelette decomposition is performed within coloured noise, while in vaguelette–wavelet decomposition the coefficients are independent. The results of Johnstone & Silverman (1997) imply that the behaviour attained by these estimators is within a constant of the best possible. The constant is affected by the dependence between the coefficients, and the theory indicates that, of the two wavelet-based estimators, the vaguelette–wavelet decomposition will be closer to being uniformly minimax, if we measure performance only in terms of mean squared error in coefficient space. However, in the case of vaguelette–wavelet decomposition, the non-orthogonality of the vaguelette basis allows an extra factor in the bounding constant when we consider the integrated squared error of the functions themselves.

To conclude, our proposed vaguelette–wavelet decomposition approach has two features that are conceptually attractive. First, for independent errors the thresholding is performed within white noise. Thresholding treats every coefficient separately, and, while arguments such as those presented by Johnstone & 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 characterisation of vaguelette–wavelet decomposition 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, though further research is needed to understand in what situations and problems the wavelet–vaguelette decomposition or the vaguelette–wavelet decomposition is to be preferred.

Acknowledgement

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.

Appendix

Proof of Theorem 1

Setting up the wavelet–vaguelette decomposition case in sequence space. 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 modify our notation slightly to draw connections between the two approaches and to avoid confusion about the $\sqrt{n}$ factor introduced by the discrete wavelet transform. The condition that a set $\mathscr{F}$ has bounded $B^s_{p,q}$ norm is equivalent to a condition on the coefficients θ_{jk} of the form

$$\begin{cases} \sum_{j=0}^{\infty} 2^{js'q} \|\theta_j\|_p^q \leqslant C^q & \text{for } q<\infty, \\ \sup_{j\geqslant 0} 2^{js'} \|\theta_j\|_p \leqslant C & \text{for } q=\infty, \end{cases} \tag{A1}$$

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 & 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 § 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

$$\hat{\theta}_{jk}=\begin{cases} \delta_{\lambda\sigma_j}(X_{jk}) & \text{for } j<J, \\ 0 & \text{for } j\geqslant J. \end{cases} \tag{A2}$$

By the orthogonality of the wavelet basis,

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

The vaguelette–wavelet decomposition case in sequence space. 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 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 (7) 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 condition (A1) 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 (3), 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 squared error of $\hat{f}$.

The maximum risk of the sequence space problem. The above discussion 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 (A1), with the estimator $\hat{\theta}$ being defined as in (A2) 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 § 9.3 of Johnstone & Silverman (1997) shows that the tail sum

$$\sum_{j=J}^{\infty}\sum_k (\hat{\theta}_{jk}-\theta_{jk})^2 = \sum_{j=J}^{\infty}\sum_k \theta_{jk}^2 \tag{A3}$$

is $O(n^{-2s})$ if $p \geqslant 2$, 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 & Johnstone (1994) for the mean squared error of a single coefficient implies that, for $j < J$,

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

so that, since the sum (A3) for $j \geqslant J$ is $o(n^{-r})$,

$$\begin{aligned} E\|\hat{\theta}-\theta\|^2 &\leqslant \{1+2(1+2\alpha)\log n\}\left\{n^{-(1+2\alpha)}\sum_{j=0}^{J-1} 2^j\sigma_j^2 + \sum_{j=0}^{J-1}\sum_k \min(\theta_{jk}^2, \sigma_j^2)\right\} + o(n^{-r}) \\ &< \{1+2(1+2\alpha)\log n\}\{n^{-1}(2^{1+2\alpha}-1)^{-1} + S_2\} + o(n^{-r}), \end{aligned} \tag{A4}$$

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 § 9.3 of Johnstone & Silverman (1997), which shows that for θ in $\mathscr{F}$

$$S_2 \leqslant \sum_{j=0}^{J-1} M_p(\sigma_0 2^{\alpha j}, C2^{-js'}; 2^j), \tag{A5}$$

where

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

Define ζ by $2^\zeta = (Cn^{1/2})^{1/(s'+\alpha+1/p)}$. Then for $j \leqslant \zeta$ the summands in (A5) are proportional to $2^{(1+2\alpha)j}$, a geometrically increasing sequence. For $j \geqslant \zeta$ the summands are proportional to $2^{-2s'j}$ if $p \geqslant 2$ 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 $\mathscr{F}$, S_2 is bounded by a constant multiple of $2^{(1+2\alpha)\zeta}\sigma_0^2$. This is proportional to n^{-r}, since

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

by definition. Substituting back into (A4) completes the proof of Theorem 1.

References

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

Daubechies, I. (1992). *Ten Lectures on Wavelets.* Philadelphia: Society for Industrial and Applied Mathematics.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition. *Appl. Comp. Harm. Anal.* **2**, 101–26.

Donoho, D. L. & Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.

Donoho, D. L. & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.* **90**, 1200–24.

Donoho, D. L. & Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* To appear.

Goldenshluger, A. & Nemirovski, A. (1997). On spatial adaptive estimation of nonparametric regression. *Math. Meth. Statist.* **6**, 135–70.

Johnstone, I. M. & Silverman, B. W. (1990). Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18**, 251–80.

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

Johnstone, I. M. & Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc.* B **59**, 319–51.

Kolaczyk, E. D. (1996). A wavelet shrinkage approach to tomographic image reconstruction. *J. Am. Statist. Assoc.* **91**, 1079–90.

Lepskii, O. (1990). On a problem of adaptive estimation in gaussian white noise. *Theory Prob. Applic.* **35**, 454–66.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Patt. Anal. Mach. Intel.* **11**, 674–93.

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

Nason, G. P. & Silverman, B. W. (1994). The discrete wavelet transform in S. *J. Comp. Graph. Statist.* **3**, 163–91.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with Discussion). *Statist. Sci.* **1**, 502–27.

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

[*Received August* 1995. *Revised June* 1997]